Lumping Markov Chains with Silent Steps

Jasen Markovski, Nikola Trčka
Department of Mathematics and Computer Science,
Technische Universiteit Eindhoven,
P.O. Box 513, NL-5600 MB Eindhoven, The Netherlands

April 3, 2006

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.

1 Introduction

Markov chains (see e.g. [12, 6]) 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 processes 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 processes is called *lumping* [18, 4]. Lumping is a method based on the aggregation of states that exhibit the same behavior. It produces a smaller Markov process 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 [17, 3]). Although some of these algebras incorporate generally distributed stochastic delays (e.g. [9, 2]), the most widely used are the ones that restrict to exponential distributions (e.g. [14, 16]) due to the memoryless property. Typically, the employed model is some kind of extension of Markov processes 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 process, analysis is performed by standard techniques.

For the stochastic process algebra IMC (stands for *Interactive Markov chain*) [14], the extension of Markov processes 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.

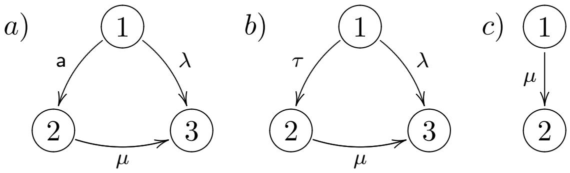


Figure 1: An IMC, its corresponding Markov chain with silent steps and the induced Markov chain

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.

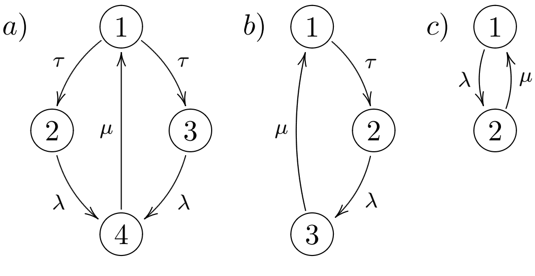


Figure 2: Three equivalent Markov chains with silent steps

Next, observe the process in Fig. 2a. In state 1 this process exhibits classical nondeterminism, 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* [8]. 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:

$$\begin{tikzcd}
\text{M.c. with f.t.} \arrow[r, "\tau\to\infty"'] \arrow[d, "\tau\textit{-lumping}"] & \text{M.p.} \arrow[d, "\begin{array}{c}\textit{ordinary}\\ \textit{lumping}\end{array}"] \\
\begin{array}{c}\tau\text{-lumped}\\ \text{M.c. with f.t.}\end{array} \arrow[r, "\tau\to\infty"'] & \begin{array}{c}\text{lumped}\\ \text{M.p.}\end{array}
\end{tikzcd}$$

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:

$$\begin{tikzcd}
\text{M.c. with f.t.} \arrow[r, phantom, "\sim"] \arrow[d, "\begin{array}{c}\textit{induced}\\ \tau\textit{-lumping}\end{array}"'] & \text{M.c. with f.t.} \arrow[d, "\begin{array}{c}\textit{induced}\\ \tau\textit{-lumping}\end{array}"] \\
\begin{array}{c}\tau\text{-lumped}\\ \text{M.c. with f.t.}\end{array} \arrow[r, phantom, "\sim"] & \begin{array}{c}\tau\text{-lumped}\\ \text{M.c. with f.t.}\end{array}
\end{tikzcd}$$

2 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 $\mathcal{P}=\big\{S\big\}$ and $\mathcal{P}=\big\{\{i\}\mid i\in S\big\}$ 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 $\mathcal{P}=\big\{S\big\}$ a distributor is a vector with elements that sum up to 1; for the trivial partitioning $\mathcal{P}=\big\{\{i\}\mid i\in S\big\}$ there exists only one distributor (I).

Example 1 Let $\mathcal{S}=\{1,2,3\}$ and $\mathcal{P}=\big\{\{1,2\},\{3\}\big\}$. 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 $\mathcal{P}$.

