Approximation of Markov Processes by Lower Dimensional Processes

Ioannis Tzortzis, Charalambos D. Charalambous, Themistoklis Charalambous,
Christoforos N. Hadjicostis and Mikael Johansson

***Abstract*— In this paper, we investigate the problem of aggregating a given finite-state Markov process by another process with fewer states. The aggregation utilizes total variation distance as a measure of discriminating the Markov process by the aggregate process, and aims to maximize the entropy of the aggregate process invariant probability, subject to a fidelity described by the total variation distance ball. An iterative algorithm is presented to compute the invariant distribution of the aggregate process, as a function of the invariant distribution of the Markov process. It turns out that the approximation method via aggregation leads to an optimal aggregate process which is a hidden Markov process, and the optimal solution exhibits a water-filling behavior. Finally, the algorithm is applied to specific examples to illustrate the methodology and properties of the approximations.**

I. Introduction

Finite-state Markov processes are often employed to model physical phenomena in many diverse areas, such as machine learning, information theory (lossy compression), speech processing and system biology. However, in many such applications, the state space of the Markov chains are prohibitively large, for performing simulations or training of the models. One approach to overcome the large number of states is to approximate the Markov chain by a lower dimensional Markov chain, with respect to certain measures of discrepancy, or to approximate the distribution of the high dimensional Markov chain by a reduced dimensional Markov Chain. Such methods are described using relative entropy as a measure of approximation in [1]–[3] (and references therein). In these papers, the assumption often imposed is that the approximating process is also a Markov process. However, from lossy compression of Markov sources in Information Theory [4], it is already known that the approximating process subject to a fidelity of reproduction is not Markov, but it is a hidden Markov process.

In this paper, motivated by information theoretic lossy compression techniques, we propose an alternative methodology to approximate finite-state Markov processes by reduced state aggregate processes, without a priori imposing the assumption that the approximating process is also a Markov process. As a measure of fidelity or discrepancy metric between the Markov process and the aggregate process we use the total variation distance between their invariant distributions. Then, we formulate the approximation problem, we discuss the optimal solution, and we present an iterative algorithm to compute the invariant distribution of the aggregate process, including specific examples to illustrate the concepts. The formulation is based on maximizing the entropy (Jayne's maximum entropy [5]) of the invariant distribution of the lower dimensional process, subject to a fidelity criterion defined by the total variation distance metric, between the invariant distributions of the higher dimensional Markov process and that of the lower dimensional process.

This specific formulation leads to an optimal approximation algorithm described via aggregation of the states (i.e., by grouping certain states of the original Markov process) to obtain the approximating reduced state process, which is indeed a hidden Markov process. This formulation is equivalent to finding the minimum description length [6] of the aggregate process, and it is related to minimizing the average code word length of approximating the Markov process, subject to a fidelity criterion.

The main contributions of this paper are the following:

1) an iterative algorithm to compute the invariant distribution of the aggregate process;
2) extremum measures which exhibit a water-filling behavior and solve the approximation problem;
3) examples which illustrate the methodology and the properties of the approximation.

The paper is organized as follows. Section II presents the mathematical formulation as an optimization problem. Section III presents the solution of the optimization problem. Section IV presents examples illustrating the approximation methodology. Finally, Section V concludes by discussing the most important results obtained in this paper.

II. Problem Formulation

A. Description of the problem

Consider a discrete-time homogeneous Markov process $\{X_t : t = 0, 1, \ldots\}$ with state-space $\mathcal{X}$ of finite cardinality $|\mathcal{X}| = N$, and transition probability matrix P with elements $\{p_{ij} : i, j = 1, \ldots, N\}$ defined by

$$p_{ij} \overset{\triangle}{=} \mathbb{P}(X_{t+1} = j | X_t = i), \quad i, j \in \mathcal{X}, \quad t = 0, 1, \ldots. \tag{1}$$

The Markov process is assumed to be irreducible, aperiodic having a unique invariant distribution $\mu = [\mu_1 \ \mu_2 \ldots \mu_N]$ satisfying

$$\mu = \mu P. \tag{2}$$

I. Tzortzis, C. D. Charalambous and C. N. Hadjicostis are with the Department of Electrical Engineering, University of Cyprus, Nicosia, Cyprus. Emails: {tzortzis.ioannis, chadcha, chadjic}@ucy.ac.cy.

T. Charalambous and M. Johansson are with the School of Electrical Engineering, Royal Institute of Technology (KTH), Stockholm, Sweden. Emails: {themisc, mikaelj}@kth.se.

The objective is to approximate the Markov process $\{X_t : t = 0, 1, \dots, \}$ by another, not necessarily Markov process $\{Y_t : t = 0, 1, \dots, \}$ with finite-state space $\mathcal{Y} \subseteq \mathcal{X}$ of finite cardinality $|\mathcal{Y}| = M \leq |\mathcal{X}| = N$, and invariant distribution $\nu = [\nu_1 \, \nu_2 \dots \nu_M]$, with respect to an appropriate measure of proximity between the original Markov process $\{X_t : t = 0, 1, \dots, \}$ and the approximating process $\{Y_t : t = 0, 1, \dots, \}$, called the *discrepancy measure*. We consider the total variation distance between two distributions as a measure of discrepancy, and the entropy as a pay-off of grading the performance of the approximation. Depending on the nature of the approximation, one may also invoke other pay-off functionals. Note that, we do not impose any assumption on the approximating process to be a Markov process.

