Lumping Markov Chains with Silent Steps

Jasen Markovski and Nikola Trčka

Department of Mathematics and Computer Science
Technische Universiteit Eindhoven
P.O. Box 513, NL-5600 MB Eindhoven, The Netherlands
{j.markovski,n.trcka}@tue.nl

***Abstract*— A silent step in a dynamic system is a step that is considered unobservable and that can be eliminated. We define a Markov chain with silent steps as a class of Markov chains parameterized with a special real number τ. When τ goes to infinity silent steps become immediate, i.e. timeless, and therefore unobservable. To facilitate the elimination of these steps while preserving performance measures, we introduce a notion of lumping for the new setting. To justify the lumping we first extend the standard notion of ordinary lumping to the setting of discontinuous Markov chains, processes that can do infinitely many transitions in finite time. Then, we give a direct connection between the two lumpings for the case when τ is infinite. The results of this paper can serve as a correctness criterion and a method for the elimination of silent (τ) steps in Markovian process algebras.**

I. Introduction

Markov chains (see e.g. [1], [2]) have established themselves as very powerful, yet fairly simple, models for performance analysis. There exists a well-developed and vast mathematical theory to support these models. Efficient methods have been found to deal with Markov chains with millions of states. They all facilitate performance evaluation using different schemes to save storage space and enable faster calculations. However, although alleviated, the state space explosion problem is not completely resolved and many real world problems still cannot be feasibly solved.

One of the most important optimization techniques for the reduction of the complexity of Markov chains is called *lumping* [3], [4]. Lumping is a method based on the aggregation of states that exhibit the same behavior. It produces a smaller Markov chain that retains the same performance characteristics as the original one.

Over the past few years several stochastic process algebras have been developed in order to allow for a compositional modeling of both qualitative and quantitative aspects of systems (for an overview see [5], [6]). Although some of these algebras incorporate generally distributed stochastic delays (e.g. [7], [8]), the most widely used are the ones that restrict to exponential distributions (e.g. [9], [10]) due to the memoryless property. Typically, the employed model is some kind of extension of Markov chains with action labels. When a system is modeled, all action information is discarded and the system is reduced by lumping. Then, on the resulting Markov chain, analysis is performed by standard techniques.

Jasen Markovski is supported by Bsik project BRICKS AFM 3.2
Nikola Trčka is supported by the Netherlands Organization for Scientific Research (NWO) project 612.064.205

For the stochastic process algebra IMC (stands for *Interactive Markov chain*) [9], the extension of Markov chains with actions is orthogonal, i.e. actions and stochastic delays are not combined, but interleaved (see Fig. 1a). The elimination of action information from the model is done together with its aggregation; all actions are first renamed into silent steps and then the model is minimized using a suitably extended notion of weak bisimulation. This bisimulation treats interaction between (exponentially) delayable transitions the same way as ordinary lumpability does, but the interaction of delayable and silent steps is based on the intuitive fact that silent steps are timeless and therefore always have priority over delayable ones.

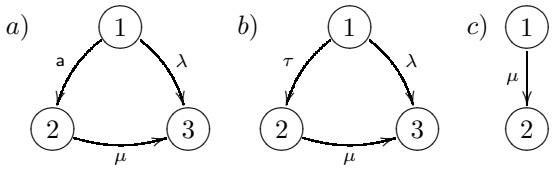


Fig. 1. An IMC, its corresponding Markov chain with silent steps and the induced Markov chain

To give an example, consider the IMC depicted in Fig. 1a. If this model is considered closed, i.e. if it does not interact with the environment, the action a can be renamed into the silent step τ and, what we call a Markov chain with silent steps is obtained (Fig. 1b). Now, assume that the process starts from state 1. The transition from state 1 to state 3 takes time distributed according to the exponential distribution of rate λ. However, as the transition from state 1 to state 2 is determined by a silent step τ; it does not take any time, and so, due to the race-condition policy, it must be taken as soon as the process enters state 1. Thus, the process in state 1 does not actually have a choice and always takes the left transition, entering state 2. From state 2, there is only one possibility, to enter state 3 after an exponential delay of rate μ. The execution of the silent step cannot be observed and one sees only the transition from state 2 to state 3. Therefore, according to the intuition, the process in Fig. 1b is performance-equivalent to the one in Fig. 1c.

Next, observe the process in Fig. 2a. In state 1 this process

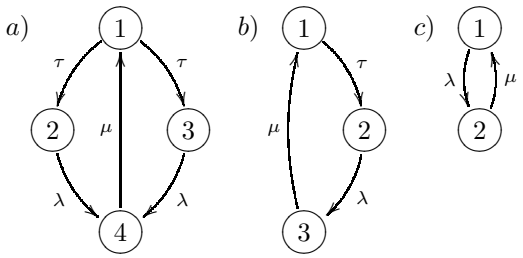
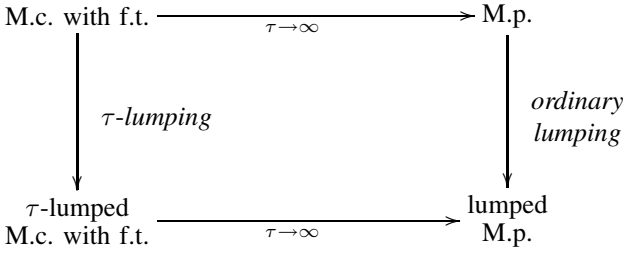


Fig. 2. Three equivalent Markov chains with silent steps

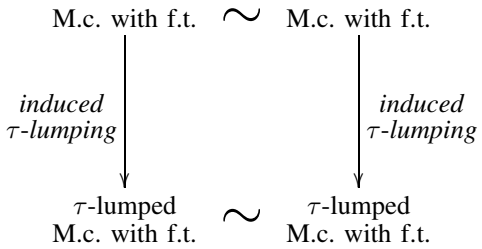
exhibits classical non-determinism, i.e. the probability of executing the left (right) transition is not determined. However, if we observe the behavior of the states 2 and 3, we easily notice that it is the same. More precisely, no matter which transition is taken from state 1, after performing a silent step and then delaying exponentially with rate λ, the process enters state 4. This suggests that the process in Fig. 2a is equivalent to the ones in Fig. 2b and Fig. 2c.

The main goal of this paper is to give a mathematical underpinning for the elimination of silent steps. We propose a new approach to reduction of Markov chains with silent steps. We treat them as more general Markov chains and extend the notion of lumping to the new setting. The lumping is shown to correspond to the above intuition. Moreover, staying in the domain of stochastic processes, the performance properties of Markov chains with silent steps are automatically defined and, therefore, we can speak of the correctness of the reductions. The approach goes in two steps.

First we extend the standard Markov chain model by assuming that some transitions are parameterized with a special (large) real number τ and call the notion a Markov chain with fast transitions (Definition 4). Formalizing the idea that silent steps do not take any time, we observe the parameterized process as τ tends to infinity, making therefore the parameterized transitions immediate. The limit process may do infinitely many transitions in a finite amount of time, i.e. may be *discontinuous* [11]. A Markov chain that can behave discontinuously we call a Markov process. In standard literature this model is usually considered pathological and we only use it to justify our results. We define a notion of ordinary lumping for Markov processes (Definition 3) and, based on that, a new notion of lumping for Markov chains with fast transitions, called τ-lumping (Definition 5). We justify the latter notion by showing that the following diagram commutes:



In the second step, we treat a Markov chain with silent steps as a class of Markov chains with fast transitions that have the same structure but different weights assigned to silent steps (this is achieved by introducing a relation $\sim$). We define a notion of lumping, called $\tau_\sim$-lumping, directly for Markov chains with silent steps, and show that it is a proper lifting of τ-lumping to equivalence classes. In other words, we show that $\tau_\sim$-lumping induces a τ-lumping such that the following diagram commutes:



We only give proof sketches. For detailed proofs of our results we refer to the full version of this paper [12].

II. Preliminaries

All vectors are column vectors if not indicated otherwise. $\mathbf{1}^n$ denotes the vector of n 1's. $\mathbf{0}^{n\times m}$ denotes the $n \times m$ zero matrix. I^n denotes the $n \times n$ identity matrix. When it is clear from the context, we omit the n and m. We write $A > 0$ (resp. $A \geq 0$) when all elements of a matrix or a vector A are greater than (resp. greater than or equal to) zero. By $\operatorname{diag}(A_1, \ldots, A_n)$ we denote a block matrix with blocks $A_1, \ldots, A_n$ on the diagonal and $\mathbf{0}$'s elsewhere.

