# The Average Cost of Markov Chains Subject to Total Variation Distance Uncertainty☆☆

A. A. Malikopoulos[a,1,*], C. D. Charalambous[b,1], I. Tzortzis[b,1]

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

**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.

*Keywords:* Stochastic optimal control, controlled Markov chain, average cost, total variation distance.

## 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, 2, 3, 4] and references therein). A significant amount of research has been also reported for the problem with finite state and action spaces [5, 6, 7, 8, 9, 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 prove 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. Hordijk [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, 15, 16, 17, 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 under the award number DE–AR0000796. This support is gratefully acknowledged.

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

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, 25, 26, 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 $t = 0, 1, ...$, 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 map the state space to the control space defined as the control law. Each sequence $\pi$ of the functions $\phi_i$, $\pi = \{\phi_1, ..., \phi_{|\mathcal{S}|}\}$, 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 $\pi^*$ 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 $\pi = \{\phi_1, \phi_2, ..., \phi_{|\mathcal{S}|}\}$, the MC $\{X_t|t = 1, 2, ...\}$ 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 $\mathbf{1} = [1, 1, ..., 1]^T$ 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 $l(\pi) = [l(1, \phi_1), l(2, \phi_2), ..., l(i, \phi_i), ...l(\mathcal{S}, \phi_{|\mathcal{S}|})]^T$ 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 $\Pi$ 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 $\pi$, and we denote them simply by $J$, $\mu$, 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 $\Gamma$ is compact there exists a cost-per-stage that minimizes the functional $J$. The following section provides the solution of the above optimization problem.

## 3. Discrepancy Measure: Total Variation distance

### *3.1. Total variation distance*

The Markov process has a single ergodic class (Assumption 2.1), and thus a unique invariant distribution $\mu = [\mu_1,\ \mu_2, \ldots, \mu_{|\mathcal{S}|}]$. The objective is to approximate the Markov process $\{X_t : t = 0, 1, \ldots\}$ by another, non necessarily Markov process, $\{Y_t : t = 0, 1, \ldots\}$ with state space $\bar{\mathcal{S}} \subseteq \mathcal{S}$, and invariant distribution $\bar{\nu} = [\bar{\nu}_1,\ \bar{\nu}_2, \ldots, \bar{\nu}_{|\bar{\mathcal{S}}|}]$, with respect to an appropriate measure of proximity between the original Markov process $\{X_t : t = 0, 1, \ldots\}$ and the approximating process $\{Y_t : t = 0, 1, \ldots\}$, called the *discrepancy measure*.

The distance metric we use to define the discrepancy between two distributions is the TV distance, to allow distributions defined on different state spaces. To this end, we introduce TV distance for general spaces, as follows. Let $\mathcal{M}_1(\mathcal{S})$ denote the set of probability measures on $\mathcal{B}(\mathcal{S})$. The TV distance is a metric, $||\cdot||_{TV} : \mathcal{M}_1(\mathcal{S}) \times \mathcal{M}_1(\mathcal{S}) \to [0, \infty)$ defined by [36]

$$||\alpha - \mu||_{TV} \triangleq \sup_{A \in \mathcal{B}(\mathcal{S})} |\alpha(A) - \mu(A)| \tag{5}$$

where $\alpha, \mu \in \mathcal{M}_1(\mathcal{S})$. With respect to this metric, $(\mathcal{M}_1(\mathcal{S}), ||\cdot||_{TV})$ is a complete metric space. Given a probability measure $\mu \in \mathcal{M}_1(\mathcal{S})$ define the fidelity set via the ball, with respect to the TV distance, centered at the measure $\mu \in \mathcal{M}_1(\mathcal{S})$, having radius $R \in [0, 2]$, by

$$\mathbb{B}_R(\mu) \triangleq \big\{\nu \in \mathcal{M}_1(\mathcal{S}) : ||\nu - \mu||_{TV} \leq R\big\}. \tag{6}$$

By the properties of the distance metric then $||\nu - \mu||_{TV} \leq ||\nu||_{TV} + ||\mu||_{TV} = 2$, hence $R$ is further restricted to the interval $[0, 2]$.

### *3.2. Approximation problem based on maximum entropy principle*

Consider the finite alphabet case $(\mathcal{S}, \mathcal{M})$, with cardinality $|\mathcal{S}|$, $\mathcal{M} = 2^{|\mathcal{S}|}$. Thus, $\nu$ and $\mu$ are point mass distributions on $\mathcal{S}$. Define the set of probability vectors on $\mathcal{S}$ by

$$\mathbb{P}(\mathcal{S}) \triangleq \big\{p = (p_1, \ldots, p_{|\mathcal{S}|}) : p_i \geq 0, i \in \mathcal{S}, \sum_{i \in \mathcal{S}} p_i = 1\big\}.$$

Thus, $p \in \mathbb{P}(\mathcal{S})$ is a probability vector in $\mathbb{R}_+^{|\mathcal{S}|}$. Also, let $l \triangleq \{l_1, \ldots, l_{|\mathcal{S}|}\}$ so that $l \in \mathbb{R}_+^{|\mathcal{S}|}$ (e.g., set of non-negative vectors of dimension $|\mathcal{S}|$). Given the invariant distribution $\mu \in \mathbb{P}(\mathcal{S})$ and a parameter $R \in [0, 2]$ define the long-run expected average cost criterion with respect to the invariant distribution $\{\nu_i : i \in \mathcal{S}\} \in \mathbb{B}_R(\mu) \subset \mathbb{P}(\mathcal{S})$ by

$$\mathbb{L}(\nu) = \sum_{i \in \mathcal{S}} l_i \nu_i, \quad l \in \mathbb{R}_+^{|\mathcal{S}|}. \tag{7}$$

The objective is to approximate $\mu \in \mathbb{P}(\mathcal{S})$ by $\nu \in \mathbb{B}_R(\mu)$ by solving the maximization problem defined by

$$\mathbb{L}(\nu^*) = \max_{\substack{\nu \in \mathbb{B}_R(\mu) \\ \mu = \mu P}} \mathbb{L}(\nu), \quad \forall R \in [0, 2]. \tag{8}$$

Problem (8) is a non-decreasing concave function of $R$, and for $R \leq R_{\max}$ the inequality constraint holds with equality, where $R_{\max}$ is the smallest non-negative number belonging to $[0, 2]$ such that $\mathbb{L}(\nu^*)$ is constant in $[R_{\max}, 2]$ (for more details see [37]). Hence, Problem (8) is a convex optimization problem on the space of probability measures.