B. Motivation of total variation distance as a measure of discrepancy

Related work on approximating a Markov process by a lower dimensional Markov process [1]–[3], utilizes relative entropy distance (or Kullback-Leibler distance) as a measure of approximation. Below, we summarize the reasons which motivated us to employ total variation distance as a measure of approximation between the invariant distributions of the high dimensional Markov process and the lower dimensional process.

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

1) Relative Entropy distance: [7] The relative entropy of $\alpha \in \mathcal{M}_1(\Sigma)$ with respect to $\beta \in \mathcal{M}_1(\Sigma)$ is a mapping $\mathbb{D}(\alpha||\beta) : \mathcal{M}_1(\Sigma) \times \mathcal{M}_1(\Sigma) \longrightarrow [0, \infty]$ defined by

$$\mathbb{D}(\alpha||\beta) \triangleq \begin{cases} \int_\Sigma \log(\frac{\alpha(dx)}{\beta(dx)})\alpha(dx) & \text{if } \log(\frac{\alpha}{\beta}) \in L_1(\alpha) \\ & \text{and } \alpha << \beta, \\ +\infty & \text{otherwise.} \end{cases} \tag{3}$$

where $\alpha << \beta$ denotes absolute continuity of $\alpha \in \mathcal{M}_1(\Sigma)$ with respect to $\beta \in \mathcal{M}_1(\Sigma)$.[1] It is well known that $\mathbb{D}(\alpha||\beta) \geq 0, \forall \alpha, \beta \in \mathcal{M}_1(\Sigma)$, while $\mathbb{D}(\alpha||\beta) = 0 \Leftrightarrow \alpha = \beta, a.e.$.

2) Total variation distance: [8], [9] The total variation distance is a metric, $|| \cdot ||_{TV} : \mathcal{M}_1(\Sigma) \times \mathcal{M}_1(\Sigma) \to [0, \infty)$ defined by

$$||\alpha - \beta||_{TV} \triangleq \sup_{P \in \mathcal{P}(\Sigma)} \sum_{F_i \in P} |\alpha(F_i) - \beta(F_i)|, \tag{4}$$

where $\alpha, \beta \in \mathcal{M}_1(\Sigma)$ and $\mathcal{P}(\Sigma)$ denotes the collection of all finite partitions of Σ. Given a probability measure $\mu \in \mathcal{M}_1(\Sigma)$ define the fidelity set via the ball, with respect to the variation distance, centered at the measure $\mu \in \mathcal{M}_1(\Sigma)$, having radius $R \in [0, 2]$, by

$$\mathbb{B}_R(\beta) \triangleq \Big\{ \alpha \in \mathcal{M}_1(\Sigma) : ||\alpha - \beta||_{TV} \leq R \Big\}. \tag{5}$$

The two extreme cases are $R = 0$ implying $\alpha = \beta, a.e.$, and $R = 2$ implying that the support sets of α and β denoted by $\text{supp}(\alpha)$ and $\text{supp}(\beta)$, respectively, are non-overlapping, that is, $\text{supp}(\alpha) \cap \text{supp}(\beta) = \emptyset$. One of the most interesting properties of total variation distance ball is that any admissible $\nu \in \mathbb{B}_R(\mu)$ may not be absolutely continuous with respect to μ. Consequently, any approximating distribution $\nu \in \mathbb{B}_R(\mu)$ can be defined on a smaller alphabet than distribution μ, that is, $\text{supp}(\nu) \subseteq \text{supp}(\mu)$. By Pinsker's inequality, distance in total variation of probability measures is a lower bound on relative entropy distance, that is,

$$||\alpha - \beta||_{TV} \leq \sqrt{2\mathbb{D}(\alpha||\beta)}, \quad \alpha, \beta \in \mathcal{M}_1(\Sigma). \tag{6}$$

Hence, for any fixed $\beta \in \mathcal{M}_1(\Sigma)$ then $\{\alpha \in \mathcal{M}_1(\Sigma) : \mathbb{D}(\alpha||\beta) \leq r^2/2\} \subseteq \mathbb{B}_r(\beta)$.

This means that even for those measures which satisfy $\alpha << \beta$, the set described by relative entropy is a subset of the much larger total variation distance set. Moreover, by the definition of relative entropy (3), for any finite $r \in [0, \infty]$, and fixed $\beta \in \mathcal{M}_1(\Sigma)$, any set described by relative entropy consists of only those measures $\alpha \in \mathcal{M}_1(\Sigma)$ which are absolutely continuous with $\beta \in \mathcal{M}_1(\Sigma)$. This property of relative entropy rules out the possibility of measures $\alpha \in \mathcal{M}_1(\Sigma)$ and $\beta \in \mathcal{M}_1(\Sigma)$ to be defined on different state-spaces. It is one of the main disadvantages of employing relative entropy as a measure of approximation between a high dimensional Markov process and a lower dimensional process. An alternative, is to use a partition function to lift the lower dimensional process to the state-space of the original Markov process; however the drawback of this approach lies in assuming a partition function [3].