Partitioning is a central notion in the definition of lumping.

Definition 1 (Partitioning): Let $\mathcal{S}$ be a set. A set $\mathcal{P} = \{C_1, \ldots, C_N\}$ is a *partitioning* of $\mathcal{S}$ if $\mathcal{S} = C_1 \cup \ldots \cup C_N$, $C_i \neq \emptyset$ and $C_i \cap C_j = \emptyset$ for $i \neq j$.

The partitionings $\mathcal{P} = \{S\}$ and $\mathcal{P} = \{\{i\} \mid i \in S\}$ are called *trivial*.

With every partitioning $\mathcal{P} = \{C_1, \ldots, C_N\}$ of $\mathcal{S} = \{1, \ldots, n\}$ we associate the following matrices. The matrix $V \in \mathbb{R}^{n\times N}$ defined as

$$V[i,j] = \begin{cases} 0, & i \notin C_j \\ 1, & i \in C_j \end{cases}$$

is called the *collector* matrix for $\mathcal{P}$. Its j-th column has 1's for elements corresponding to states in C_j and has zeroes otherwise. Note that $V \cdot \mathbf{1} = \mathbf{1}$. For the trivial partitionings, we have $V = \mathbf{1}$ and $V = I$.

A matrix $U \in \mathbb{R}^{N\times n}$ such that $U \geq 0$ and $UV = I^{N\times N}$ is a *distributor* matrix for $\mathcal{P}$. It can be readily seen that U is actually any matrix of which the elements of the i-th row that correspond to elements in C_i sum up to one while the other elements of the row are 0. For the trivial partitioning $\mathcal{P} = \{S\}$ a distributor is a vector with elements that sum up to 1; for the trivial partitioning $\mathcal{P} = \{\{i\} \mid i \in S\}$ there exists only one distributor (I).

Example 1: Let $\mathcal{S} = \{1, 2, 3\}$ and $\mathcal{P} = \{\{1, 2\}, \{3\}\}$. Then $V = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $U = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} & 0 \\ 0 & 0 & 1 \end{pmatrix}$ is an example for a distributor matrix for $\mathcal{P}$.

III. Lumping Markov Processes

In this section we define Markov processes and a notion of ordinary lumping for them. Since we drop the usual requirement that a Markov process is continuous, we generalize the existing theory of lumpability [13].

A. Markov Processes

A Markov process is a finite-state continuous-time stochastic process that is homogeneous and satisfies the Markov property [1], [2]. It is known that a Markov process with an ordered state space is completely determined by a transition matrix (called *its* transition matrix) and a vector that gives the starting probabilities of the process for each state (called the *initial probability* vector).

Definition 2 (Transition matrix): A matrix $P(t) \in \mathbb{R}^{n\times n}$, $(t > 0)$ is called a *transition matrix* iff

1) $P(t) \geq 0$,
2) $P(t) \cdot \mathbf{1} = \mathbf{1}$ and
3) $P(t+s) = P(t) \cdot P(s)$ for all $s > 0$.

If $\lim_{t\to 0} P(t)$ is equal to the identity matrix, then $P(t)$ is considered *continuous*, otherwise it is *discontinuous*. Note that the limit always exists [1].

Example 2: Let $0 \leq p \leq 1$ and $\lambda \geq 0$. Then

$$P(t) = \begin{pmatrix} (1-p)\cdot e^{-p\lambda t} & p\cdot e^{-p\lambda t} & 1-e^{-p\lambda t} \\ (1-p)\cdot e^{-p\lambda t} & p\cdot e^{-p\lambda t} & 1-e^{-p\lambda t} \\ 0 & 0 & 1 \end{pmatrix}$$

is a transition matrix. It is discontinuous because

$$\lim_{t\to 0} P(t) = \begin{pmatrix} 1-p & p & 0 \\ 1-p & p & 0 \\ 0 & 0 & 1 \end{pmatrix} \neq I.$$

The following theorem [11], [14] gives a convenient characterization of a transition matrix that does not depend on t.

Theorem 1: Let $(\Pi, Q) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n}$ be such that:

1) $\Pi \geq 0,\ \Pi\cdot\mathbf{1} = \mathbf{1},\ \Pi^2 = \Pi$,
2) $\Pi Q = Q\Pi = Q$,
3) $Q\cdot\mathbf{1} = \mathbf{0}$ and
4) $Q + c\Pi \geq 0$ for some $c \geq 0$.

Then $P(t) = \Pi e^{Qt}$ is a transition matrix. Moreover, the converse also holds: For any transition matrix $P(t)$ there exists a unique pair (Π, Q) that satisfies Conditions 1–4 and such that $P(t) = \Pi e^{Qt}$.

Proof: See [11], [14]. ∎

Note that, if $P(t) = \Pi \cdot e^{Qt}$ is continuous, then it follows that $\Pi = I$ and that Q is a *generator* matrix, i.e. a square matrix of which the non-diagonal elements are non-negative and each diagonal element is the additive inverse of the sum of the non-diagonal elements of the same row.

Our results do not depend on the initial probability vector nor on the exact nature of states. So, when we speak of Markov processes, we actually mean the class of processes with the same transition matrix but with possibly different sets of states and initial probability vectors. This allows us to identify a Markov process that has the transition matrix $P(t) = \Pi\cdot e^{Qt} \in \mathbb{R}^{n\times n}$ with the pair $(\Pi, Q) \in \mathbb{R}^{n\times n}\times\mathbb{R}^{n\times n}$ and to refer to the indices $\{1,\ldots,n\}$ as its states. A Markov process is called (dis)continuous if its transition matrix is (dis)continuous. In standard literature, it is always assumed that $\Pi = I$ [1], [2]. We call continuous Markov processes *Markov chains*.

We now explain the behavior of a Markov process $(\Pi, Q) \in \mathbb{R}^{n\times n}\times\mathbb{R}^{n\times n}$. Note that, after a suitable renumbering of the states, Π gets the following form [11]:

$$\Pi = \begin{pmatrix} \Pi_1 & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Pi_2 & \ldots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \Pi_M & \mathbf{0} \\ \overline{\Pi}_1 & \overline{\Pi}_2 & \ldots & \overline{\Pi}_M & \mathbf{0} \end{pmatrix}$$

where for all $1 \leq i \leq M$, $\Pi_i = \mathbf{1}\cdot\mu_i$ and $\overline{\Pi} = \delta_i\cdot\mu_i$ for a *row* vector $\mu_i > 0$ such that $\mu_i\cdot\mathbf{1} = 1$ and a vector $\delta_i \geq 0$ such that $\sum_{i=1}^{M}\delta_i = \mathbf{1}$. This numbering determines a partitioning $\mathcal{E} = \{E_1,\ldots,E_M,T\}$ of $\mathcal{S} = \{1,\ldots,n\}$ (called the *ergodic partitioning*) into *ergodic* classes, $E_1,\ldots,E_M$, determined by $\Pi_1,\ldots,\Pi_M$, and into a class of *transient* states, T, determined by $\overline{\Pi}_1,\ldots,\overline{\Pi}_M$.

In an ergodic class a Markov process spends a non-zero amount of time switching rapidly among its elements. This time is exponentially distributed and determined by the matrix Q. If the ergodic class contains one state only, then Q has the form of a generator in that state, and $Q[i,j]$ for $i \neq j$ is interpreted as the rate from i to j. For every ergodic class E_i, the vector μ_i is the vector of *ergodic probabilities* and, for each state in E_i, it holds the probability that the process is in that state. If a Markov process is continuous, i.e. if it is a Markov chain, then every ergodic class E_i must contain exactly one state and therefore $\mu_i = \begin{pmatrix}1\end{pmatrix}$.

In a transient state the process spends no time (with probability one) and goes immediately to an ergodic class (and stays trapped there). The vector δ_i holds the *trapping probabilities* from transient states to the ergodic class E_i and $\delta_i[j] > 0$ iff state j can be trapped in some ergodic class E_i. A Markov chain cannot have transient states.