Consider Jayne's maximum entropy principle[1] then, the approximation problem can be formulated as follows: maximize the entropy of $\{\nu_i : i \in \mathcal{S}\}$ subject to TV fidelity set, defined by

$$\max_{\substack{\nu \in \mathbb{B}_R(\mu) \\ \mu = \mu P}} H(\nu), \quad H(\nu) \triangleq -\sum_{i \in \mathcal{S}} \log(\nu_i)\nu_i. \tag{9}$$

Problem (9) is of interest when the concept of insufficient reasoning (e.g., Jayne's maximum entropy principle [38]) is applied to construct a model for $\nu \in \mathbb{P}(\mathcal{S})$, subject to information quantified via the fidelity set defined by the variation distance between $\nu$ and $\mu$.

[1] The maximum entropy principle states that, subject to precisely stated prior data, the probability distribution which best represents the current state of knowledge is the one with largest entropy.

It can be shown that the maximum entropy approximation problem (9) is precisely equivalent to the problem of finding the approximating distribution corresponding to the minimum description code word length, also called as universal coding problem [39], as follows. Let $\{l_i : i \in \mathcal{S}\}$ denote the positive codeword lengths corresponding to each symbol of the approximating distribution, which satisfy the Kraft inequality of lossless Shannon codes $\sum_{i\in\mathcal{S}} D^{-l_i} \leq 1$, where the code word alphabet is $D$-ary (unless specified otherwise $\log(\cdot) \triangleq \log_D(\cdot)$). Then, by the Von-Neumann's theorem[2] we have that

$$\min_{l\in\mathbb{R}_+^{|\mathcal{S}|}:\sum_{i\in\mathcal{S}} D^{-l_i}\leq 1} \max_{\substack{\nu\in\mathbb{B}_R(\mu)\\ \mu=\mu P}} \sum_{i\in\mathcal{S}} l_i\nu_i$$

$$= \max_{\substack{\nu\in\mathbb{B}_R(\mu)\\ \mu=\mu P}} \min_{l\in\mathbb{R}_+^{|\mathcal{S}|}:\sum_{i\in\mathcal{S}} D^{-l_i}\leq 1} \sum_{i\in\mathcal{S}} l_i\nu_i = \max_{\substack{\nu\in\mathbb{B}_R(\mu)\\ \mu=\mu P}} H(\nu). \quad (10)$$

Hence, for $l_i \triangleq -\log\nu_i$, $\forall i \in \mathcal{S}$, the approximated Problem (8) is equivalent to (9). The interpretation is that the maximum entropy approximation corresponds to the minimum description length [39] of a class of distributions described by the TV ball.

### *3.3. Solution of the approximation problem*

In this section, we draw upon the results of [37] to find the solution of Problem (8), and consequently the solution of Problem (9). First, we identify the support sets and their corresponding values.

Let us define the maximum and minimum values of the sequence $\{l_1, \ldots, l_{|\mathcal{S}|}\}$ by $l_{\max} \triangleq \max_{i\in\mathcal{S}} l_i$, $l_{\min} \triangleq \min_{i\in\mathcal{S}} l_i$, and its corresponding support sets by

$$\mathcal{S}^0 \triangleq \{i \in \mathcal{S} : l_i = l_{\max}\}, \quad \mathcal{S}_0 \triangleq \{i \in \mathcal{S} : l_i = l_{\min}\}.$$

For all remaining sequence, $\{l_i : i \in \mathcal{S} \setminus \mathcal{S}^0 \cup \mathcal{S}_0\}$, denote the set of indices for which $l$ achieves its $(k+1)^{th}$ smallest value by $S_k$, $k \in \{1, 2, \ldots, |\mathcal{S}\setminus\mathcal{S}^0\cup\mathcal{S}_0|\}$ (till all the elements of $\mathcal{S}$ are exhausted) and denote the corresponding values of the sequence of $\mathcal{S}_k$ sets by $l(\mathcal{S}_k)$,

For $l \in B^+(\mathcal{S})$, and $\mu \in \mathbb{P}(\mathcal{S})$ fixed, we show in [40], that the optimal probabilities of Problem (8) are given by

$$\nu^*(\mathcal{S}^0) = \min\Big(\frac{1}{|\mathcal{S}|}, \frac{\sum_{i\in\mathcal{S}^0}\mu_i + \frac{R}{2}}{|\mathcal{S}^0|}\Big), \quad (11a)$$

$$\nu^*(\mathcal{S}_0) = \max\Big(\frac{1}{|\mathcal{S}|}, \frac{\sum_{i\in\mathcal{S}_0}\mu_i - \frac{R}{2}}{|\mathcal{S}_0|}\Big), \quad (11b)$$

$$\nu^*(\mathcal{S}_k) = \sum_{i\in\mathcal{S}_k}\mu_i, \quad k = 1, 2, \ldots, r, \quad (11c)$$