Let $\mathcal{P}_1=\{C_1,\ldots,C_N\}$ be a partitioning of $\mathcal{S}$ and let $\mathcal{P}_2=\{D_1,\ldots,D_M\}$ be a partitioning of $\mathcal{P}_1$. The *composition* of $\mathcal{P}_1$ and $\mathcal{P}_2$ is a partitioning of $\mathcal{S}$ defined as:

$$\mathcal{P}_1\circ\mathcal{P}_2=\{L_1,\ldots,L_M\},\ L_i=\bigcup_{C\in D_i}C.$$

Example 2 Let $\mathcal{S}=\{1,\ldots,6\}$ and let $\mathcal{P}_1=\big\{\{1,2\},\{3,4,5\},\{6\}\big\}$ and $\mathcal{P}_2=\Big\{\big\{\{1,2\}\big\},\big\{\{3,4,5\},\{6\}\big\}\Big\}$. Then $\mathcal{P}_1\circ\mathcal{P}_2=\big\{\{1,2\},\{3,4,5,6\}\big\}$.

Note that $V_{\mathcal{P}_1\circ\mathcal{P}_2}=V_{\mathcal{P}_1}V_{\mathcal{P}_2}$.

3 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 [20].

3.1 Markov Processes

A Markov process is a finite-state continuous-time stochastic process that is homogeneous and satisfies the Markov property [12, 6]. 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 [12].

Example 3 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 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 $(\Rightarrow)$ Assume that (Π, Q) satisfies 1–4 and let $P(t) = \Pi e^{Qt}$. We check the three conditions of Definition 2.

1. Let $c \geq 0$ be such that $Q + c\Pi \geq 0$. Note that matrices $c\Pi$ and $Q + c\Pi$ commute. This implies that
$$\Pi e^{Qt} = \Pi e^{-c\Pi t + (Q+c\Pi)t} = \Pi e^{-c\Pi t} e^{(Q+c\Pi)t}.$$
Clearly, $e^{(Q+c\Pi)t} \geq 0$ so, to prove that $P(t) \geq 0$, it remains to show that $\Pi e^{-c\Pi t} \geq 0$.

   From $\Pi^2 = \Pi$ it easily follows that $\Pi^n = \Pi$ for all $n \geq 1$. Then,
$$\begin{aligned}
\Pi e^{-c\Pi t} &= \Pi \cdot \sum_{n=0}^{\infty} \frac{(-ct)^n \Pi^n}{n!} \\
&= \Pi \cdot (I + \sum_{n=1}^{\infty} \frac{(-ct)^n \Pi^n}{n!}) \\
&= \Pi \cdot (I + \sum_{n=1}^{\infty} \frac{(-ct)^n \Pi}{n!}) \\
&= \Pi \cdot (I + \Pi \cdot (e^{-ct} - 1)) \\
&= \Pi \cdot e^{-ct} \geq 0.
\end{aligned}$$

2. For the second condition, we have
$$\begin{aligned}
P(t) \cdot \mathbf{1} &= \Pi e^{Qt} \cdot \mathbf{1} \\
&= \Pi \cdot \sum_{n=0}^{\infty} \frac{Q^n t^n}{n!} \cdot \mathbf{1} \\
&= \Pi \cdot (I + \sum_{n=1}^{\infty} \frac{Q^n t^n}{n!}) \cdot \mathbf{1} \\
&= \Pi \cdot (\mathbf{1} + \sum_{n=1}^{\infty} \frac{(Q^n \cdot \mathbf{1}) t^n}{n!}) \\
&= \Pi \cdot (\mathbf{1} + \mathbf{0}) \\
&= \mathbf{1}.
\end{aligned}$$

3. From $\Pi Q = Q\Pi$ it follows that $\Pi Q^n = Q^n \Pi$ for all $n \geq 0$. Using this, we derive
$$\begin{aligned}
\Pi e^{Qt} &= \Pi \cdot \sum_{n=0}^{\infty} \frac{Q^n t^n}{n!} \\
&= \sum_{n=0}^{\infty} \frac{Q^n t^n}{n!} \cdot \Pi \\
&= e^{Qt}\Pi.
\end{aligned}$$
   Thus,
$$\begin{aligned}
P(t) \cdot P(s) &= \Pi e^{Qt} \Pi e^{Qs} \\
&= \Pi^2 e^{Qt} e^{Qs} \\
&= \Pi e^{Q(t+s)} \\
&= P(t+s).
\end{aligned}$$