Example 3: a) For $0 < p < 1$, $\lambda > 0$, the pair (Π, Q) defined as:

$$\Pi = \begin{pmatrix} 1-p & p & 0 \\ 1-p & p & 0 \\ 0 & 0 & 1 \end{pmatrix}\quad Q = \begin{pmatrix} -p(1-p)\lambda & -p^2\lambda & p\lambda \\ -p(1-p)\lambda & -p^2\lambda & p\lambda \\ 0 & 0 & 0 \end{pmatrix}$$

is a (discontinuous) Markov process. Its has two ergodic classes $E_1 = \{1,2\}$ and $E_2 = \{3\}$ and no transient states. The corresponding ergodic probability vectors are $\mu_1 = \begin{pmatrix}1-p & p\end{pmatrix}$ and $\mu_2 = \begin{pmatrix}1\end{pmatrix}$. In the first two states the process exhibits non-continuous behavior. It constantly switches among those states and it is found in the first one with probability $1-p$ and in the second one with probability p. We will see later that the amount of time the process spends switching is exponentially distributed with the rate $p\lambda$.

b) Let, for $0 < p < 1$ and $\lambda, \mu, \rho > 0$, (Π, Q) be defined

as:

$$\Pi = \begin{pmatrix} 0 & p & 1-p & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{and}$$

$$Q = \begin{pmatrix} 0 & -p\lambda & -(1-p)\mu & p\lambda + (1-p)\mu \\ 0 & -\lambda & 0 & \lambda \\ 0 & 0 & -\mu & \mu \\ \rho & 0 & 0 & -\rho \end{pmatrix}.$$

The ergodic partitioning is $E_1 = \{2\}$, $E_2 = \{3\}$, $E_3 = \{4\}$ and $T = \{1\}$ (note that the numbering does not make the ergodic partitioning explicit since the transient state precedes the ergodic ones). We have $\mu_i = \big(1\big)$ for all $i = 1, 2, 3$ and $\delta_1 = \big(p\big)$, $\delta_2 = \big(1-p\big)$ and $\delta_3 = \big(0\big)$. If the process is in the state 1, then with probability p it is trapped in the state 2, the only state in the ergodic class E_1, and with probability $1-p$ it is trapped in the state 3. It cannot be trapped in the state 4.

B. Ordinary Lumping

We now define a notion of lumping for Markov processes and prove some standard theorems for the new, more general, setting.

Definition 3 (Ordinary lumping): A partitioning $\mathcal{P}$ of $\{1, \ldots, n\}$ is called an *ordinary lumping* of a Markov process $(\Pi, Q) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n}$ iff

$$VU\Pi V = \Pi V \quad \text{and} \quad VUQV = QV$$

where V and U are respectively the collector and a distributor matrix for $\mathcal{P}$.

The lumping condition does not depend on the particular choice of the non-zero elements of U. Suppose that $VU\Pi V = \Pi V$ and that there exists $U' \geq 0$ such that $U'V = I$. Then $VU'\Pi V = VU'VU\Pi V = VU\Pi V = \Pi V$. Similarly, $VU'QV = QV$. The condition actually says that the rows of ΠV (resp. QV) that correspond to the states that belong to the same class must be equal [3]. Intuitively, this means that the states in the same class behave in the same way when transiting to other classes.

Note also that the partitioning $\mathcal{P} = \{S\}$ is always an ordinary lumping. However, there is usually some reward structure imposed on the process that forbids the trivial case. In this paper we abstract from rewards since they can be straightforwardly added.

Theorem 2: Let (Π, Q) be a Markov process and let $\mathcal{P} = \{C_1, \ldots, C_N\}$ be an ordinary lumping of (Π, Q). Define

$$\hat{\Pi} = U\Pi V \quad \text{and} \quad \hat{Q} = UQV.$$

Then $(\hat{\Pi}, \hat{Q}) \in \mathbb{R}^{N\times N} \times \mathbb{R}^{N\times N}$ is a Markov process.

Proof: [Sketch] We prove that the conditions of Theorem 1 hold. The derivation of Conditions 1–3 is straightforward. To derive Condition 4 we use the same $c \geq 0$ as the one for which $Q + c\Pi \geq 0$. ∎

The definition of $(\hat{\Pi}, \hat{Q})$ does not depend on a particular distributor matrix U. To show this, let U' be another distributor matrix for $\mathcal{P}$. Then $U'\Pi V = U'VU\Pi V = U\Pi V$. Similarly, $U'QV = UQV$.

If $\mathcal{P}$ is an ordinary lumping of (Π, Q) and $\hat{\Pi}$ and $\hat{Q}$ are defined as in the preceding theorem, then we say that (Π, Q) *lumps to* $(\hat{\Pi}, \hat{Q})$ (with respect to $\mathcal{P}$). We write $(\Pi, Q) \overset{\mathcal{P}}{\rightsquigarrow} (\hat{\Pi}, \hat{Q})$ when $\mathcal{P}$ is an ordinary lumping of (Π, Q) and (Π, Q) lumps to $(\hat{\Pi}, \hat{Q})$ with respect to $\mathcal{P}$.

Note that, if $(\Pi, Q) \overset{\mathcal{P}}{\rightsquigarrow} (\hat{\Pi}, \hat{Q})$ and (Π, Q) is a Markov chain, then $\hat{\Pi} = U\Pi V = UIV = I$ and by Theorem 1, $\hat{Q}$ is a generator matrix. In this case, our notion coincides with the known definition of ordinary lumping for Markov chains proposed in [13].

Example 4: a) Let (Π, Q) be the Markov process from Example 3a. Then $\mathcal{P} = \big\{\{1, 2\}, \{3\}\big\}$ is an ordinary lumping and the lumped process $(\hat{\Pi}, \hat{Q})$ is defined by:

$$\hat{\Pi} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \hat{Q} = \begin{pmatrix} -p\lambda & p\lambda \\ 0 & 0 \end{pmatrix}.$$

Note that, in this case, the lumped process is a Markov chain. This example also shows how a whole ergodic class can constitute a lumping class. It is not hard to show that an ergodic class is always a correct lumping class.

b) Let (Π, Q) be the Markov process from Example 3b. If $\lambda \neq \mu$, by checking the lumping condition for all possible partitionings, we conclude that this Markov process does not have a non-trivial lumping. The states 2 and 3 cannot be joined in a class because they have different rates leading to the state 4. The state 1 cannot be joined together with the state 2 because 2 cannot reach the state 3 whereas the state 1 can. Similarly, 1 cannot be joined together with the state 3.

For $\lambda = \mu$ however, the partitioning $\mathcal{P} = \big\{\{1\}, \{2, 3\}, \{4\}\big\}$ is an ordinary lumping and, with respect to it, (Π, Q) lumps to $(\hat{\Pi}, \hat{Q})$ defined as:

$$\hat{\Pi} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \hat{Q} = \begin{pmatrix} 0 & -\lambda & \lambda \\ 0 & -\lambda & \lambda \\ \rho & 0 & -\rho \end{pmatrix}.$$

If $\lambda = \mu$, also the partitioning $\mathcal{P} = \big\{\{1, 2, 3\}, \{4\}\big\}$ is an ordinary lumping. With respect to this partitioning (Π, Q) lumps to $(\hat{\Pi}, \hat{Q})$ defined as:

$$\hat{\Pi} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \hat{Q} = \begin{pmatrix} -\lambda & \lambda \\ \rho & -\rho \end{pmatrix},$$

which is a Markov chain.

The following theorem reflects the conditions of Definition 3 to the corresponding transition matrix.

Theorem 3: Let (Π, Q) be a Markov process and let $P(t) = \Pi e^{Qt}$ $(t > 0)$, be its transition matrix. Let $\mathcal{P}$ be an ordinary lumping of (Π, Q). Then

$$VUP(t)V = P(t)V.$$

Proof: [Sketch] The equality can be derived directly by using that $\Pi Q^n = Q^n$ and $VUQ^nV = Q^nV$ for all $n \geq 1$. ∎

The following theorem shows that the transition matrix of the lumped process can also be obtained directly from the transition matrix of the original process.

Theorem 4: Let $(\Pi, Q) \overset{\mathcal{P}}{\rightsquigarrow} (\hat{\Pi}, \hat{Q})$. Let $P(t) = \Pi e^{Qt}$ and $\hat{P}(t) = \hat{\Pi} e^{\hat{Q}t}$ $(t > 0)$ be the transition matrices of (Π, Q) and $(\hat{\Pi}, \hat{Q})$ respectively. Then

$$\hat{P}(t) = UP(t)V.$$

Proof: [Sketch] The equality is derived directly by using that $(UQV)^n = UQ^nV$ for all $n \geq 0$, and that $VUQ^nV = Q^nV$ for all $n \geq 1$. ∎

IV. Lumping Markov Chains with Fast Transitions

In this section we introduce an extension to Markov chains by letting them perform steps of (drastically) different scales. In the limit these processes become Markov processes. We define a notion of lumping for the new model.

A. Markov Chains with Fast Transitions

A Markov chain with fast transitions is defined as a pair of generator matrices; the first matrix represents the normal (slow) transitions, while the second matrix represents the (*speed* of) fast transitions.