Motivated by the above issues, the approximation method proposed in this paper is based on total variation distance as a measure of discriminating the Markov process and the approximated process.

C. Aggregation based on maximum entropy principle

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

$$\mathbb{P}(\Sigma) \triangleq \Big\{ p = (p_1, \dots, p_{|\Sigma|}) : p_i \geq 0, i \in \Sigma, \sum_{i \in \Sigma} p_i = 1 \Big\}.$$

Thus, $p \in \mathbb{P}(\Sigma)$ is a probability vector in $\mathbb{R}_+^{|\Sigma|}$. Also, let $\ell \triangleq \{\ell_1, \dots, \ell_{|\Sigma|}\} \in \mathbb{R}_+^{|\Sigma|}$ (e.g., set of non-negative vectors of dimension $|\Sigma|$).

Given the invariant distribution $\mu \in \mathbb{P}(\Sigma)$ and a parameter $R \in [0, 2]$ define the average pay-off with respect to the stationary distribution $\{\nu_i : i \in \Sigma\} \in \mathbb{B}_R(\mu) \subset \mathbb{P}(\Sigma)$ by

$$\mathbb{L}(\nu) = \sum_{i \in \Sigma} \ell_i \nu_i, \quad \ell \in \mathbb{R}_+^{|\Sigma|}. \tag{7}$$

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

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

[1] If $\beta(A) = 0$ for some measurable set A then $\alpha(A) = 0$.

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 defined later and 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 [14]). Hence, Problem (8) is a convex optimization problem on the space of probability measures. The solution of (8) is obtained by choosing the parameters $\ell_i \triangleq -\log \nu_i, \forall i \in \Sigma$, hence optimization (8) becomes equivalent to the problem of finding the approximating distribution corresponding to the minimum description code word length [6].

Consider Jayne's maximum entropy principle; then, the approximation problem can be formulated as follows: maximize the entropy of $\{\nu_i : i \in \Sigma\}$ subject to total variation fidelity set, defined by

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

Problem (9) is of interest when the concept of insufficient reasoning (e.g., Jayne's maximum entropy principle[2] [5]) is applied to construct a model for $\nu \in \mathbb{P}(\Sigma)$, subject to information quantified via the fidelity set defined by the variation distance between ν and μ.

It is not difficult to show 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 [6], [15], as follows. Let $\{\ell_i : i \in \Sigma\}$ denote the positive codeword lengths corresponding to each symbol of the approximating distribution, which satisfy the Kraft inequality of lossless Shannon codes $\sum_{i \in \Sigma} D^{-\ell_i} \leq 1$, where the code word alphabet is D-ary (unless specified otherwise $\log(\cdot) \triangleq \log_D(\cdot)$). Then, by the Von-Neumann's theorem, which holds due to compactness and convexity of the constraints, we have that

$$\begin{aligned} &\min_{\ell \in \mathbb{R}_+^{|\Sigma|} : \sum_{i \in \Sigma} D^{-\ell_i} \leq 1} \; \max_{\substack{\nu \in \mathbb{B}_R(\mu) \\ \mu = \mu P}} \sum_{i \in \Sigma} \ell_i \nu_i \\ &= \max_{\substack{\nu \in \mathbb{B}_R(\mu) \\ \mu = \mu P}} \; \min_{\ell \in \mathbb{R}_+^{|\Sigma|} : \sum_{i \in \Sigma} D^{-\ell_i} \leq 1} \sum_{i \in \Sigma} \ell_i \nu_i = \max_{\substack{\nu \in \mathbb{B}_R(\mu) \\ \mu = \mu P}} H(\nu). \end{aligned}$$

Hence, for $\ell_i \triangleq -\log \nu_i, \forall i \in \Sigma$, the optimization (8) is equivalent to optimization (9).

III. Solution of the Aggregation Problem

We draw upon the results of [14] to find the solution of optimization (8), and consequently the solution of (9). First, we identify the support sets and their corresponding values.

Define the maximum and minimum values of the sequence $\{\ell_1, \ldots, \ell_{|\Sigma|}\} \in \mathbb{R}_+^{|\Sigma|}$ by $\ell_{\max} \triangleq \max_{i \in \Sigma} \ell_i$, $\ell_{\min} \triangleq \min_{i \in \Sigma} \ell_i$, and its corresponding support sets by

$$\Sigma^0 \triangleq \{i \in \Sigma : \ell_i = \ell_{\max}\}, \;\; \Sigma_0 \triangleq \{i \in \Sigma : \ell_i = \ell_{\min}\}. \tag{10}$$

For all remaining elements of the sequence, $\{\ell_i : i \in \Sigma \setminus \Sigma^0 \cup \Sigma_0\}$, define recursively the set of indices for which ℓ achieves its $(k+1)$st smallest value by Σ_k, where $k \in \{1, 2, \ldots, |\Sigma \setminus \Sigma^0 \cup \Sigma_0|\}$, till all the elements of Σ are exhausted (i.e., k is at most $|\Sigma \setminus \Sigma^0 \cup \Sigma_0|$), and the corresponding values of the sequence on the Σ_k sets by $\ell(\Sigma_k)$.