($\Leftarrow$) For the proof of the opposite direction suppose $P(t)$ is a transition matrix. Define

$$\Pi = \lim_{t\to 0} P(t) \quad \text{and} \quad Q = \lim_{h\to 0} \frac{P(h) - \Pi}{h}.$$

Now, it is not hard to check that Conditions 1–4 hold. See [8, 15] for the proof that the above limits exist and for the uniqueness proof. ■

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 $\Pi = I$ [12, 6]. 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 [8]:

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

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 4 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 $\mu_i = \begin{pmatrix}1\end{pmatrix}$ for all $i = 1,2,3$ and $\delta_1 = \begin{pmatrix}p\end{pmatrix}$, $\delta_2 = \begin{pmatrix}1-p\end{pmatrix}$ and $\delta_3 = \begin{pmatrix}0\end{pmatrix}$. 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.

3.2 Ordinary Lumping

We now define a notion of lumping for Markov processes.

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 actually says that the rows of ΠV (resp. QV) that correspond to the states that belong to the same class must be equal [18]. It does not depend on the particular choice of the non-zero elements of U. To prove this, 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, $VU'QV = QV$.

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 We show that the four conditions of Theorem 1 hold.

1. Since $U \geq 0$, $V \geq 0$ and $\Pi \geq 0$, we have $\hat{\Pi} = U\Pi V \geq 0$. Also,

$$\begin{aligned} \hat{\Pi} \cdot \mathbf{1} &= U\Pi V \cdot \mathbf{1} \\ &= U\Pi \cdot \mathbf{1} \\ &= U \cdot \mathbf{1} \\ &= \mathbf{1} \end{aligned}$$

and, since $VU\Pi V = \Pi V$,

$$\begin{aligned} \hat{\Pi}^2 &= U\Pi VU\Pi V \\ &= U\Pi\Pi V \\ &= U\Pi V \\ &= \hat{\Pi}. \end{aligned}$$

2. For the second condition we have

$$\begin{aligned} \hat{\Pi}\hat{Q} &= U\Pi VUQV \\ &= U\Pi QV \\ &= UQV \\ &= \hat{Q} \end{aligned} \qquad \text{and} \qquad \begin{aligned} \hat{Q}\hat{\Pi} &= UQVU\Pi V \\ &= UQ\Pi V \\ &= UQV \\ &= \hat{Q}. \end{aligned}$$

3. We calculate,

$$\begin{aligned} \hat{Q} \cdot \mathbf{1} &= UQV \cdot \mathbf{1} \\ &= UQ \cdot \mathbf{1} \\ &= U \cdot \mathbf{0} \\ &= \mathbf{0}. \end{aligned}$$

4. Let c be such that $Q + c\Pi \geq 0$. Then

$$\hat{Q} + c\hat{\Pi} = UQV + cU\Pi V = U(Q + c\Pi)V \geq 0. \qquad \blacksquare$$

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 [20].

Example 5 a. Let (Π, Q) be the Markov process from Example 4a. 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 4b. 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 It is not hard to show (e.g. by induction on n) that $\Pi Q^n = Q^n$ and $VUQ^nV = Q^nV$

for all $n \geq 1$. Using this,

$$\begin{aligned}
VUP(t)V &= VU\Pi e^{Qt}V \\
&= VU\Pi \sum_{n=0}^{\infty} \frac{Q^n t^n}{n!} V \\
&= VU\Pi \left(I + \sum_{n=1}^{\infty} \frac{Q^n t^n}{n!} \right) V \\
&= VU\Pi V + \sum_{n=1}^{\infty} \frac{VU\Pi Q^n V t^n}{n!} \\
&= \Pi V + \sum_{n=1}^{\infty} \frac{VUQ^n V t^n}{n!} \\
&= \Pi V + \sum_{n=1}^{\infty} \frac{Q^n V t^n}{n!} \\
&= \Pi V + \sum_{n=1}^{\infty} \frac{\Pi Q^n V t^n}{n!} \\
&= \Pi \left(I + \sum_{n=1}^{\infty} \frac{Q^n t^n}{n!} \right) V \\
&= \Pi \sum_{n=0}^{\infty} \frac{Q^n t^n}{n!} V \\
&= \Pi e^{Qt} V = P(t)V. \qquad \blacksquare
\end{aligned}$$