Definition 4 (Markov chain with fast transitions): Let Q_λ and Q_τ be generator matrices. The *Markov chain with fast transitions* determined by Q_λ and Q_τ, denoted (Q_λ, Q_τ), is a function that assigns to each $\tau > 0$ the Markov chain $(I, Q_\lambda + \tau Q_\tau)$.

We picture a Markov chain with fast transitions (Q_λ, Q_τ) by the usual visual representation of the generator matrix $Q_\lambda + \tau Q_\tau$ (see Fig. 3).

If Q is a generator matrix, then $\Pi = \lim_{t\to\infty} e^{Qt}$ is called the *ergodic projection* of Q. It is proven in [1] that the limit always exists; moreover it is known (see [15] and the references therein) that Π is actually the unique matrix such that $\Pi \geq 0$, $\Pi \cdot \mathbf{1} = \mathbf{0}$, $\Pi^2 = \Pi$, $\Pi Q = Q\Pi = \mathbf{0}$ and $\operatorname{rank}(\Pi) + \operatorname{rank}(Q) = n$. The following theorem shows that, when $\tau \to \infty$, a Markov chain with fast transitions becomes a Markov process and that, in this case, the behavior of the Markov chain with fast transitions depends only on the ergodic projection of the matrix that models the fast transitions and not on the matrix itself.

Theorem 5: Let $P_\tau(t) = e^{(Q_\lambda + \tau Q_\tau)t}$. Then

$$\lim_{\tau\to\infty} P_\tau(t) = \Pi e^{Qt} \quad (t > 0)$$

where $\Pi = \lim_{t\to\infty} e^{Q_\tau t}$ is the ergodic projection of Q_τ and $Q = \Pi Q_\lambda \Pi$. In addition, (Π, Q) satisfies Conditions 1–4 of Theorem 1.

Proof: See [16] for the first proof, or [17] for a proof written in more modern terms. See [11] for the proof that convergence is also uniform. ∎

When (Π, Q) is the limit of (Q_λ, Q_τ) we write $(Q_\lambda, Q_\tau) \to_\infty (\Pi, Q)$. In this situation, we also define the ergodic partitioning of (Q_λ, Q_τ) to be the ergodic partitioning of (Π, Q).

The ergodic partitioning of (Q_λ, Q_τ) can also be obtained differently. We write $i \to j$ if $Q_\tau[i, j] > 0$, i.e. if there is a direct fast transition from i to j. Let $\twoheadrightarrow$ denote the reflexive-transitive closure of $\to$. If $i \twoheadrightarrow j$ we say that j is *reachable* from i. If $i \twoheadrightarrow j$ and $j \twoheadrightarrow i$ we say that i and j *communicate* and write $i \leftrightarrows j$. Now, it can be shown (see [1]) that every ergodic class is actually a closed class of communicating states, closed meaning that for all i inside the class there does not exist j outside the class such that $i \to j$.

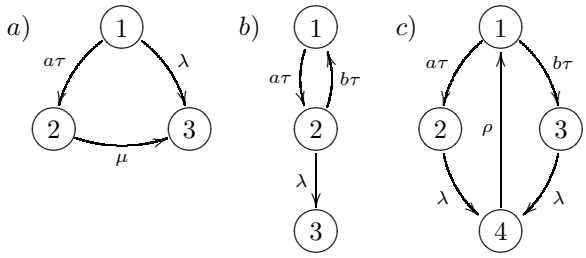


Fig. 3. Markov chains with fast transitions from Example 5

Example 5: a) Consider a Markov chain with fast transitions (Q_λ, Q_τ) depicted in Fig. 3a. It is defined with

$$Q_\lambda = \begin{pmatrix} -\lambda & 0 & \lambda \\ 0 & -\mu & \mu \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad Q_\tau = \begin{pmatrix} -a & a & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

The transition from state 1 to state 2 is fast and has the speed a. The other two transitions are normal.
The limit of (Q_λ, Q_τ) is obtained as follows:

$$\Pi = \lim_{t\to\infty} e^{Q_\tau t} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and}$$

$$Q = \Pi Q_\lambda \Pi = \begin{pmatrix} 0 & -\mu & \mu \\ 0 & -\mu & \mu \\ 0 & 0 & 0 \end{pmatrix}.$$

The ergodic partitioning is $E_1 = \{2\}$, $E_2 = \{3\}$ and $T = \{1\}$.

b) Consider the Markov chain with fast transitions depicted in Fig. 3b. The limit of this Markov chain with fast transitions is the Markov process from Example 3a (for $p = \frac{a}{a+b}$).

c) The limit of the Markov chain with fast transitions in Fig. 3c is the Markov process of Example 3b (for $p = \frac{a}{a+b}$ and $\lambda = \mu$).

B. τ-lumping

We now define a special notion of lumping for Markov chains with fast transitions. The notion is based on the notion of ordinary lumping for Markov processes: a partitioning is a lumping of a Markov chain with fast transitions if it is an ordinary lumping of its limit.

Definition 5 (τ-lumping): A partitioning $\mathcal{P}$ of $\{1,\ldots,n\}$ is called a *τ-lumping* of a Markov chain with fast transitions $(Q_\lambda, Q_\tau) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n}$ if it is an ordinary lumping of the Markov process (Π, Q) where $(Q_\lambda, Q_\tau) \to_\infty (\Pi, Q)$.

As for Markov processes, we give a definition of the lumped process by multiplying Q_λ and Q_τ with the collector matrix and a distributor matrix. Since ordinary lumping for Markov processes is closed under Markov chains, this technique gives a Markov chain with fast transitions as a result. However, since the lumping condition does not hold for Q_λ and Q_τ, but only for Π and Q, the definition of the lumped process may depend on the choice for a distributor. We define a special distributor and show that it is correct in the sense that it gives a lumped process of which the limit is the lumped version of the limit of the original Markov chain with fast transitions.

Definition 6: Let $\mathcal{P} = \{C_1 \ldots, C_N\}$ be a *τ-lumping* of a Markov chain with fast transitions (Q_λ, Q_τ) and let $\Pi = \lim_{t\to\infty} e^{Q_\tau t}$. Define $W \in \mathbb{R}^{N\times n}$ as

$$W[k,i] = \begin{cases} 0, & i \notin C_k \\ \frac{\Pi[i,i]}{\sum_{j\in C_k} \Pi[j,j]}, & i \in C_k, \sum_{j\in C_k} \Pi[j,j] > 0 \\ \frac{1}{|C_k|}, & i \in C_k, \sum_{j\in C_k} \Pi[j,j] = 0 \end{cases}$$

for $1 \le k \le N$. Define $\hat{Q}_\lambda, \hat{Q}_\tau \in \mathbb{R}^{N\times N}$ as

$$\hat{Q}_\lambda = WQ_\lambda V \text{ and } \hat{Q}_\tau = WQ_\tau V.$$

We say that (Q_λ, Q_τ) *τ-lumps* to $(\hat{Q}_\lambda, \hat{Q}_\tau)$ (with respect to $\mathcal{P}$).

Let us explain the form of W. We consider it as a matrix that gives weights to the elements of Q_λ and Q_τ. The weights are normalized to fit the form of a distributor. States that belong to ergodic classes are identified by the fact that their diagonal elements in Π are greater than zero. The transient states have diagonal elements in Π equal to zero. An exponential rate that goes out of a state in an ergodic class is weighted according to its ergodic probability. The transient states do not influence the ergodic probabilities, so transient states that are lumped together with states from ergodic classes are assigned zero weight. We have complete freedom when lumping transient states with other transient states because they play no role when τ goes to infinity. We choose to assign them equal weights.