For $\ell \in \mathbb{R}_+^{\Sigma}$, and $\mu \in \mathbb{P}(\Sigma)$, it is shown in [14], that the solution of optimization (8) is given by

$$\mathbb{L}(\nu^*) = \ell_{\max} \nu^*(\Sigma^0) + \ell_{\min} \nu^*(\Sigma_0) + \sum_{k=1}^{r} \ell(\Sigma_k) \nu^*(\Sigma_k) \tag{11}$$

where r is the number of Σ_k sets which is at most $|\Sigma \setminus \Sigma^0 \cup \Sigma_0|$. Moreover, the optimal probabilities are obtained via a water-filling solution, as follows

$$\nu^*(\Sigma^0) \triangleq \sum_{i \in \Sigma^0} \nu_i^* = \sum_{i \in \Sigma^0} \mu_i + \frac{\alpha}{2}, \tag{12a}$$

$$\nu^*(\Sigma_0) \triangleq \sum_{i \in \Sigma_0} \nu_i^* = \Big(\sum_{i \in \Sigma_0} \mu_i - \frac{\alpha}{2} \Big)^+, \tag{12b}$$

$$\nu^*(\Sigma_k) \triangleq \sum_{i \in \Sigma_k} \nu_i^* = \Big(\sum_{i \in \Sigma_k} \mu_i - \Big(\frac{\alpha}{2} - \sum_{j=1}^{k} \sum_{i \in \Sigma_{j-1}} \mu_i \Big)^+ \Big)^+, \tag{12c}$$

$$\alpha = \min\left(R, R_{\max}\right), \qquad R_{\max} \triangleq 2(1 - \sum_{i \in \Sigma^0} \mu_i), \tag{12d}$$

where $k = 1, 2, \ldots, r$.

The optimal probabilities (12a)-(12c), can be expressed in matrix form by

$$\nu^* = \mu Q \tag{13}$$

where $\nu^* \triangleq [\nu_1^* \;\; \nu_2^* \;\; \ldots \;\; \nu_N^*]$ denotes the invariant distribution of the process $\{Y_t, t = 0, 1, \ldots\}$, and the dimensions of Q matrix depends on the value of total variation parameter R. In Section III-A, we provide a technique for constructing the desired Q matrix for optimization (9).

Remark III.1 *Note that this reduction can be made for a state of the distribution of the Markov process (instead of the invariant distribution), given that R is provided, i.e.,*

$$\mathbb{P}(Y(t)=j) = \sum_{i=1}^{N} \mathbb{P}(Y(t)=j | X(t)=i) \mathbb{P}(X(t)=i) \tag{14}$$

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

$$\nu(t+1) = \mu(t+1)Q = \mu(t)PQ = \mu(0)P^t Q, \tag{15}$$

where the resulting stochastic matrix PQ gives the probability $\mathbb{P}(Y(t+1) = j)$ of the hidden 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 hidden 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 approximating process can be computed by (15). This can be useful when we want to observe the evolution of a reduced set of symbols instead of the state sequence of the original Markov process.

[2] 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.

A. *The Q matrix of optimization (9)*

Here, we give an algorithm to construct the Q matrix for solving optimization (9). Before giving the algorithm, we introduce some notation.

Let r denote the number of Σ_k sets, that is, $1 \leq r \leq |\Sigma \setminus \Sigma^0 \cup \Sigma_0|$ (note that, set Σ_0 is excluded). Furthermore, let r^+ and r^- denote the number of μ_i, $i \in \Sigma$, such that $\mu_i \geq \frac{1}{|\Sigma|}$ and $\mu_i < \frac{1}{|\Sigma|}$, respectively, and in addition $\mu_i \neq \mu_j$ for all $i \neq j$, $i, j \in \Sigma$.

Remark III.2 *The initialization step of Algorithm III.3 is performed by letting $R = 0$. In this case, $\nu_i = \mu_i$, for all $i \in \Sigma$, and hence, $\ell_i \triangleq -\log \nu_i = -\log \mu_i$.*

Algorithm III.3

1) *Initialization step:*
   a) *Arrange μ_i, $i \in \Sigma$, in a descending order and let $R = 0$.*
   b) *Identify the support sets Σ^0, Σ_0 and Σ_k for all $k \in \{1, 2, \ldots, |\Sigma \setminus \Sigma^0 \cup \Sigma_0|\}$.*
   c) *Calculate the value of r, r^- and r^+.*

For any $R \in [0, 2]$:

2) *Step.1 (Indicator functions):*
   a) *For $k = 1, 2 \ldots, r^- {-} 1$ let*
   $$\mu_-^R(\Sigma_k) \triangleq \frac{\sum_{i \in \cup_{j=0}^{k-1} \Sigma_j} \mu_i - R/2}{\sum_{j=0}^{k-1} |\Sigma_j|}.$$
   Define
   $$I_-^{\Sigma_k} \triangleq \begin{cases} 1 & \text{if } \mu_-^R(\Sigma_k) \leq \frac{\sum_{i \in \Sigma_k} \mu_i}{|\Sigma_k|}, \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$
   For $k = r^-$ let
   $$\mu_-^R(\Sigma_{r^-}) \triangleq \frac{\sum_{i \in \cup_{j=0}^{r^- -1} \Sigma_j} \mu_i - R/2}{\sum_{j=0}^{r^- -1} |\Sigma_j|}.$$
   Define
   $$I_-^{\Sigma_{r^-}} \triangleq \begin{cases} 1 & \text{if } \mu_-^R(\Sigma_{r^-}) \leq \frac{1}{|\Sigma|}, \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$
   b) *For $k = 1, 2 \ldots, r^+ {-} 1$ let*
   $$\mu_+^R(\Sigma_k) \triangleq \frac{\sum_{i \in \Sigma \setminus \cup_{j=r}^{k-1} \Sigma_{r-j}} \mu_i + R/2}{|\Sigma \setminus \cup_{j=r}^{k-1} \Sigma_{r-j}|},$$
   Define
   $$I_+^{\Sigma_k} \triangleq \begin{cases} 1 & \text{if } \mu_+^R(\Sigma_k) \geq \frac{\sum_{i \in \Sigma_{r-k+1}} \mu_i}{|\Sigma_{r-k+1}|}, \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$
   For $k = r^+$ let
   $$\mu_+^R(\Sigma_{r^+}) \triangleq \frac{\sum_{i \in \Sigma \setminus \cup_{j=r}^{r^+ -1} \Sigma_{r-j}} \mu_i + \frac{R}{2}}{|\Sigma \setminus \cup_{j=r}^{r^+ -1} \Sigma_{r-j}|}.$$
   Define
   $$I_+^{\Sigma_{r^+}} \triangleq \begin{cases} 1 & \text{if } \mu_+^R(\Sigma_{r^+}) \geq \frac{1}{|\Sigma|}, \\ 0 & \text{otherwise.} \end{cases} \tag{19}$$

3) *Step.2 (The $Q^\dagger$ matrix):*
   Let $Q^\dagger$ be an $(|\Sigma|) \times (2 + r)$ matrix.
   a) *The elements of the first column are given as follows.*
      i) *For all $i \in \Sigma_0$, let the $(Q^\dagger)_{i,1}$ be equal to*
      $$\frac{1 - R/2}{|\Sigma_0| + \sum_{j=1}^{r^\downarrow -1} I_-^{\Sigma_j} |\Sigma_j|} \left(I_-^{\Sigma_{r^-}}\right)^{\mathsf{c}} + \frac{I_-^{\Sigma_{r^-}}}{|\Sigma|}. \tag{20}$$
      ii) *For all $i \in \Sigma_k$, $k = 1, 2, \ldots, r^- {-} 1$, let the $(Q^\dagger)_{i,1}$ be equal to*
      $$\frac{I_-^{\Sigma_k} - R/2}{|\Sigma_0| + \sum_{j=1}^{r^\downarrow -1} I_-^{\Sigma_j} |\Sigma_j|} \left(I_-^{\Sigma_{r^-}}\right)^{\mathsf{c}} + \frac{I_-^{\Sigma_{r^-}}}{|\Sigma|}. \tag{21}$$
      iii) *Let all the remaining elements be equal to*
      $$\frac{-R/2}{|\Sigma_0| + \sum_{j=1}^{r^\downarrow -1} I_-^{\Sigma_j} |\Sigma_j|} \left(I_-^{\Sigma_{r^-}}\right)^{\mathsf{c}} + \frac{I_-^{\Sigma_{r^-}}}{|\Sigma|}. \tag{22}$$
   b) *The elements of the last column are given by*
      i) *For all $i \in \Sigma^0$, let the $(Q^\dagger)_{i,r+2}$ be equal to*
      $$\frac{1 + R/2}{|\Sigma^0| + \sum_{j=1}^{r^\uparrow -1} I_+^{\Sigma_j} |\Sigma_{r-j+1}|} (I_+^{\Sigma_{r^+}})^{\mathsf{c}}. \tag{23}$$
      ii) *For all $i \in \Sigma_{r-k+1}$, $k = 1, 2, \ldots, r^\uparrow - 1$ let the $(Q^\dagger)_{i,r+2}$ be equal to*
      $$\frac{I_+^{\Sigma_k} + R/2}{|\Sigma^0| + \sum_{j=1}^{r^\uparrow -1} I_+^{\Sigma_j} |\Sigma_{r-j+1}|} (I_+^{\Sigma_{r^+}})^{\mathsf{c}}. \tag{24}$$
      iii) *Let all the remaining elements be equal to*
      $$\frac{R/2}{|\Sigma^0| + \sum_{j=1}^{r^\uparrow -1} I_+^{\Sigma_j} |\Sigma_{r-j+1}|} (I_+^{\Sigma_{r^+}})^{\mathsf{c}}. \tag{25}$$
   c) *The elements of all remaining columns are given by*
      i) *For all $i \in \Sigma_k$, $k = 1, 2, \ldots, r^- {-} 1$ let*
      $$(Q^\dagger)_{i,z} = \frac{(I_-^{\Sigma_k})^{\mathsf{c}}}{|\Sigma_k|}, \tag{26}$$
      where $z = 1 + k$ denotes the zth column. Let all the remaining elements of the zth column be equal to zero. However, if $I_-^{\Sigma_k} = 1$, then let all the elements of the zth column be equal with the corresponding elements of the first column, that is,
      $$(Q^\dagger)_{1,z} = (Q^\dagger)_{1,1}, (Q^\dagger)_{2,z} = (Q^\dagger)_{2,1}, \ldots, (Q^\dagger)_{|\Sigma|,z} = (Q^\dagger)_{|\Sigma|,1}. \tag{27}$$
      ii) *For all $i \in \Sigma_{r-k+1}$, $k = 1, 2, \ldots, r^+ {-} 1$ let*
      $$(Q^\dagger)_{i,z} = \frac{(I_+^{\Sigma_k})^{\mathsf{c}}}{|\Sigma_k|}, \tag{28}$$
      where $z = r + 2 - k$ denotes the zth column. Let all the remaining elements of the zth column be equal to zero. However, if $I_+^{\Sigma_k} = 1$, then let all the elements of the zth column be equal with the corresponding elements of the last column, that is,