where $r$ is the number of $\mathcal{S}_k$ sets which is at most $|\mathcal{S} \setminus \mathcal{S}^0 \cup \mathcal{S}_0|$. The optimal probabilities $\nu^*(\cdot)$ are obtained iteratively as a function of the TV parameter $R \in [0, 2]$, and we assume that $\forall i \in \mathcal{S}^0$, $\nu_i^* = \nu^*(\mathcal{S}^0)$, $\forall i \in \mathcal{S}_0$, $\nu_i^* = \nu^*(\mathcal{S}_0)$ and $\forall i \in \mathcal{S}_k$, $\nu_i^* = \nu^*(\mathcal{S}_k)$. In particular, for $R = 0$ (initialization step) we have that $\nu_i^* = \mu_i$, $\forall i \in \mathcal{S}$, and hence the identification of the support sets of $S$ is obtained based on the values of $\mu_i$, $\forall i \in \mathcal{S}$, that is, using the fact that $l_i = -\log\nu_i^* = -\log\mu_i$, $\forall i \in \mathcal{S}$. As $R$ increases we evaluate $\nu_i^*$, $\forall i \in \mathcal{S}$, given by (11) based on the identified support sets of the initialization step until a merge occurs (i.e., see Figure 4). In this case, a new identification of the support sets is required, and the procedure is repeated until the optimal probabilities become equal to $\nu^*(\cdot) = 1/|\mathcal{S}|$, where thereafter the solution is constant.

**Remark 3.1.** *By the relation $l_i = -\log\nu_i^*$, $\forall i \in \mathcal{S}$, elements $i, j \in \mathcal{S}$ belong to the same support set only if $l_i = l_j \Leftrightarrow \nu_i^* = \nu_j^*$.*

The optimal probabilities (11a)-(11c) can be expressed in matrix form by

$$\nu^* = \mu Q = \mu P Q \quad (12)$$

where the dimensions of $Q$ matrix depends on the value of TV parameter $R$. In [40] an algorithm is provided for constructing the desired $Q$ matrix. Intuitively, as $R$ increases, the maximizing distribution $\nu^*$ exhibits a water-filling solution, with the property the states of $\mu \in \mathbb{P}(\mathcal{S})$ are aggregated together to form a new partition of $\mathcal{S}$. However, due to the water-filling behavior of the solution, the maximizing distribution $\nu^* \in \mathbb{P}(\mathcal{S})$ is not the invariant distribution of a Markov process. Indeed, the process associated with distribution $\nu^*$ is a hidden Markov process. Consider a process $\{Y_t : t = 0, 1, \ldots\}$ taking values in $\bar{\mathcal{S}} \subseteq \mathcal{S}$. Define

$$\mathbb{P}(Y(t) = j) \triangleq \sum_{j\in\bar{\mathcal{S}}} \mathbb{P}(Y(t) = j | X(t) = i)\mathbb{P}(X(t) = i). \quad (13)$$

By denoting $\mu(t) \triangleq \mathbb{P}(X(t) = i)$ and $\nu(t) \triangleq P(Y(t) = j)$ we have

$$\nu(t+1) = \mu(t+1)Q = \mu(t)PQ = \mu(0)P^tQ, \quad (14)$$

where the resulting stochastic matrix $PQ$ gives the probability $\mathbb{P}(Y(t+1) = j)$ of the process $\{Y_t, t = 0, 1, \ldots\}$, given the state of the distribution of the Markov process $\{X_t, t = 0, 1, \ldots\}$ at time $t$. The dimension of the mapping $PQ$, which relates the process $\{Y_t, t = 0, 1, \ldots\}$ to the Markov process $\{X_t, t = 0, 1, \ldots\}$, depends on the value of the parameter $R$. As a result, once the mapping $PQ$ is computed the state of the approximated process can be computed by (14). This can be useful when we want to observe the evolution of a reduced process instead of the original Markov process.

### *3.4. Approximation of a hidden Markov process by a Markov process*

A futher approximation of the resulting finite state hidden Markov process $\{Y_t : t = 0, 1, \ldots\}$ by a finite state

[2] Equation (10) follows from compactness and convexity of the constraints and convexity of $\sum l_i\nu_i$ for a fixed $\nu \in \mathbb{B}_R(\mu)$ and concavity for a fixed $l \in \mathbb{R}_+^{|\mathcal{S}|}$ (and continuity).

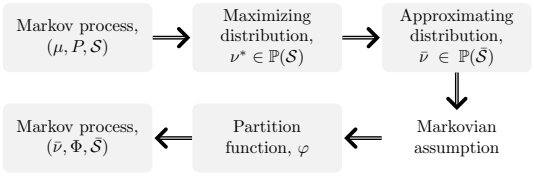



Figure 1: Approximation procedure.

Markov process of reduced order is provided in [40], by minimizing the Kullback-Leibler (KL) divergence. In summary, the approximation procedure consists of the following steps (see Figure 1): Given a finite state Markov process $(\mu, P, \mathcal{S})$, a lower dimensional distribution $\bar{\nu} \in \mathbb{P}(\bar{\mathcal{S}})$ with corresponding process $\{Y_t : t = 0, 1, \ldots\}$ and state space $\bar{\mathcal{S}} \subseteq \mathcal{S}$, is obtained from the maximizing distribution $\nu^* \in \mathbb{P}(\mathcal{S})$ by defining a partition function as follows.

**Definition 3.2.** *Let $\mathcal{S}$ and $\bar{\mathcal{S}}$ be two finite dimensional state-spaces with $|\bar{\mathcal{S}}| < |\mathcal{S}|$. Define a surjective (partition) function $\varphi : \mathcal{S} \mapsto \bar{\mathcal{S}}$ as follows:*

1. *For all $i, j \in \mathcal{S}^0$, $\varphi(i) = \varphi(j) = \ell \in \bar{\mathcal{S}}$ and $\bar{\nu}_\ell = |\mathcal{S}^0|\nu^*(\mathcal{S}^0)$.*
2. *For all $i, j \in \mathcal{S}_0$, $\varphi(i) = \varphi(j) = m \in \bar{\mathcal{S}}$ and $\bar{\nu}_m = |\mathcal{S}_0|\nu^*(\mathcal{S}_0)$.*
3. *For all $i, j \in \mathcal{S}_k$, $\varphi(i) = \varphi(j) = n_k \in \bar{\mathcal{S}}$ and $\bar{\nu}_{n_k} = |\mathcal{S}_k|\nu^*(\mathcal{S}_k)$, $k = 1, 2, \ldots, |\mathcal{S} \setminus \mathcal{S}^0 \cup \mathcal{S}_0|$.*