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 By induction, it is not hard to prove $(UQV)^n = UQ^nV$ for all $n \geq 0$. Using this and that $VUQ^nV = Q^nV$ for all $n \geq 1$, we have the following derivation:

$$\begin{array}{rcl}
\hat{P}(t) & = & \hat{\Pi}e^{\hat{Q}t} \\
& = & U\Pi V e^{UQVt} \\
& = & U\Pi V \sum_{n=0}^{\infty} \frac{(UQV)^n t^n}{n!} \\
& = & U\Pi V \sum_{n=0}^{\infty} \frac{UQ^nVt^n}{n!} \\
& = & U\Pi \sum_{n=0}^{\infty} \frac{VUQ^nVt^n}{n!} \\
& = & U\Pi \sum_{n=0}^{\infty} \frac{Q^nVt^n}{n!} \\
& = & U\Pi (\sum_{n=0}^{\infty} \frac{Q^nt^n}{n!})V \\
& = & U\Pi e^{Qt}V \\
& = & UP(t)V.
\end{array}$$ ■

We finish this section by giving some relational properties of ordinary lumping. The relation 'lumps to' is clearly reflexive (set $U = V = I$). We show that it is also transitive.

Theorem 5 (Transitivity of ordinary lumping) Suppose $(\Pi, Q) \overset{\mathcal{P}_1}{\rightsquigarrow} (\hat{\Pi}, \hat{Q})$ and $(\hat{\Pi}, \hat{Q}) \overset{\mathcal{P}_2}{\rightsquigarrow} (\check{\Pi}, \check{Q})$. Then $(\Pi, Q) \overset{\mathcal{P}_1 \circ \mathcal{P}_2}{\rightsquigarrow} (\check{\Pi}, \check{Q})$.

Proof Let V_i and U_i be respectively the collector and a distributor matrix associated with $\mathcal{P}_i$, $i \in \{1, 2\}$. Let $U = U_2U_1$ and $V = V_1V_2$. Recall that V and U are the collector and a distributor for $\mathcal{P}_1 \circ \mathcal{P}_2$. From $(\Pi, Q) \overset{\mathcal{P}_1}{\rightsquigarrow} (\hat{\Pi}, \hat{Q})$ we have $V_1U_1\Pi V_1 = \Pi V_1$ and $\hat{\Pi} = U_1\Pi V_1$. From $(\hat{\Pi}, \hat{Q}) \overset{\mathcal{P}_2}{\rightsquigarrow} (\check{\Pi}, \check{Q})$, we have $V_2U_2\hat{\Pi}V_2 = \hat{\Pi}V_2$ and $\check{\Pi} = U_2\hat{\Pi}V_2$. Then,

$$\begin{array}{rcl}
VU\Pi V & = & V_1V_2U_2U_1\Pi V_1V_2 \\
& = & V_1V_2U_2\hat{\Pi}V_2 \\
& = & V_1\hat{\Pi}V_2 \\
& = & V_1U_1\Pi V_1V_2 \\
& = & \Pi V_1V_2 \\
& = & \Pi V
\end{array}$$

and

$$\begin{array}{rcl}
\check{\Pi} & = & U_2\hat{\Pi}V_2 \\
& = & U_2U_1\Pi V_1V_2 \\
& = & U\Pi V.
\end{array}$$

Similarly, $VUQV = QV$ and $\check{Q} = UQV$. ■

4 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.