$$(Q^{\dagger})_{1,z} = (Q^{\dagger})_{1,|\Sigma|}, (Q^{\dagger})_{2,z} = (Q^{\dagger})_{2,|\Sigma|}, \dots, \\ (Q^{\dagger})_{|\Sigma|,z} = (Q^{\dagger})_{|\Sigma|,|\Sigma|}. \quad (29)$$

4) *Step.4:*
If any of the columns of matrix $Q^{\dagger}$ are equal, then merge them by adding them. Matrix Q is defined to be the matrix after the merging of all equal columns.

IV. EXAMPLES

A. Example illustrating Algorithm III.3

Here, we provide an example in order to explain each step of Algorithm III.3, which is to be implemented for the optimal solution of approximation problems based on relative entropy.

Let $\mu = [\mu_1\ \mu_2\ \mu_3\ \mu_4]$, where $\mu_1 > \mu_2 > \mu_3 > \mu_4$, and also assume that $\mu_1 > \mu_2 > \frac{1}{|\Sigma|}$ and $\mu_4 < \mu_3 < \frac{1}{|\Sigma|}$, where $|\Sigma| = 4$. For simplicity of presentation, it is assumed that the optimum probabilities ν_i^*, $i \in \Sigma$, as a function of R are known and they are given by Fig.1.

Initialization step. For $R = 0$, and from Remark III.2, we conclude that $\ell_1 < \ell_2 < \ell_3 < \ell_4$, and therefore the support sets are equal to $\Sigma^0 = \{4\}$, $\Sigma_0 = \{1\}$, $\Sigma_1 = \{2\}$ and $\Sigma_2 = \{3\}$. The number of the Σ_k sets is equal to $r = 2$. The number of μ_i, $i \in \Sigma$, which are greater (or equal) than $\frac{1}{|\Sigma|} = 0.25$ (and also $\mu_i \neq \mu_j$, $i, j \in \Sigma$) is $r^- = 2$. Similarly, the number of μ_i which are strictly smaller than $\frac{1}{|\Sigma|} = 0.25$ (and also not equal to each other) is also $r^+ = 2$.

Step.1 From (16)-(17), the indicator functions $I_-^{\Sigma_1}$ and $I_-^{\Sigma_2}$ are given by

$$I_-^{\Sigma_1} \triangleq \begin{cases} 1 & \text{if } \mu_1 - \frac{R}{2} \leq \mu_2, \\ 0 & \text{otherwise,} \end{cases} \quad I_-^{\Sigma_2} \triangleq \begin{cases} 1 & \text{if } \frac{\mu_1+\mu_2-\frac{R}{2}}{2} \leq 0.25, \\ 0 & \text{otherwise,} \end{cases}$$

and from (18)-(19), the indicator functions $I_+^{\Sigma_1}$ and $I_+^{\Sigma_2}$ are given by

$$I_+^{\Sigma_1} \triangleq \begin{cases} 1 & \text{if } \mu_4 + \frac{R}{2} \geq \mu_3, \\ 0 & \text{otherwise,} \end{cases} \quad I_+^{\Sigma_2} \triangleq \begin{cases} 1 & \text{if } \frac{\mu_3+\mu_4+\frac{R}{2}}{2} \geq 0.25, \\ 0 & \text{otherwise.} \end{cases}$$

The behavior of the indicator functions for values of $R \in [0, 2]$ is as shown in Fig.1. For $0 \leq R < R_1$, that is, before a merge occurs, all indicator functions are equal to zero. If a merge occurs the respective indicator function becomes equal to one, until for some $R \geq R_2$, where all indicator functions are equal to one.