Example 6: a) Consider the Markov chain with fast transitions depicted in Fig. 3a. We show that $\{\{1,2\},\{3\}\}$ is its τ-lumping and that the process τ-lumps to the one in Fig. 4a. We obtain

$$V = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad W = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

The conditions for τ-lumping hold:

$$VW\Pi Q_\lambda \Pi V = \begin{pmatrix} -\mu & \mu \\ -\mu & \mu \\ 0 & 0 \end{pmatrix} = \Pi Q_\lambda \Pi V$$

$$\text{and} \quad VW\Pi V = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} = \Pi V.$$

The lumped process is defined by the following two matrices and is indeed depicted in Fig. 4a:

$$\hat{Q}_\lambda = WQ_\lambda V = \begin{pmatrix} -\mu & \mu \\ 0 & 0 \end{pmatrix}, \ \hat{Q}_\tau = WQ_\tau V = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

This example illustrates how, in transient states, fast transitions have priority over slow transitions.

b) Consider the Markov chain with fast transitions depicted in Fig. 3b. It is easily checked that $\{\{1,2\},\{3\}\}$ is a τ-lumping of this Markov chain with fast transitions. We obtain

$$W = \begin{pmatrix} \frac{b}{a+b} & \frac{a}{a+b} & 0 \\ 0 & 0 & 1 \end{pmatrix}, \ \hat{Q}_\lambda = \begin{pmatrix} -\frac{a\,\lambda}{a+b} & \frac{a\,\lambda}{a+b} \\ 0 & 0 \end{pmatrix}, \ \hat{Q}_\tau = \mathbf{0}.$$

So, the process τ-lumps to the one in Fig. 4b.
This example shows that when two ergodic states with different slow transition rates are lumped together, the resulting state is ergodic and it can perform the same slow transition but with an adapted rate. The example also shows that the Markov chain with fast transitions of Fig. 3b spends an exponentially distributed amount of time with rate $\frac{a\,\lambda}{a+b}$ switching between the state 1 and the state 2.

c) Example 4b shows that for the Markov chain with fast transitions depicted in Fig. 3c, the partitionings $\mathcal{P} = \{\{1\},\{2,3\},\{4\}\}$ and $\mathcal{P} = \{\{1,2,3\},\{4\}\}$ are τ-lumpings. For the first partitioning we have

$$W = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \ \hat{Q}_\lambda = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -\lambda & \lambda \\ \rho & 0 & -\rho \end{pmatrix},$$

$$\hat{Q}_\tau = \begin{pmatrix} -a-b & a+b & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

For the second partitioning we obtain

$$W = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \ \hat{Q}_\lambda = \begin{pmatrix} -\lambda & \lambda \\ \rho & -\rho \end{pmatrix}, \ \hat{Q}_\tau = \mathbf{0}.$$

The two lumped Markov chains with fast transitions are depicted in Fig. 4c and Fig. 4d respectively.
This example shows that τ-lumping need not eliminate all silent steps (Fig. 4c). It also shows how transient states can be lumped with ergodic states, resulting in an ergodic state (Fig. 4d).

The following example shows some Markov chains with fast transitions that are minimal in the sense that they only admit the trivial τ-lumpings.

Example 7: a) Consider the Markov chain with fast transitions in Fig. 5a. From Example 4b it directly follows that, for $\lambda \neq \mu$, this Markov chain with fast transitions does not have a non-trivial lumping.

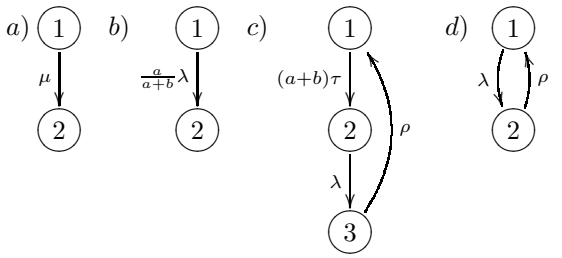


Fig. 4. τ-lumped Markov chains with fast transitions – Example 6

b) The Markov chain with fast transitions in Fig. 5b also has only the trivial lumpings (unless $\lambda = \mu$ and then the states 3 and 4 can form a lumping class).
c) The Markov chain with fast transitions in Fig. 5c has only the trivial lumpings if $\lambda \neq \mu$ and $b \neq c$. If $\lambda = \mu$ then the states 3 and 4 can form a lumping class. If $b = c$ then the states 1 and 2 constitute a lumping class.

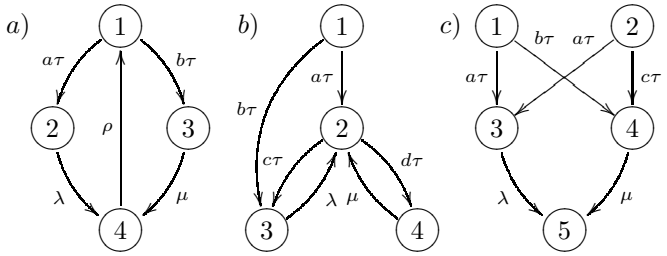


Fig. 5. Markov chains with fast transitions without non-trivial τ-lumpings – Example 7

The following lemmas are used to support a proof that τ-lumping as defined by Definition 5 is sound. Lemma 1 justifies a more refined numbering of states that allows for a comprehensive matrix manipulation in the proofs. It expresses an important connection between the ergodic partitioning and the lumping partitioning. If two lumping classes contain states from the same ergodic class, then whenever one of the lumping classes contains states from another ergodic class, the other must also contain states from that ergodic class.

Lemma 1: Let (Q_λ, Q_τ) be a Markov chain with fast transitions and let $\mathcal{E} = \{E_1, \ldots, E_M, T\}$ be its ergodic partitioning. Let $\mathcal{P} = \{C_1, \ldots, C_N\}$ be a τ-lumping of (Q_λ, Q_τ). Then, for all $1 \leq i, j \leq M$ and $1 \leq k, \ell \leq N$, if $E_i \cap C_k \neq \emptyset$, $E_i \cap C_\ell \neq \emptyset$ and $E_j \cap C_k \neq \emptyset$, then $E_j \cap C_\ell \neq \emptyset$.

Proof: [Sketch] We analyze the rows of ΠV for the states that belong to $E_i \cap C_k$, $E_j \cap C_k$ and $E_i \cap C_\ell$. Recall that the lumping condition $VU\Pi V = \Pi V$ implies that all rows of ΠV that correspond to a lumping class must be equal. Recall that the rows of Π that correspond to an ergodic class are equal. ∎

Let $\mathcal{P} = \{C_1, \ldots, C_N\}$ be a lumping and let $\mathcal{E} = \{E_1, \ldots, E_M, T\}$ be the ergodic partitioning. Let $C_1, \ldots, C_L$ contain states from ergodic classes (and possibly some transient states too) and let $C_{L+1}, \ldots, C_N$ consist only of transient states. By Lemma 1 we can rearrange $C_1, \ldots, C_N$ and $E_1, \ldots, E_M$ and divide them into S blocks as follows. Let $E_{i1}, \ldots, E_{ie_i}$ and $C_{i1}, \ldots, C_{ic_i}$ $(1 \leq i \leq S)$ denote the ergodic and lumping classes such that, for all $1 \leq j \leq e_i$, $1 \leq k \leq c_i$, $E_{ij} \cap C_{ik} \neq \emptyset$, and that E_{ij} has no common elements with other partitioning classes. Note that $L = \sum_{i=1}^{S} c_i$. We then renumber states such that those that belong to an ergodic class with a lower index precede those that belong to an ergodic class with a higher index (assuming the lexicographic order). We also renumber transient states to divide them into those that are lumped together with some states from ergodic classes and those that are lumped only with other transient states.

The effect of the renumbering is that the matrices Π, V and W get the following forms:

$$\Pi = \begin{pmatrix} \Pi_1 & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Pi_2 & \ldots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \Pi_S & \mathbf{0} & \mathbf{0} \\ \overline{\Pi}_1 & \overline{\Pi}_2 & \ldots & \overline{\Pi}_S & \mathbf{0} & \mathbf{0} \\ \widetilde{\Pi}_1 & \widetilde{\Pi}_2 & \ldots & \widetilde{\Pi}_S & \mathbf{0} & \mathbf{0} \end{pmatrix}$$

$$\begin{aligned} \Pi_i &= \operatorname{diag}(\Pi_{i1}, \ldots, \Pi_{ie_i}) & \Pi_{ij} &= \mathbf{1}^{|E_{ij}|} \cdot \mu_{ij} \\ \overline{\Pi}_i &= \begin{pmatrix} \overline{\Pi}_{i1} & \ldots & \overline{\Pi}_{ie_i} \end{pmatrix} & \overline{\Pi}_{ij} &= \overline{\delta}_{ij} \cdot \mu_{ij} \\ \widetilde{\Pi}_i &= \begin{pmatrix} \widetilde{\Pi}_{i1} & \ldots & \widetilde{\Pi}_{ie_i} \end{pmatrix} & \widetilde{\Pi}_{ij} &= \widetilde{\delta}_{ij} \cdot \mu_{ij}, \end{aligned}$$

where the matrices $\overline{\Pi}_i$ and $\widetilde{\Pi}_i$ respectively represent the transient states that are lumped together with ergodic classes and the ones that are lumped only with other transient states; the vectors $\overline{\delta}_{ij}$ and $\widetilde{\delta}_{ij}$ are the corresponding restrictions of the vector δ_{ij}.