*where the values assigned to indices $\ell, m$ and $n_k$, for all $k = 1, 2, \ldots, |\mathcal{S} \setminus \mathcal{S}^0 \cup \boldsymbol{S}_0|$, are selected from $\bar{\mathcal{S}}$ so that $\bar{\nu}$. form an ascending order, i.e., if $\bar{\nu}_\ell < \bar{\nu}_m < \bar{\nu}_{n_1} < \cdots < \bar{\nu}_{n_k}$, then $\ell < m < n_1 < \cdots < n_k$.*

Then under the restriction that the lower dimensional process is also a finite state Markov process $(\bar{\nu}, \Phi, \bar{\mathcal{S}})$, and by utilizing the partition function $\varphi$, a transition probability matrix $\Phi$ is found which minimizes the KL divergence rate between $P$ and $\hat{\Phi}$

$$D^{\varphi}(P||\Phi) = \sum_{i,j \in \mathcal{S}} \mu_i P_{ij} \log\Big(\frac{P_{ij}}{\hat{\Phi}_{ij}}\Big). \tag{15}$$

The lifted version of the lower dimensional MC $\Phi$, denoted by $\hat{\Phi}$, is defined by (i.e., see [41])

$$\hat{\Phi}_{ij} = \frac{\mu_j}{\sum_{k \in \psi(j)} \mu_k} \Phi_{\varphi(i)\varphi(j)}, \quad i, j \in \mathcal{S}$$

where $\psi(j)$ denotes the set of elements belonging to the same set as the $jth$ element. The solution of $\Phi$ is given by

$$\Phi_{kl} = \frac{u^{(k)} \Pi P u^{(\ell)'}}{\bar{\nu}_k}, \quad k, \ell \in \bar{\mathcal{S}} \tag{16}$$

where $\Pi = diag(\nu^*)$, $u^{(k)'}$ is the transpose of $u^{(k)}$, and $u^{(k)}$ is a $1 \times |\mathcal{S}|$ row vector defined by

$$u_i^{(k)} = \begin{cases} 1 & \text{if } \varphi(i) = k \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

Next, we provide an example in order to explain in detail the approximation procedure.

**Example 3.3.** *Let $(\mu, P, \mathcal{S})$ denote a finite state Markov process, with $|\mathcal{S}| = 5$, and $\mu \triangleq [\mu_1, \mu_2, \mu_3, \mu_4, \mu_5] = [0.05, 0.1, 0.15, 0.3, 0.4]$. For simplicity of presentation let us assume that the optimal probabilities $\nu_i^*$, $i \in S$, as a function of $R$ are as shown in Figure 4(a).*
*Initialization step. For $R = 0$, we have that, $v_i^* = \mu_i$, $\forall i \in \mathcal{S}$, and by the relation $\ell_i = -\log \nu_i^*$, $\forall i \in \mathcal{S}$, the support sets of $\mathcal{S}$ are given by $\mathcal{S}^0 = \{1\}$, $\mathcal{S}_0 = \{5\}$, $\mathcal{S}_1 = \{4\}$, $\mathcal{S}_2 = \{3\}$ and $\mathcal{S}_3 = \{2\}$.*

*For the sake of this example, we choose $R = 0.2$, and we proceed with the identification of the new support sets. In particular, by Figure 4(a) (at $R = 0.2$), $\nu_1^* = \nu_2^* = 0.125$, $\nu_3^* = \mu_3 = 0.15$ and $\nu_4^* = \nu_5^* = 0.3$. By the relation $\ell_i = -\log \nu_i^*$, $\forall i \in \mathcal{S}$, we have that $l_1 = l_2 > l_3 > l_4 = l_5$, and hence, the new support sets are given by $\mathcal{S}^0 = \{1, 2\}$, $\mathcal{S}_0 = \{4, 5\}$ and $\mathcal{S}_1 = \{3\}$. By (11), the optimal probabilities are*

$$\nu^*(\mathcal{S}^0) = \min(0.2, \frac{0.15 + 0.1}{2}) = 0.125$$
$$\nu^*(\mathcal{S}_0) = \max(0.2, \frac{0.7 - 0.1}{2}) = 0.3, \quad \nu^*(\mathcal{S}_1) = 0.15.$$

*By Definition 3.2, (a) for elements $\{1, 2\} \in \mathcal{S}^0$, $\varphi(1) = \varphi(2) = 2 \in \bar{\mathcal{S}}$ and $\bar{\nu}_2 = |\mathcal{S}^0|\nu^*(\mathcal{S}^0) = 0.25$, (b) for elements $\{4, 5\} \in \mathcal{S}_0$, $\varphi(4) = \varphi(5) = 3 \in \bar{\mathcal{S}}$ and $\bar{\nu}_3 = 0.6$, (c) for element $\{3\} \in \mathcal{S}_1$, $\varphi(3) = 1 \in \bar{\mathcal{S}}$ and $\bar{\nu}_1 = |\mathcal{S}_1|\nu^*(\mathcal{S}_1) = 0.15$. The transition probability matrix $\Phi$ of the lower dimension Markov process $(\bar{\nu}, \Phi, \bar{\mathcal{S}})$ is given by*

$$\Phi = \begin{bmatrix} P_{33} & P_{31}+P_{32} & P_{34}+P_{35} \\ .5(P_{13}+P_{23}) & .5(P_{11}+P_{12}+P_{21}+P_{22}) & .5(P_{14}+P_{15}+P_{24}+P_{25}) \\ .5(P_{43}+P_{53}) & .5(P_{41}+P_{42}+P_{51}+P_{52}) & .5(P_{44}+P_{45}+P_{54}+P_{55}) \end{bmatrix}.$$

In the next section we apply the approximation results to solve Problem P1.