Step.2 Let Q be an 4×4 matrix. For $0 \leq R < R_1$,

$$Q^{\dagger} = \begin{pmatrix} 1-R/2 & 0 & 0 & R/2 \\ -R/2 & 1 & 0 & R/2 \\ -R/2 & 0 & 1 & R/2 \\ -R/2 & 0 & 0 & 1+R/2 \end{pmatrix}$$

and since no equal columns exist then $Q = Q^{\dagger}$. For $R_1 \leq R < R_2$,

$$Q^{\dagger} = \begin{pmatrix} \frac{1-R/2}{2} & \frac{1-R/2}{2} & R/4 & R/4 \\ \frac{1-R/2}{2} & \frac{1-R/2}{2} & R/4 & R/4 \\ -R/4 & -R/4 & \frac{1+R/2}{2} & \frac{1+R/2}{2} \\ -R/4 & -R/4 & \frac{1+R/2}{2} & \frac{1+R/2}{2} \end{pmatrix}.$$

and hence

$$Q = \begin{pmatrix} 1-R/2 & R/2 \\ 1-R/2 & R/2 \\ -R/2 & 1+R/2 \\ -R/2 & 1+R/2 \end{pmatrix}.$$

For $R \geq R_2$,

$$Q^{\dagger} = \begin{pmatrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{pmatrix} \Longrightarrow Q = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

Note that, the dimension of matrix Q is based on the value of total variation distance parameter R. For $0 < R \leq R_1$ its dimension is equal to $|\mathcal{X}| \times (2 + r)$. Whenever two columns become equal (that is, an indicator function is activated) they are merged, until for some $R \geq R_2$, where matrix Q is transformed into column vector of dimension $|\mathcal{X}| \times 1$. Once matrix Q is constructed, as a function of parameter R, then by (13) the solution of optimization (9) is readily available. Moreover, by Remark (III.1), the probability mass flow from the original Markov chain to the hidden Markov chain is also readily available, as we show next.

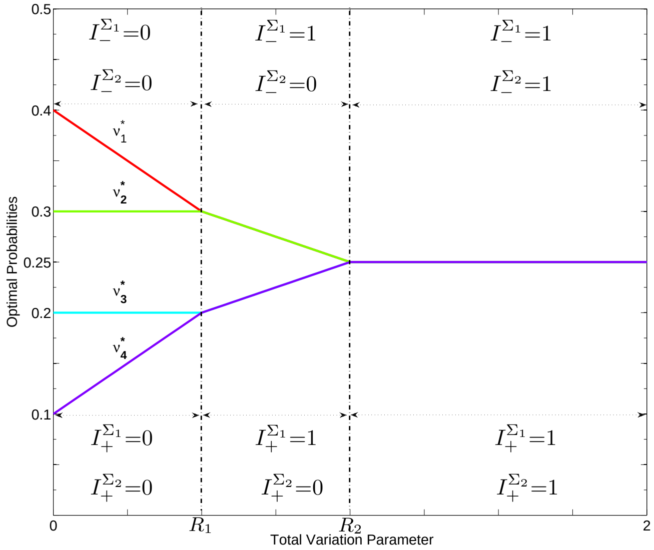
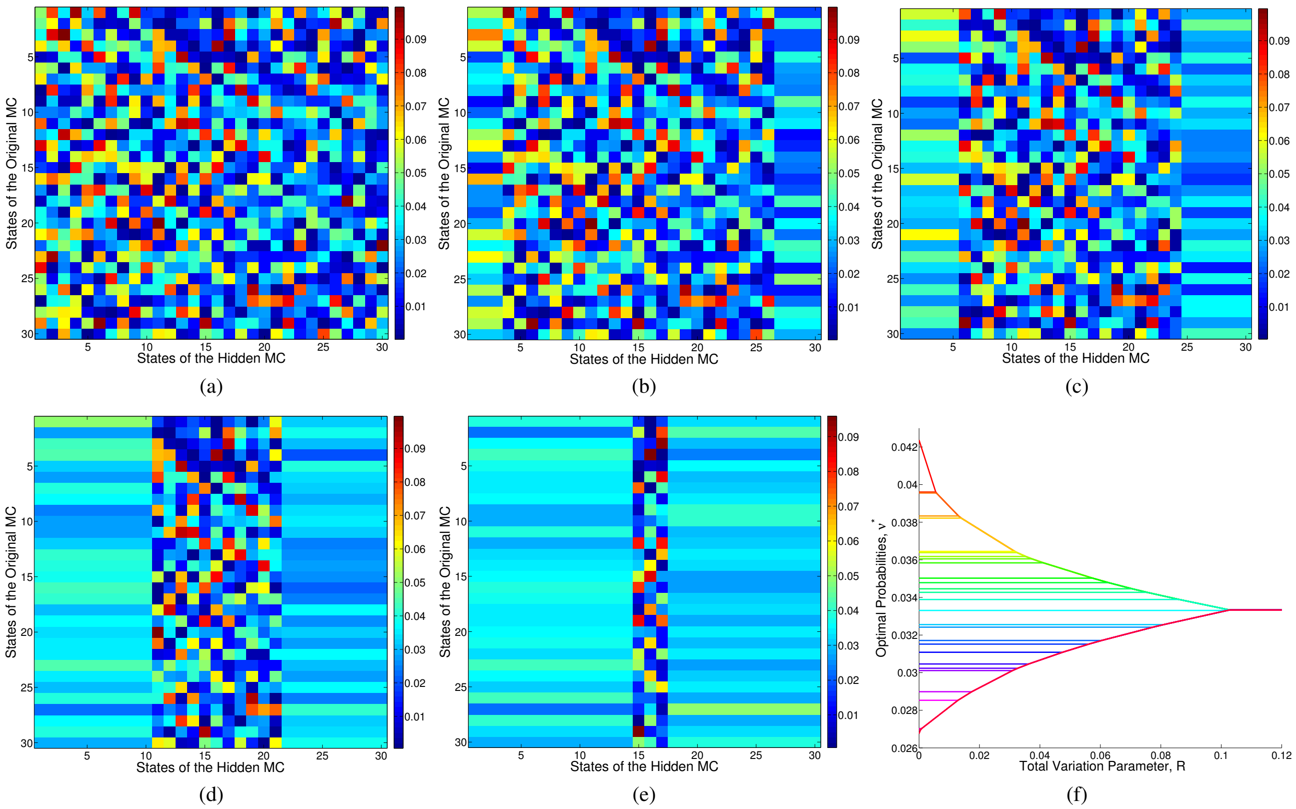