The collector matrix V associated with $\mathcal{P}$ now has the following form:

$$V = \begin{pmatrix} V_1 & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & V_2 & \ldots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & V_S & \mathbf{0} \\ \overline{V}_1 & \overline{V}_2 & \ldots & \overline{V}_S & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & \widetilde{V} \end{pmatrix} \qquad V_i = \begin{pmatrix} V_{i1} \\ \vdots \\ V_{ie_i} \end{pmatrix}$$

$$\begin{aligned} V_{ij} &= \operatorname{diag}\left(\mathbf{1}^{|E_{i1} \cap C_{i1}|}, \ldots, \mathbf{1}^{|E_{ie_i} \cap C_{ic_i}|}\right) \\ \overline{V}_i &= \operatorname{diag}\left(\mathbf{1}^{|T \cap C_{i1}|}, \ldots, \mathbf{1}^{|T \cap C_{ic_i}|}\right) \\ \widetilde{V} &= \operatorname{diag}\left(\mathbf{1}^{|T \cap C_{L+1}|}, \ldots, \mathbf{1}^{|T \cap C_N|}\right). \end{aligned}$$

The matrix W of Definition 6 has the following form:

$$W = \begin{pmatrix} W_1 & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & W_2 & \ldots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & W_S & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} & \widetilde{W} \end{pmatrix}$$

$$\begin{aligned} W_i &= \begin{pmatrix} W_{i1} & \ldots & W_{ie_i} \end{pmatrix} \\ \widetilde{W} &= \operatorname{diag}(\widetilde{w}_{L+1}, \ldots, \widetilde{w}_N) \end{aligned}$$

where

$$W_{ij} = \operatorname{diag}\left(\frac{\mu_{ij}^{(1)}}{\sum_{k=1}^{e_i}\mu_{ik}^{(1)}\cdot\mathbf{1}},\ldots,\frac{\mu_{ij}^{(c_i)}}{\sum_{k=1}^{e_i}\mu_{ik}^{(c_i)}\cdot\mathbf{1}}\right)$$

and

$$\widetilde{w}_i = \left(\frac{1}{|C_i|}\ \cdots\ \frac{1}{|C_i|}\right) \in \mathbb{R}^{1\times|C_i|}.$$

The following lemma gives an important property of the matrix W.

Lemma 2: Let Π, V and W be as in Definition 6. Then

$$\Pi V W \Pi = \Pi V W.$$

Proof: [Sketch] It suffices to show that $X_i V_i W_i \Pi_i = X_i V_i W_i$, for all $X_i \in \{\Pi_i, \overline{\Pi}_i, \widetilde{\Pi}_i\}$ and $1 \le i \le S$. This is done by showing that $\mu_{ij} V_{ij} W_{ik} \Pi_{ik} = \mu_{ij} V_{ij} W_{ik}$ for $1 \le j, k \le e_i$. ∎

The following theorem shows the correctness of Definition 6.

Theorem 6: Suppose $(Q_\lambda, Q_\tau) \overset{\mathcal{P}}{\rightsquigarrow}_\tau (\hat{Q}_\lambda, \hat{Q}_\tau)$, $(Q_\lambda, Q_\tau) \to_\infty (\Pi, Q)$ and $(\Pi, Q) \overset{\mathcal{P}}{\rightsquigarrow} (\hat{\Pi}, \hat{Q})$. Then

$$(\hat{Q}_\lambda, \hat{Q}_\tau) \to_\infty (\hat{\Pi}, \hat{Q}).$$

Proof: [Sketch] Recall that $\hat{\Pi}$ is the ergodic projection of $\hat{Q}_\tau$ iff $\hat{\Pi} \ge 0$, $\hat{\Pi}\cdot\mathbf{1} = \mathbf{1}$, $\hat{\Pi}^2 = \hat{\Pi}$, $\hat{\Pi}\hat{Q}_\tau = \hat{Q}_\tau\hat{\Pi} = \mathbf{0}$ and $\operatorname{rank}(\hat{\Pi}) + \operatorname{rank}(\hat{Q}_\tau) = N$. The first three conditions follow from Theorems 1 and 2. The fourth condition follows from Lemma 2. We prove that $\operatorname{rank}(\hat{\Pi}) + \operatorname{rank}(\hat{Q}_\tau) = N$ by showing that $\operatorname{rank}(\hat{\Pi}) = S$ and $\operatorname{rank}(\hat{Q}_\tau) = N - S$. Finally, we use Lemma 2 again to derive $\hat{\Pi}\hat{Q}_\lambda\hat{\Pi} = \hat{Q}$. ∎

V. Lumping Markov Chains with Silent Steps

We define a Markov chain with silent steps to be a Markov chain with fast transitions in which the speeds of the fast transitions are considered not known. In other words, a Markov chain with silent steps is obtained by abstracting from the speeds in a Markov chain with fast transitions. We give a notion of lumping that satisfies the following criterion: the lumping is good if it induces a τ-lumping for all possible speeds of fast transitions and, moreover, the slow transitions in the lumped process do not depend on those speeds.

A. Markov Chains With Silent Steps

First, we introduce an equivalence on matrices.

Definition 7 (Matrix grammar): Two matrices $A, B \in \mathbb{R}^{n\times n}$ are said to have the *same grammar*, denoted $A \sim B$, if for all $1 \le i, j \le n$, $A[i,j] = 0$ iff $B[i,j] = 0$.

Example 8: For $a, b, c \ne 0$, matrices $\left(\begin{smallmatrix} a & a \\ b & 0 \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} a & b \\ c & 0 \end{smallmatrix}\right)$ have the same grammar.

A Markov chain with silent steps is a class of Markov chains with fast transitions of which the generator matrices that model fast transitions have the same grammar; abstraction from the speeds is achieved by identifying generator matrices that have the same grammar.

Definition 8 (Markov chain with silent steps): A *Markov chain with silent steps* is a pair $(Q_\lambda, [Q_\tau]_\sim)$ where (Q_λ, Q_τ) is a Markov chain with fast transitions.

If $(Q_\lambda, [Q_\tau]_\sim)$ is a Markov chain with silent steps, it is visualized as the Markov chain with fast transitions (Q_λ, Q_τ) but omitting the speeds on τ transitions. Note that the notions of reachability, communication and ergodic partitioning are speed independent, and so they carry over to the setting of Markov chains with silent steps naturally.

B. $\tau_\sim$-lumping

In this section we introduce a notion of lumping for Markov chains with silent steps, called $\tau_\sim$-lumping, and show that it is a proper lifting of τ-lumping to equivalence classes of the relation $\sim$. First we give an example that shows that not every τ-lumping can be taken for $\tau_\sim$-lumping.

Example 9: a) Consider the Markov chain with silent steps depicted in Fig. 6a. The Example 6b shows that the partitioning $\mathcal{P} = \{\{1,2\},\{3\}\}$ is a τ-lumping for every possible speeds given to the silent transitions. However, the slow transition in the lumped process depends on the speed of the fast transitions.

b) Consider the Markov chain with silent steps depicted in Fig. 6b. The Example 7c shows, that although for some speeds the partitioning $\{\{1,2\},\{3\},\{4\}\}$ is a τ-lumping, it need not be so for some other speeds.

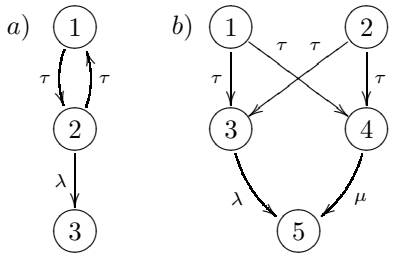


Fig. 6. Markov chains with silent steps – Example 9

Carefully restricting to the cases when τ-lumping is "speed independent" we come up with the following definition for $\tau_\sim$-lumping.