4.1 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 [12] that the limit always exists; moreover it is known (see [1] 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 $\text{rank}(\Pi) + \text{rank}(Q) = n$. 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 6 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 [5] for the first proof, or [19] for a proof written in more modern terms. See [8] 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 [12]) 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 $i \to j$.

Example 6 a. Consider a Markov chain with fast transitions (Q_λ, Q_τ) depicted in Fig. 3a. It is defined with

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

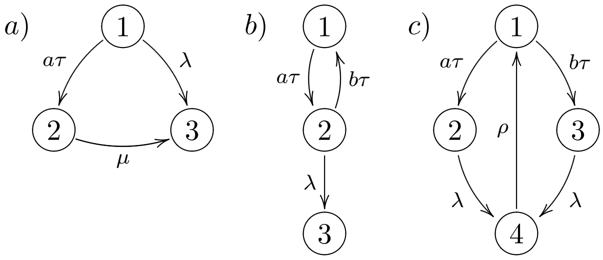


Figure 3: Markov chains with fast transitions from Example 6

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 4a (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 4b (for $p = \frac{a}{a+b}$ and $\lambda = \mu$).

4.2 τ-lumping

We now define a special notion of lumping for Markov chains with fast transitions introduced in the previous section. 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 7 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

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

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 5b 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

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

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

For the second partitioning we obtain

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

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

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).

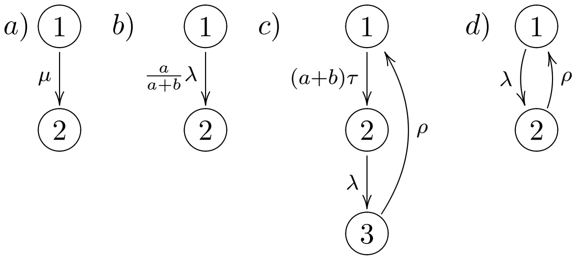


Figure 4: τ-lumped Markov chains with fast transitions – Example 7

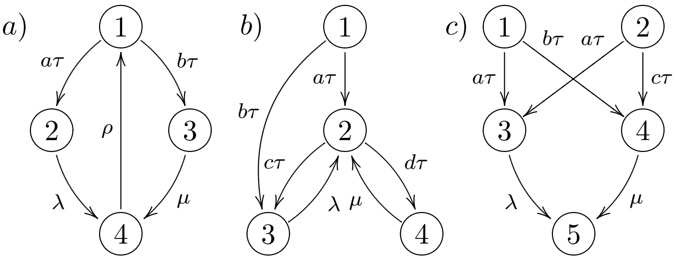


Figure 5: Markov chains with fast transitions without non-trivial τ-lumpings – Example 8

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.

Definition 6 induces the following diagram:

$$\begin{tikzcd}
(Q_\lambda, Q_\tau) \arrow[r, "\tau\to\infty"'] \arrow[d, "\tau\text{-}lumping"] & (\Pi, Q) \arrow[d, "\begin{array}{l}ordinary\\ lumping\end{array}"] \\
(\hat{Q}_\lambda, \hat{Q}_\tau) & (\hat{\Pi}, \hat{Q})
\end{tikzcd} .$$

For the definition to be sound, we have to show that the diagram can be closed, i.e. that

$$(\hat{Q}_\lambda, \hat{Q}_\tau) \xrightarrow[\tau\to\infty]{} (\hat{\Pi}, \hat{Q}) .$$

We first show some properties of the matrix W. For that we need a more refined numbering of states. The required renumbering is based on the following lemma. The lemma 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 Let Π be the ergodic projection of Q_τ. Let the numbering be such that makes the ergodic partitioning explicit and that, in each ergodic class, states in a lumping class with a lower index precede states in a lumping class with a higher index. Then, Π_i is a square matrix of the following form:

$$\Pi_i = \mathbf{1} \cdot \mu_i = \mathbf{1} \cdot (\mu_i^{(1)} \ldots \mu_i^{(N)})$$

where, for some $1 \leq k \leq N$, $\mu_i^{(k)}$ is the restriction of the vector μ_i to the elements that belong in C_k. Note that it can be the empty vector.