## 4. Optimal Control Policy

### 4.1. Equilibrium Control Policy

In this section, we show that applying the solution of (9) to the Problem P1 yields the optimal control policy. Namely, the control policy endows a invariant distribution that yields higher probability at the states with low cost, and lower probability at the states with high cost.

**Definition 4.1.** *A control policy $\bar{\pi} = \{\phi_1, ..., \phi_i, ..., \phi_{|\bar{\mathcal{S}}|}\}$ is an equilibrium control policy (ECP) if it yields a pair of vectors $\mu^*$ and $l^*$ in the following form*

$$l^*(1, \phi_1) \leq \ldots \leq l^*(i, \phi_i) \leq \ldots \leq l^*(\bar{\mathcal{S}}, \phi_{\bar{\mathcal{S}}}) \tag{18}$$
$$\mu_1^* \geq \ldots \geq \mu_i^* \geq \ldots \geq \mu_{|\bar{\mathcal{S}}|}^*, \quad \forall i \in \bar{\mathcal{S}}. \tag{19}$$

Thus if an ECP exists, it yields higher probability to the states with low cost and lower probability to the states with the high cost.

**Proposition 4.2.** *[23] An ECP $\bar{\pi}$ for the average cost $J$ exists, if for all control policies, $J$ is convex with respect to the cost function $l$ and its epigraph is closed for each $\mu$.*

**Theorem 4.3.** *[23] The ECP $\bar{\pi}$ is an optimal control policy, namely*

$$J(\bar{\pi}) = J^*(\pi^*) = \mu^*(\bar{\pi}) \cdot l^*(\bar{\pi}). \tag{20}$$

*4.2. Optimality Equation Subject to a Distance Uncertainty*

We now present the main results of this paper.

**Theorem 4.4.** *The ECP is an optimal control policy in Problem P1.*

*Proof.* Let the invariant distribution within the TV distance be[3] $\bar{\nu} = \mu + \xi$. We set $\mathbb{M}(\pi) = \Phi(\pi) - \mathbb{I}$, where $\Phi$ is the transition probability matrix and $\mathbb{I}$ is the identity matrix. Let

$$\mathbf{1} \cdot \psi(\pi) = l(\pi) + \mathbb{M}(\pi) \cdot q, \qquad \forall \pi \in \Pi, \tag{21}$$

where $\mathbf{1} = \big(1, 1, ..., 1\big)^T$, $\psi(\pi) \in \mathbb{R}$, and $q \in \mathbb{R}^{|\bar{\mathcal{S}}|}$ such that $\mathbb{M}(\pi) \cdot q > 0$. Multiplying the above equation by $\bar{\nu}(\pi)$ from the left we have

$$\psi(\pi) = \bar{\nu}(\pi) \cdot l(\pi) + \bar{\nu}(\pi) \cdot \mathbb{M}(\pi) \cdot q \tag{22}$$
$$= \bar{\nu}(\pi) \cdot l(\pi) + \bar{\nu}(\pi) \cdot \big(\Phi(\pi) - \mathbb{I}\big) \cdot q \tag{23}$$
$$= \bar{\nu}(\pi) \cdot l(\pi) + \bar{\nu}(\pi) \cdot \Phi(\pi) \cdot q - \bar{\nu}(\pi) \cdot q \tag{24}$$
$$= \bar{\nu}(\pi) \cdot l(\pi) + \bar{\nu}(\pi) \cdot q - \bar{\nu}(\pi) \cdot q = \bar{\nu}(\pi) \cdot l(\pi) \tag{25}$$

since $\mathbb{M}(\pi) = \Phi(\pi) - \mathbb{I}$ and $\bar{\nu}(\pi) = \bar{\nu}(\pi) \cdot \Phi(\pi)$. Hence, $\psi(\pi)$ is the long-run expected average cost corresponding to the control policy $\pi$.

From the Definition 4.1 of the ECP, $\bar{\pi}$,

$$l^*(\bar{\pi})(i, \phi_i) \leq l(\pi)(i, \phi_i), \ \forall i \in \bar{\mathcal{S}}, \forall \pi \in \Pi, \tag{26}$$

and since $\mathbb{M}(\pi) \cdot q > 0$, (26) can be written in matrix form

$$l^*(\bar{\pi}) \leq l(\pi) + \mathbb{M}(\pi) \cdot q = \mathbf{1} \cdot \psi(\pi), \tag{27}$$

where $\psi(\pi)$ is the long-run expected average cost corresponding to any control policy $\pi \in \Pi$. Multiplying (27) by $\bar{\nu}^*(\bar{\pi})$ from the left we have

$$\psi(\bar{\pi}) = \bar{\nu}^*(\bar{\pi}) \cdot l^*(\bar{\pi}) \leq \psi(\pi), \qquad \forall \pi \in \Pi. \tag{28}$$

Thus the ECP is the optimal control policy that minimizes the long-run expected average cost. □

The ECP provides the optimal solution of the average cost problem when there is an uncertainty regarding the invariant distribution but we know that it belongs to a given set. The ECP can be seen as a solution concept. 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. For instance, we can design a controller with the aim to achieve a higher probability for the states with low cost and lower probability for the states with high cost.

The next result shows that the ECP yields the "true" optimal control policy corresponding to the problem when there is no uncertainty about the invariant distribution.

[3] $\xi$ is a finite signed measure which integrates to zero.