Definition 9 ($\tau_\sim$-lumping): Let $(Q_\lambda, [Q_\tau]_\sim) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n}$ be a Markov chain with silent steps and let $\{E_1, \ldots, E_M, T\}$ be its ergodic partitioning. Let $\mathcal{P}$ be a partitioning of $\{1,\ldots,n\}$. Let, for all $i \in \{1,\ldots,n\}$, $\operatorname{erg}(i) = \{j \in \bigcup_{1\le k\le M} E_k \mid i \twoheadrightarrow j\}$ be the set of all ergodic states reachable from the state i. Let for all $C \in \mathcal{P}$, $\operatorname{erg}(C)$ denote $\bigcup_{i\in C}\operatorname{erg}(i)$. We say that $\mathcal{P}$ is a *$\tau_\sim$-lumping* of $(Q_\lambda, [Q_\tau]_\sim)$ iff

1) for all $C \in \mathcal{P}$ at least one of the following holds:
   a) $\operatorname{erg}(C) \subseteq D$, for some $D \in \mathcal{P}$.
   b) $\operatorname{erg}(C) = E_i$, for some $1 \le i \le M$.
   c) $C \subseteq T$ and $i \to j$, for *exactly* one $i \in C$ and some $j \notin C$;

   and
2) for all $C \in \mathcal{P}$, all $i, j \in C \cap \left(\bigcup_{1\le k\le M} E_k\right)$ and all $D \in \mathcal{P}$ such that $C \ne D$,

$$\sum_{\ell\in D} Q_\lambda[i,\ell] = \sum_{\ell\in D} Q_\lambda[j,\ell].$$

Condition 1a says that the ergodic states reachable by silent transitions from the states in C are all in the same lumping class. Condition 1b says that the ergodic states reachable by silent transitions from the states in C constitute an ergodic class. Condition 1c says that C is a set of transient states with precisely one (silent) exit. Conditions 1a and 1b overlap when $E_i \subseteq D$. If, in addition, C contains only transient states and has only one exit, all three conditions overlap. Condition 2 says that every ergodic state in C must have the same accumulative rate to every other lumping class.

We now show that a $\tau_\sim$-lumping of a Markov chain with silent steps induces a grammar preserving τ-lumping of any Markov chain with fast transitions to which it corresponds.

Theorem 7: Suppose $(Q_\lambda, [Q_\tau]_\sim) \overset{\mathcal{P}}{\rightsquigarrow}_{\tau_\sim} (\hat{Q}_\lambda, [\hat{Q}_\tau]_\sim)$. Then $(Q_\lambda, Q_\tau) \overset{\mathcal{P}}{\rightsquigarrow}_{\tau} (\hat{Q}_\lambda, \hat{Q}_\tau)$, and for all $Q'_\tau \sim Q_\tau$ it holds that $(Q_\lambda, Q'_\tau) \overset{\mathcal{P}}{\rightsquigarrow}_{\tau} (\hat{Q}_\lambda, \hat{Q}'_\tau)$ and $\hat{Q}'_\tau \sim \hat{Q}_\tau$.

Proof: [Sketch] We assume that $(Q_\lambda, Q'_\tau) \rightarrow_\infty (\Pi, Q)$. We prove that $VU\Pi V = \Pi V$ by showing that the vector $\Pi^{(C,D)} \cdot \mathbf{1}$ has all elements equal, for all $C, D \in \mathcal{P}$, where $\Pi^{(C,D)}$ is the restriction of Π to the elements of C row-wise and the elements of D column-wise. Condition 1a of Definition 9 implies that $\Pi^{(C,F)} \cdot \mathbf{1} = \mathbf{1}$, for $F = D$. Condition 1b implies that $\Pi^{(C,F)} = \Pi^{(F,F)} = \mathbf{1} \cdot \mu_j^{(F)}$, for $F \cap E_i \neq \emptyset$. Condition 1c implies that $\Pi^{(C,F)} = \Pi^{(F,F)} = \mathbf{1} \cdot x$, for some row vector $x \neq \mathbf{0}$ and $F \cap \mathrm{erg}(i) \neq \emptyset$. We note that $\Pi^{(C,F)} \cdot \mathbf{1} = \mathbf{0}$ everywhere else.

To derive $VUQV = QV$ we first assume that the renumbering is such that $Q_\lambda = \left(\begin{smallmatrix} Q_E & Q_{ET} \\ Q_{TE} & Q_T \end{smallmatrix}\right)$ and $V = \left(\begin{smallmatrix} V_E & \mathbf{0} \\ V_{TE} & V_T \end{smallmatrix}\right)$. Condition 2 written in matrix form is now $V_E U_E \left(\begin{smallmatrix} Q_E & Q_{ET} \end{smallmatrix}\right) V = \left(\begin{smallmatrix} Q_E & Q_{ET} \end{smallmatrix}\right) V$, where U_E is a distributor matrix corresponding to (the collector matrix) V_E. Note that $Q = \Pi Q_\lambda \Pi = \Pi \left(\begin{smallmatrix} Q_E & Q_{ET} \\ \mathbf{0} & \mathbf{0} \end{smallmatrix}\right) \Pi$ and $\Pi V = \Pi \left(\begin{smallmatrix} V_E & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{smallmatrix}\right)$.

Finally, we assume that $\hat{Q}_\lambda = WQ_\lambda V$, $\hat{Q}_\tau = WQ_\tau V$, $\hat{Q}'_\lambda = W'Q_\lambda V$ and $\hat{Q}'_\tau = W'Q'_\tau V$, where V is the collector implied by $\mathcal{P}$ and W, W' are the corresponding (special) distributor matrices. We retain the same renumbering as above and write $W = \left(\begin{smallmatrix} W_E & W_T \end{smallmatrix}\right)$, $W' = \left(\begin{smallmatrix} W'_E & W'_T \end{smallmatrix}\right)$. That $\hat{Q}_\lambda = \hat{Q}'_\lambda$ follows from the fact that $W'_T = W_T$ and that $W \left(\begin{smallmatrix} Q_E & Q_{ET} \\ \mathbf{0} & \mathbf{0} \end{smallmatrix}\right) V = W' \left(\begin{smallmatrix} Q_E & Q_{ET} \\ \mathbf{0} & \mathbf{0} \end{smallmatrix}\right) V$. We show that $\hat{Q}'_\tau \sim \hat{Q}_\tau$ by considering $\hat{Q}_\tau[k,\ell] = \sum_{i \in C_k, j \in C_\ell} W[k,i] Q_\tau[i,j] V[j,\ell]$ and demonstrate that $\hat{Q}_\tau[k,\ell] = 0$ iff $Q_\tau[i,j] = 0$ for all $i \in C_k$, $j \in C_\ell$. ■

Now, if $(Q_\lambda, Q_\tau) \overset{\mathcal{P}}{\rightsquigarrow}_{\tau} (\hat{Q}_\lambda, \hat{Q}_\tau)$ we say that $(Q_\lambda, [Q_\tau]_\sim)$ $\tau_\sim$-lumps to $(\hat{Q}_\lambda, [\hat{Q}_\tau]_\sim)$ (with respect to $\mathcal{P}$) and denote it by $(Q_\lambda, [Q_\tau]_\sim) \overset{\mathcal{P}}{\rightsquigarrow}_{\tau_\sim} (\hat{Q}_\lambda, [\hat{Q}_\tau]_\sim)$.

We give an example of $\tau_\sim$-lumpings.

Example 10: Consider the Markov chains with silent steps depicted in Fig. 7. For each one of them we give a $\tau_\sim$-lumping and for each lumping class we show which option of Condition 1 of Definition 9 holds. The corresponding lumped Markov chains with silent steps are depicted in Fig. 8.

a) For the Markov chain with silent steps depicted in Fig. 7a the partitioning $\mathcal{P} = \{\{1,2\},\{3\}\}$ is a $\tau_\sim$-lumping. For the lumping class $\{1,2\}$ Condition 1a in Definition 9 is satisfied. For the class $\{3\}$ both Conditions 1a and 1b are satisfied.

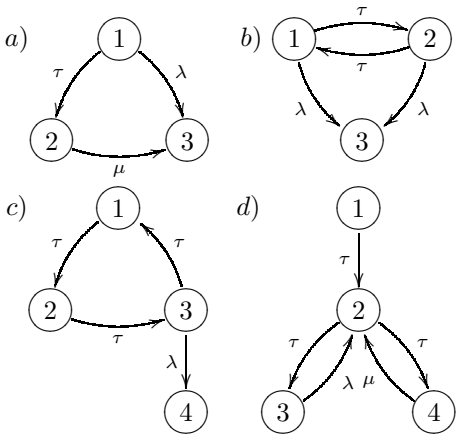


Fig. 7. Markov chains with silent steps with non-trivial $\tau_\sim$-lumpings – Example 10

b) For the Markov chain with silent steps in Fig. 7b $\mathcal{P} = \{\{1,2\},\{3\}\}$ is a $\tau_\sim$-lumping. For both lumping classes Conditions 1a and 1b are satisfied.

c) For the Markov chain with silent steps in Fig. 7c $\mathcal{P} = \{\{1,2\},\{3\},\{4\}\}$ is a $\tau_\sim$-lumping. For the lumping classes $\{1,2\}$ and $\{4\}$ both Conditions 1a and 1b are satisfied. For the class $\{3\}$ only Condition 1b is satisfied.

d) For the Markov chain with silent steps in Fig. 7d $\mathcal{P} = \{\{1,2\},\{3\},\{4\}\}$ is a $\tau_\sim$-lumping. For the classes $\{3\}$ and $\{4\}$ both Conditions 1a and 1b are satisfied. Since $\{1,2\}$ contains only transient states, for this class only Condition 1c is satisfied.