Let V be the collector matrix associated with $\mathcal{P}$; then

$$V = \begin{pmatrix} V_1 \\ \vdots \\ V_M \\ \cdots \end{pmatrix} \quad \text{and} \quad V_i = \operatorname{diag}\left(\mathbf{1}^{|E_i \cap C_1|}, \ldots, \mathbf{1}^{|E_i \cap C_N|}\right),\ 1 \leq i \leq M$$

(transient states are not important for this lemma).

Define, for every $1 \leq i \leq M$, a row vector σ_i as:

$$\sigma_i[k] = \begin{cases} \mu_i^{(k)} \cdot \mathbf{1}, & E_i \cap C_k \neq \emptyset \\ 0, & E_i \cap C_k = \emptyset \end{cases}$$

for $1 \leq k \leq n$. Note that $\sigma_i[k] > 0$ iff $E_i \cap C_k \neq \emptyset$. Then

$$\Pi V = \begin{pmatrix} \Pi_1 V_1 & \mathbf{0} \\ \vdots & \vdots \\ \Pi_M V_M & \mathbf{0} \\ \cdots & \end{pmatrix} \quad \text{and} \quad \Pi_i V_i = \mathbf{1} \cdot \sigma_i,\ 1 \leq i \leq M.$$

Suppose now $E_i \cap C_k \neq \emptyset$, $E_i \cap C_\ell \neq \emptyset$ and $E_j \cap C_k \neq \emptyset$ for some $1 \leq i, j \leq M$ and $1 \leq k, \ell \leq N$. This implies that $\sigma_i[k] > 0$, $\sigma_i[\ell] > 0$ and $\sigma_j[k] > 0$. By the lumping condition all rows of $\Pi_i V_i$ and $\Pi_j V_j$ that correspond to the class C_k must be equal. Since both $\Pi_i V_i$ and $\Pi_j V_j$, are matrices that consist of equal rows, we have that $\sigma_i[\ell] = \sigma_i[k] = \sigma_j[k] = \sigma_j[\ell]$. Therefore, $\sigma_j[\ell] > 0$. We conclude that $E_j \cap C_\ell \neq \emptyset$. ■

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.

We give an example of such renumbering.

Example 9 Consider the Markov chain with fast transitions depicted in Fig. 6a. Its ergodic partitioning is $\mathcal{E} = \{E_1, E_2, E_3, T\}$ where $E_1 = \{2,5\}$, $E_2 = \{6,8\}$, $E_3 = \{4,7\}$ and $T = \{1,3\}$ (note that the ergodic classes can also be numbered differently). Let $\mathcal{P} = \{C_1, C_2, C_3, C_4\}$ where $C_1 = \{1\}$, $C_2 = \{2,4\}$, $C_3 = \{5,7\}$ and $C_4 = \{3,6,8\}$. It is directly checked that $\mathcal{P}$ is a τ-lumping. Note that the ergodic classes E_1 and E_3 share states from the lumping classes C_2 and C_3 and that E_2 shares states only with C_4. So, $L = 3$ and $S = 2$. Note that the transient state 3 lumps together with the ergodic states 6 and 8, and that the transient state 1 lumps alone. We renumber ergodic and lumping classes as $E_1 \mapsto E_{11}$, $E_3 \mapsto E_{12}$, $C_2 \mapsto C_{11}$, $C_3 \mapsto C_{12}$, $E_2 \mapsto E_{21}$, $C_4 \mapsto C_{21}$ and $C_1 \mapsto C_3$. Then, we renumber states as $2 \mapsto 1$, $5 \mapsto 2$, $4 \mapsto 3$, $7 \mapsto 4$, $6 \mapsto 5$, $8 \mapsto 6$, $3 \mapsto 7$, and $1 \mapsto 8$. The new Markov chain with fast transitions is depicted in 6b.