**Theorem 4.5.** *The ECP in Problem P1 corresponds to the optimal control policy in Problem P0.*

*Proof.* Let $\mu$ be the "true" invariant distribution of the MC. Let $\xi$ be the variation distance of uncertainty. Thus we have $\bar{\nu} = \mu + \xi = \mu + \xi^+ - \xi^-$. From Theorem 4.3, the ECP $\bar{\pi}$ in Problem P0 is optimal

$$J^*(\bar{\pi}) = \mu^*(\bar{\pi}) \cdot l^*(\bar{\pi}). \tag{29}$$

For the Problem P1, the average cost corresponding to the ECP, denoted $\pi'$, is

$$J(\pi') = \bar{\nu}^*(\pi') \cdot l^*(\pi'). \tag{30}$$

Since[4], $\sup \bar{\nu} = \sup \mu + \sup\{\xi^+ - \xi^-\} \Rightarrow \bar{\nu}^* = \mu^* + |\xi|$, and since, $l^*(\bar{\pi}) = l^*(\pi')$ and $\mu^*(\bar{\pi}) = \mu^*(\pi')$ we have

$$J(\pi') = (\mu^*(\pi') + |\xi|) \cdot l^*(\pi') \tag{31}$$
$$= \mu^*(\bar{\pi}) \cdot l^*(\bar{\pi}) + |\xi| \cdot l^*(\bar{\pi}). \tag{32}$$

Thus the ECP $\pi'$ for the Problem P1 corresponds to the optimal control policy $\bar{\pi}$ in Problem P0. □

In other words, even if we are not certain about the invariant distribution of the MC but we know that it is within a given set, then the ECP guarantees the optimal solution.

## 5. Application: Variable Length Lossless Coding for a Total Variation Distance Class

In this section we present an application to demonstrate that the equilibrium policy is an optimal control policy which minimizes the average cost. In particular, we consider the problem of finding uniquely decodable codes, which minimize the average code-word length, also known as universal coding problem. The problem is investigated under two possible scenarios. In particular, in Section 5.0.1 we study the variable length lossless coding problem without uncertainty, while in Section 5.0.2 we employ the approximation results of Section 3.4 and we solve the lossless coding problem under TV distance uncertainty on a reduced dimensional state space.

A source generates five different symbols $\{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5\}$ with invariant distributions endowed by the policies $\pi_1$ and $\pi_2$, given by

$$\mu(\pi_1) = [0.05, 0.10, 0.15, 0.30, 0.40]$$
$$\mu(\pi_2) = [0.075, 0.15, 0.175, 0.25, 0.35].$$

Given the TV distance parameter $R \in [0, 2]$, we define the average codeword length pay-off with respect to the true source invariant distributions endowed by the policies $\pi_1$ and $\pi_2$, denoted by $\nu^*(\pi_1)$ and $\nu^*(\pi_2)$, respectively. The objective is to find a prefix code length vector $l^* \in \mathbb{R}_+^{|\mathcal{S}|}$, satisfying Kraft inequality $\sum_{i \in \mathcal{S}} D^{-l_i}$, which minimizes the maximum average codeword length pay-off (9).

[4] $\xi^+$ and $\xi^-$ is the positive and negative variation of the finite signed measure $\xi$ defined by $\xi^+ = \max\{\xi, 0\}$ and $\xi^- = \max\{-\xi, 0\}$, respectively.

*5.0.1. Lossless Coding Without Uncertainty*

Clearly, for $R = 0$, $\nu_i^*(\pi) = \mu_i(\pi)$, $\forall i \in \mathcal{S}$. The solution of coding problem for $R = 0$ under control policies $\pi_1$ and $\pi_2$ is as shown in Figures 2 and 3, respectively. The code length vector $l^*(\pi)$ under each control policy is given by

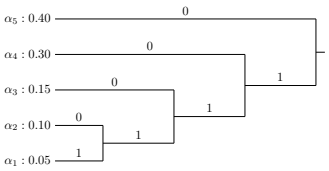
$$l^*(\pi_1) = [l_1^*, l_2^*, l_3^*, l_4^*, l_5^*] = [4, 4, 3, 2, 1]^T \quad (33)$$

$$l^*(\pi_2) = [l_1^*, l_2^*, l_3^*, l_4^*, l_5^*] = [3, 3, 2, 2, 2]^T. \quad (34)$$

The average cost is $J(\pi_1) = \nu^*(\pi_1)l^*(\pi_1) = 2.05$ and $J(\pi_2) = \nu^*(\pi_2)l^*(\pi_2) = 2.225$, and the optimal control policy can be derived by

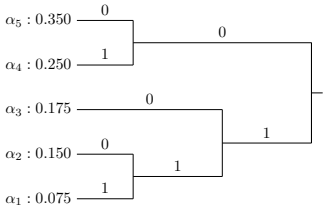
$$J^* = \inf\{J(\pi_1), J(\pi_2) | \pi_1, \pi_2 \in \Pi\} = J(\pi_1). \quad (35)$$

Hence, the control policy $\pi_1$ is the optimal control policy.




| Symbol | Code |
|---|---|
| $a_5$ | 0 |
| $a_4$ | 10 |
| $a_3$ | 110 |
| $a_2$ | 1110 |
| $a_1$ | 1111 |

Figure 2: Solution of coding problem for $R = 0$ under policy $\pi_1$.




| Symbol | Code |
|---|---|
| $a_5$ | 00 |
| $a_4$ | 01 |
| $a_3$ | 10 |
| $a_2$ | 110 |
| $a_1$ | 111 |

Figure 3: Solution of coding problem for $R = 0$ under policy $\pi_2$.

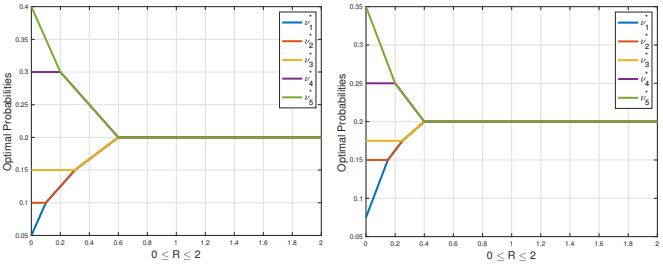
*Equilibrium Control Policy:* To demonstrate the ECP in this problem we compute $\sup_{\nu^*} \inf_{l^*} \nu^*(\pi) \cdot l^*(\pi)$ among policies $\pi_1$ and $\pi_2$.

$$\sup_{\nu^*} \inf_{l^*} \nu^*(\pi) \cdot l^*(\pi) = \sup_{\nu^*} \inf_{l^*} \begin{pmatrix} \nu^*(\pi_1) \cdot l^*(\pi_1) \\ \nu^*(\pi_2) \cdot l^*(\pi_2) \end{pmatrix}$$
$$= \sup_{\nu^*} \begin{pmatrix} [0.05, 0.1, 0.15, 0.3, 0.4] \\ [.075, .15, .175, .25, .35] \end{pmatrix} \cdot \inf_{l^*} \begin{pmatrix} [4, 4, 3, 2, 1]^T \\ [3, 3, 2, 2, 2]^T \end{pmatrix}$$
$$= [0.05, 0.1, 0.15, 0.3, 0.4] \cdot [4, 4, 3, 2, 1]^T = 2.05.$$

Thus the control policy $\pi_1$ is the ECP since it yields the maximum probability distribution to the states with low cost and lower probability distribution to the states with high cost as indicated by (18) and (19).