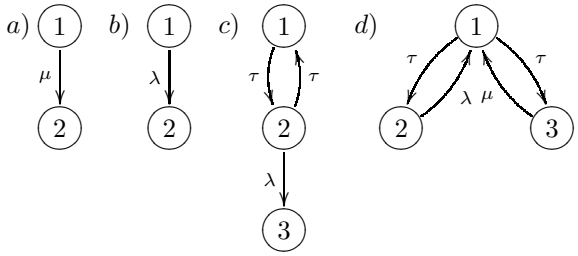


Fig. 8. $\tau_\sim$-lumped Markov chains with silent steps – Example 10

VI. Conclusions and Related Work

We presented a new approach to minimizing Markov chains with silent steps. We treated silent steps as exponentially distributed delays of which the rates tend to infinity. We extended the notion of ordinary lumping to the resulting (discontinuous) processes. Based on this theory, we provided a method for direct minimization of the original process, both, when the speed of going to infinity is given, and when it is not. The approach was illustrated in several examples which showed how the proposed definition corresponded to the intuition.

a) Related work: We discuss how our reduction technique is different to that of IMC's (when τ is the only possible action). First we do not allow silent steps to lead from a state to itself. However, as we treat them as exponential rates, they are redundant. Second, we give priority to silent steps over exponential delays only in transient states (see Example 10a) and not in ergodic states (see Example 9a). This leads to a different treatment of τ-divergence. For us, an infinite avoidance of an exponential delay is not possible. The transition must eventually be taken after an exponential delay (see Example 10b). This can be considered as some kind of fairness incorporated in the model. Third, due to the strong requirement that the lumping of Markov chains with silent steps is good if it is good for all possible speeds assigned to silent steps, our lumping does not always allow for joining states that lead to different ergodic classes (see Example 9b) unless these ergodic classes are also inside some lumping class. This means that we only disallow certain intermediate lumping steps. In all other cases, the weak bisimilarity of IMC's and $\tau_\sim$-lumping coincide.

Elimination of fast transitions in Markov processes is a subject in the field of perturbation theory. A perturbed Markov process is a Markov process in which some transitions (so-called *rare* transitions) are multiplied by a small number $\varepsilon > 0$. When considered on a time scale t/ε the perturbed process exhibits the same behavior as a Markov chain with fast transitions. Rare transitions become ordinary transitions and other transitions become fast transitions. To eliminate discontinuities in the model when $\varepsilon \to 0$, an aggregation method that eliminates all immediate transitions was introduced [18]. Later, this method was extended to all time scales [19], [11] leading to a hierarchy of simplified models. In [11], discontinuous Markov processes were used to clarify the presentation of ideas. Having another origin and motivation and not being based on lumpability, this aggregation method has several differences with our approach. First, intermediate lumping steps, i.e. steps that need not eliminate all silent steps left, like the one in Fig. 2b are not considered. Second, the focus is on eliminating only silent steps; nothing else is aggregated (contrary to joining the states 2 and 3 as in Fig. 2b). Third, the reduction can "split" states (they belong to multiple aggregation classes). This can be considered as a generalization of the lumping method but it is easily shown that it must not be allowed when lifting to Markov chains with silent steps. Fourth, it always gives a pure Markov process as a result (if, in Fig. 2a, we had ρ instead of one of the λ's, our lumping fails, while the aggregation technique does not). Fifth, to some extent, disaggregation to the exact original is possible. This is not true in our case but it is not a serious limitation if rewards are added to the model.

Fast transitions in Markov chains are also considered in other communities. An algorithm for removal of fast transitions in generalized stochastic Petri Nets is given in [20]. In [21] an algorithm for finding equilibrium probabilities in the presence of immediate transitions with known speed is developed. In [22] a reduction similar to the one in [18] for stiff Markov chains is given.

REFERENCES

[1] J. L. Doob, *Stochastic Processes*. Wiley, 1953.

[2] K. L. Chung, *Markov Chains with Stationary Probabilities*. Springer, 1967.

[3] J. G. Kemeny and J. L. Snell, *Finite Markov chains*. Springer, 1976.

[4] P. Buchholz, "Exact and ordinary lumpability in finite Markov chains," *Journal of Applied Probability*, vol. 31, pp. 59–75, 1994.

[5] J. P. Katoen and P. R. D'Argenio, "General distributions in process algebra," in *Lectures on formal methods and performance analysis: first EEF/Euro summer school on trends in computer science*, E. Brinksma, H. Hermanns, and J. Katoen, Eds. Springer, 2001, vol. 2090, pp. 375–429.

[6] M. Bravetti and P. R. D'Argenio, "Tutte le algebre insieme: Concepts, discussions and relations of stochastic process algebras with general distributions," in *Validation of Stochastic Systems - A Guide to Current Research*, ser. Lecture Notes of Computer Science, C. Baier, B. R. Haverkort, H. Hermanns, J. P. Katoen, and M. Siegle, Eds. Springer, 2004, vol. 2925, pp. 44–88.

[7] P. R. D'Argenio, "Algebras and automata for timed and stochastic systems," Ph.D. dissertation, University of Twente, 1999.

[8] M. Bravetti, "Real time and stochastic time," in *Formal Methods for the Design of Real-Time Systems*, ser. Lecture Notes of Computer Science, M. Bernardo and F. Corradini, Eds. Springer, 2004, vol. 3185, pp. 132–180.

[9] H. Hermanns, *Interactive Markov chains: The Quest for Quantified Quality*, ser. Lecture Notes of Computer Science. Springer, 2002, vol. 2428.

[10] J. Hillston, *A Compositional Approach to Performance Modelling*. Cambridge University Press, 1996.

[11] M. Coderch, A. Willsky, S. Sastry, and D. Castanon, "Hierarchical aggregation of singularly perturbed finite state Markov processes," *Stochastics*, vol. 8, pp. 259–289, 1983.

[12] J. Markovski and N. Trčka, "Lumping Markov chains with silent steps," Technische Universiteit Eindhoven, Tech. Rep. CS 06/13, 2006, Available from: `http://library.tue.nl/catalog/CSRPublication.csp`.

[13] V. Nicola, "Lumping in Markov reward processes," IBM, IBM Research Report RC 14719, 1989.

[14] E. Hille and R. S. Phillips, *Functional Analysis and Semi-Groups*. AMS, 1957.

[15] R. P. Agaev and P. Y. Chebotarev, "On determining the eigenprojection and components of a matrix," *Automated Remote Control*, vol. 63, pp. 1537–1545, 2002.

[16] S. L. Campbell, *Singular Systems of Differential Equations I*. Pitman, 1980.

[17] J. J. Koliha and T. D. Tran, "Semistable operators and singularly perturbed differential equations," *Journal of Mathematical Analysis and Applications*, vol. 231, pp. 446–458, 1999.

[18] F. Delebecque and J. P. Quadrat, "Optimal control of Markov chains admitting strong and weak interactions," *Automatica*, vol. 17, pp. 281–296, 1981.

[19] F. Delebecque, "A reduction process for perturbed Markov chains," *SIAM Journal of Applied Mathematics*, vol. 2, pp. 325–330, 1983.

[20] G. Ciardo, J. Muppala, and K. S. Trivedi, "On the solution of GSPN reward models," *Performance Evaluation*, vol. 12, pp. 237–253, 1991.

[21] W. K. Grassmann and Y. Wang, "Immediate events in Markov chains," in *Computations with Markov chains*, W. J. Stewart, Ed. Kluwer, 1995, pp. 163–176.

[22] A. Bobbio and K. S. Trivedi, "An aggregation technique for the transient analysis of stiff Markov chains," *IEEE Transactions on Computers*, vol. C-35, pp. 803–814, 1986.

VII. ACKNOWLEDGMENTS

We benefited a lot from our intensive and stimulating visit to Holger Hermanns at the Saarland University in Saarbrücken. We also thank our colleagues Jos Baeten, Bas Luttik and Erik de Vink for providing us with many useful comments.