Fig. 1. Optimal Probabilities as a function of R.

B. An illustrative example of Markov chain approximation with a large number of states

Consider a 30-state Markov chain, whose transition probability matrix is given by Fig.2(a). By employing Algorithm III.3, we solve the approximation problem based on entropy principle.

Fig.2(b)-(e) depicts the probability mass flow from the original Markov chain to the hidden Markov chain for some preselected values of R, in which the color of the ith row and jth column, as indicated by the color bar, represent the $(PQ)_{i,j}$ matrix. In particular, Fig.2(b), 2(c), 2(d) and 2(e) depict a 25, 21, 13 and a 5-state approximation, respectively. Fig.2(f) depicts the water-filling behavior of the optimal



(a) (b) (c) (d) (e) (f)

Fig. 2. Approximation results based on entropy principle of a 30-state Markov chain: Plot (a) depicts the P matrix of the original Markov chain. Plots (b)-(e) depict a 25, 21, 13 and a 5-state approximation. Plot (f) depicts the optimal probabilities ν^* as a function of total variation distance.

probabilities ν^* as a function of the total variation parameter R. In summary, the solution of approximation problem based on entropy principle is described via aggregation of states, that is, by grouping certain states of the original Markov chain to obtain the approximating hidden Markov chain.

V. Conclusions

In this work, we studied the problem of aggregating a Markov process with a large number of states by another process with fewer states. The total variation distance is introduced as a discrepancy measure, and the problem is formulated by maximizing the entropy of the approximating steady state distribution, subject to a constraint on the total variation distance metric, between the steady state distribution of the original Markov process and that of the approximating process. An iterative algorithm is proposed to approximate Markov processes by another process. It shown that the solution exhibits water-filling, and that the proposal aggregation approach ensures that the resulting process is a hidden Markov process.

References

[1] M. Vidyasagar, “Reduced-order modeling of Markov and hidden Markov processes via aggregation,” in *49th IEEE Conference on Decision and Control (CDC)*, 2010, pp. 1810–1815.

[2] Y. Xu, S. Salapaka, and C. Beck, “On reduction of graphs and Markov chain models,” in *50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC)*, 2011, pp. 2317–2322.

[3] K. Deng, P. Mehta, and S. Meyn, “Optimal Kullback-Leibler aggregation via spectral theory of Markov chains,” *IEEE Trans. Autom. Control*, vol. 56, no. 12, pp. 2793–2808, Dec. 2011.

[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley-Interscience, 2006.

[5] E. T. Jaynes, “Information theory and statistical mechanics,” *Phys. Rev.*, vol. 106, no. 4, pp. 620–630, May 1957.

[6] A. Barron, J. Rissanen, and B. Yu, “The minimum description length principle in coding and modeling,” *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2743–2760, Oct. 1998.

[7] P. Dupuis and R. S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*. New York: John Wiley & Sons, Inc., 1997.

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

[9] M. Vidyasagar, “A metric between probability distributions on finite sets of different cardinalities and applications to order reduction,” *IEEE Trans. Autom. Control*, vol. 57, no. 10, pp. 2464–2477, Oct. 2012.

[10] A. L. Gibbs and F. E. Su, “On choosing and bounding probability metrics,” *Internat. Statist. Rev*, vol. 70, no. 3, pp. 419–435, Dec. 2002.

[11] M. Pinsker, “Mathematical foundations of the theory of optimum coding of information,” *Itogi Nauki. Ser. Mat. Anal. Teor. Ver. Regulir. 1962*, pp. 197–210, 1964.

[12] I. Csiszár, “Information-type measures of difference of probability distributions and indirect observations,” *Studia Sci. Math. Hungar.*, vol. 2, pp. 299–318, 1967.

[13] J. H. B. Kemperman, *On the Optimum Rate of Transmitting Information*, ser. Lecture Notes in Mathematics. Springer-Verlag, 1969, pp. 126–169.

[14] C. D. Charalambous, I. Tzortzis, S. Loyka, and T. Charalambous, “Extremum problems with total variation distance and their applications,” *IEEE Trans. Autom. Control*, vol. 59, no. 9, pp. 2353–2368, Sept. 2014.

[15] J. Rissanen, “Modeling by shortest data description,” *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.