*5.0.2. Lossless Coding with TV Distance Uncertainty on a Reduced State Space*

Next we consider the case where the invariant distribution is within a TV distance uncertainty, and we apply the results of Section 3.4 to approximate the finite state hidden




(a) Control policy $\pi_1$ (b) Control policy $\pi_2$

Figure 4: Water-filling behavior of optimal probabilities.

| Symbol | Code |
|---|---|
| $a_3$ | 0 |
| $a_2$ | 10 |
| $a_1$ | 11 |

| Symbol | Code |
|---|---|
| $a_3$ | 0 |
| $a_2$ | 10 |
| $a_1$ | 11 |

Figure 5: Solution of coding problem for $R = 0.2$ under policy $\pi_1$ (left table) and policy $\pi_2$ (right table).

Markov process associated with $\nu^* \in \mathbb{P}(\mathcal{S})$, by a Markov process $(\bar{\nu}, \Phi, \bar{\mathcal{S}})$, $|\bar{\mathcal{S}}| = 3$, of reduced dimension. The optimal probabilities $\nu_i^*$, $\forall i \in \mathcal{S}$, under control policies $\pi_1$ and $\pi_2$, are obtained by applying (11), as shown in Figure 4. We select the TV distance parameter, $R = 0.2$. By the approximation results of Example 3.3, corresponding to control policy $\pi_1$, the invariant distribution $\bar{\nu} \in \mathbb{P}(\bar{\mathcal{S}})$ of the lower dimensional Markov process $(\bar{\nu}, \Phi, \bar{\mathcal{S}})$, $|\bar{\mathcal{S}}| = 3$, is given by $\bar{\nu}_1 = 0.15$, $\bar{\nu}_2 = 0.25$, and $\bar{\nu}_3 = 0.6$. Following the same procedure (as in Example 3.3), using the optimal probabilities obtain under control policy $\pi_2$, as depicted in Figure 4(b), we can calculate the invariant distribution of the lower dimensional Markov process corresponding to control policy $\pi_2$. In particular, the invariant distribution $\bar{\nu} \in \mathbb{P}(\bar{\mathcal{S}})$ corresponding to control policy $\pi_2$ is given by $\bar{\nu}_1 = 0.175$, $\bar{\nu}_2 = 0.325$, and $\bar{\nu}_3 = 0.5$ with $|\bar{\mathcal{S}}| = 3$.

It is not difficult to show that the solution of coding problem for $R = 0.2$ under control policies $\pi_1$ and $\pi_2$ is as shown in Figure 5. The code length vector $l^*(\pi)$ under each control policy is given by $l^*(\pi_1) = [l_1^*, l_2^*, l_3^*] = [2, 2, 1]^T$ and $l^*(\pi_2) = [l_1^*, l_2^*, l_3^*] = [2, 2, 1]^T$. The average cost is $J(\pi_1) = \nu^*(\pi_1)l^*(\pi_1) = 1.40$ and $J(\pi_2) = \nu^*(\pi_2)l^*(\pi_2) = 1.50$, and hence control policy $\pi_1$ is the optimal control policy.

*Equilibrium Control Policy:* To demonstrate the ECP in this problem we compute $\sup_{\nu^*} \inf_{l^*} \nu^*(\pi) \cdot l^*(\pi)$ among policies $\pi_1$ and $\pi_2$.

$$\sup_{\nu^*} \inf_{l^*} \nu^*(\pi) \cdot l^*(\pi) = \sup_{\nu^*} \inf_{l^*} \begin{pmatrix} \nu^*(\pi_1) \cdot l^*(\pi_1) \\ \nu^*(\pi_2) \cdot l^*(\pi_2) \end{pmatrix}$$
$$= \sup_{\nu^*} \begin{pmatrix} [0.15, 0.25, 0.6] \\ [0.175, 0.325, 0.5] \end{pmatrix} \cdot \inf_{l^*} \begin{pmatrix} [2, 2, 1]^T \\ [2, 2, 1]^T \end{pmatrix}$$
$$= [0.15, 0.25, 0.6] \cdot [2, 2, 1]^T = 1.40.$$

Hence, control policy $\pi_1$ is the ECP.

## 6. CONCLUDING REMARKS

The results presented here address the problem of minimizing the average cost per unit time in a controlled MC when the invariant distribution is within a TV distance uncertainty. We showed that the ECP is an optimal control policy that minimizes the average cost and that this solution is also optimal for the original stochastic control problem without considering uncertainty. The solution endows a invariant distribution yielding higher probability at the states with low cost and lower probability at the states with high cost.

## References


[1] E. A. Feinberg, On controlled finite state Markov processes with compact control sets, Theor. Probab. Appl. 20 (1975) 856–861.

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

[3] R. Y. Chitashvili, A controlled finite Markov chain with an arbitrary set of decisions, Theory of Probability and Its Applications 20 (4) (1975) 839–847.

[4] E. A. Feinberg, The existence of a stationary $\epsilon$ -optimal policy for a fintie Markov chain, Theory of Probability and Its Applications 23 (2) (1978) 297–313.

[5] P. Varaiya, Optimal and suboptimal stationary controls for Markov chains, IEEE Transactions on Automatic Control AC-23 (3) (1978) 388–394.

[6] D. P. Bertsekas, S. E. Shreve, Stochastic Optimal Control: The Discrete-Time Case, 1st Edition, Athena Scientific, 2007.

[7] H. J. Kushner, Introduction to Stochastic Control, Holt, Rinehart and Winston, 1971.

[8] P. R. Kumar, P. Varaiya, Stochastic systems, Prentice Hall, 1986.

[9] R. A. Howard, Dynamic Programming and Markov Processes, The MIT Press, 1960.

[10] J. L. Doob, Stochastic Processes, Wiley-Interscience, 1990.

[11] J. Bather, Optimal decision procedures for finite Markov chains. I. Examples, Advances in Applied Probability 5 (2) (1973) 328–339.

[12] E. A. Feinberg, Finite state Markov decision models with average reward criteria, Stochastic Processes and their Applications 49 (1) (1994) 159–177.

[13] A. Hordjik, Dynamic Programming and Markov Potential Theory, Mathematical Centre Tracts 51.

[14] V. S. Borkar, Controlled Markov chains and stochastic networks, SIAM Journal on Control and Optimization 21 (4) (1983) 652–665.

[15] V. S. Borkar, On minimum cost per unit time control of Markov chains, SIAM Journal on Control and Optimization 22 (6) (1984) 965–978.

[16] V. S. Borkar, Control of Markov chains with long-run average cost criterion, Stochastic Differential Systems, Stochastic Control Theory and Applications (1988) 57–77.

[17] V. S. Borkar, Control of Markov chains with long-run average cost criterion: the dynamic programming equations, SIAM Journal on Control and Optimization 27 (3) (1989) 642–657.

[18] V. S. Borkar, Controlled Markov chains with constraints, Sadhana - Academy Proceedings in Engineering Sciences 15 (pt 4-5) (1990) 405.

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

[20] R. Cavazos-Cadena, Value iteration in a class of communicating Markov decision chains with the average cost criterion, SIAM Journal on Control and Optimization 34 (6) (1996) 1848–73.

[21] A. Leizarowitz, A. J. Zaslavski, Uniqueness and stability of optimal policies of finite state Markov decision processes, Mathematics of Operations Research 32 (1) (2007) 156–167.

[22] A. A. Malikopoulos, A duality framework for stochastic optimal control of complex systems, IEEE Transactions on Automatic Control 61 (10) (2016) 2756–2765.

[23] A. A. Malikopoulos, Equilibrium Control Policies for Markov Chains, in: Proceedings of the 50th IEEE Conference on Decision and Control and European Control Conference, 2011, pp. 7093–7098.

[24] A. A. Malikopoulos, Real-Time, Self-Learning Identification and Stochastic Optimal Control of Advanced Powertrain Systems, ProQuest, 2011.

[25] A. A. Malikopoulos, Stochastic optimal control for series hybrid electric vehicles, 2013 American Control Conference (2013) 1189–1194.

[26] A. A. Malikopoulos, P. Papalambros, D. Assanis, Optimal engine calibration for individual driving styles, in: SAE Congress, 2009.

[27] A. A. Malikopoulos, P. Y. Papalambros, D. N. Assanis, Online identification and stochastic control for autonomous internal combustion engines, J. Dyn. Sys., Meas., Control 132 (2) (2010) 024504–9.

[28] A. A. Malikopoulos, Supervisory Power Management Control Algorithms for Hybrid Electric Vehicles: A Survey, IEEE Transactions on Intelligent Transportation Systems 15 (5) (2014) 1869–1885.

[29] M. Shaltout, A. A. Malikopoulos, S. Pannala, D. Chen, A consumer-oriented control framework for performance analysis in hybrid electric vehicles, IEEE Transactions on Control Systems Technology 23 (4) (2015) 1451–1464.

[30] A. A. Malikopoulos, A multiobjective optimization framework for online stochastic optimal control in hybrid electric vehicles, IEEE Transactions on Control Systems Technology 24 (2) (2016) 440–450.

[31] G. R. Grimmett, D. R. Stirzaker, Probability and Random Processes, 3rd Edition, Oxford University Press, 2001.

[32] S. M. Ross, Stochastic Processes, 2nd Edition, Wiley, 1995.

[33] F. Rezaei, C. D. Charalambous, N. Ahmed, Optimization of stochastic uncertain systems with variational norm constraints, in: 46th IEEE Conference on Decision and Control, 2007, pp. 2159–2163.

[34] C. D. Charalambous, F. Rezaei, N. Ahmed, Optimal control of uncertain stochastic systems subject to total variation distance uncertainty, in: 49th IEEE Conference on Decision and Control and European Control Conference, Atlanta, GA, 2010, pp. 1442–1447.

[35] C. D. Charalambous, I. Tzortzis, F. Rezaei, Stochastic optimal control of discrete-time systems subject to conditional distribution uncertainty, in: 50th IEEE Conference on Decision and Control and European Control Conference, Orlando, Florida, 2011, pp. 6407–6412.

[36] N. Dunford, J. Schwartz, Linear Operators: Part 1: General Theory, Interscience Publishers, Inc., New York, 1957.

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

[38] E. T. Jaynes, Information theory and statistical mechanics, Phys. Rev. 106 (1957) 620–630.

[39] J. Rissanen, Modeling by shortest data description, Automatica 14 (5) (1978) 465–471.

[40] I. Tzortzis, C. D. Charalambous, T. Charalambous, C. N. Hadjicostis, M. Johansson, Approximation of markov processes by lower dimensional processes, in: Proceedings of the 53rd IEEE Conference on Decision and Control - CDC'14, 2014, pp.4441-4446.

[41] K. Deng, P. Mehta, S. Meyn, Optimal Kullback-Leibler Aggregation via Spectral Theory of Markov Chains, IEEE Transactions on Automatic Control 56 (12) (2011) 2793–2808.