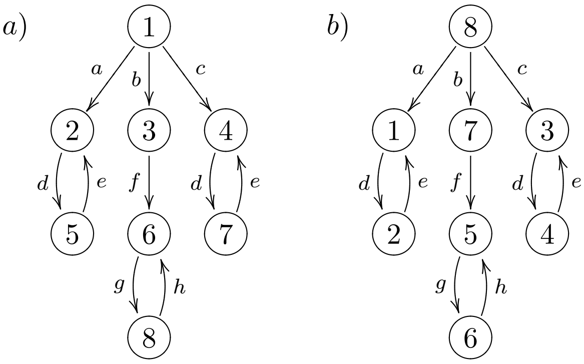


Figure 6: Markov chain with fast transitions before and after the renumbering of states – Example 9

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

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

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

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:

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

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

Let $\mu_{ij}^{(k)}$ denote the restriction of μ_{ij} to the elements of C_{ik}. Then

$$\Pi_i V_i = \begin{pmatrix} \Pi_{i1} V_{i1} \\ \vdots \\ \Pi_{ie_i} V_{ie_i} \end{pmatrix} = \begin{pmatrix} \mathbf{1}^{|E_{i1}|} \cdot \left(\mu_{i1}^{(1)} \cdot \mathbf{1} \ \dots \ \mu_{i1}^{(c_i)} \cdot \mathbf{1}\right) \\ \vdots \\ \mathbf{1}^{|E_{ie_i}|} \cdot \left(\mu_{ie_i}^{(1)} \cdot \mathbf{1} \ \dots \ \mu_{ie_i}^{(c_i)} \cdot \mathbf{1}\right) \end{pmatrix}.$$

By the lumpability condition, rows of $\Pi_i V_i$ that correspond to the same partitioning class are equal. This implies that

$$\mu_{ij}^{(\ell)} \cdot \mathbf{1} = \mu_{ik}^{(\ell)} \cdot \mathbf{1},$$

for all $1 \le j, k \le e_i$, $1 \le \ell \le c_i$. Define a row vector ρ_i as

$$\rho_i[\ell] = \mu_{ij}^{(\ell)} \cdot \mathbf{1}$$

(for any $1 \le j \le e_i$). Then

$$\mu_{ij} V_{ij} = \rho_i, \text{ for any } 1 \le j \le e_i, \text{ and } \Pi_i V_i = \mathbf{1} \cdot \rho_i.$$

The matrix W of Definition 6 has the following form:

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

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

where

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

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 To prove that $\Pi V W \Pi = \Pi V W$ it suffices to show that

$$X_i V_i W_i \Pi_i = X_i V_i W_i \text{ for all } X_i \in \{\Pi_i, \overline{\Pi}_i, \widetilde{\Pi}_i\},\ (1 \le i \le S).$$

This is equivalent to

$$\mu_{ij} V_{ij} W_{ik} \Pi_{ik} = \mu_{ij} V_{ij} W_{ik},\ (1 \le j, k \le e_i)$$

and that to

$$\rho_i W_{ik} \Pi_{ik} = \rho_i W_{ik},\ (1 \le k \le e_i).$$

We calculate

$$\rho_i W_{ik} = \frac{1}{e_i} \cdot (\rho_i[1] \dots \rho_i[c_i]) \cdot \operatorname{diag}\left(\frac{\mu_{ik}^{(1)}}{\rho_i[1]}, \dots, \frac{\mu_{ik}^{(c_i)}}{\rho_i[c_i]} \right) = \frac{1}{e_i} \cdot \mu_{ik},$$

and so

$$\begin{array}{rcl} \rho_i W_{ik} \Pi_{ik} & = & \dfrac{1}{e_i} \cdot \mu_{ik} \cdot \mathbf{1} \cdot \mu_{ik} \\ & = & \dfrac{1}{e_i} \cdot 1 \cdot \mu_{ik} \\ & = & \rho_i W_{ik}. \end{array}$$

■

We write $(Q_\lambda, Q_\tau) \overset{\mathcal{P}}{\rightsquigarrow}_\tau (\hat{Q}_\lambda, \hat{Q}_\tau)$ when $\mathcal{P}$ is a τ-lumping of the Markov chain with fast transitions (Q_λ, Q_τ) and when (Q_λ, Q_τ) τ-lumps to $(\hat{Q}_\lambda, \hat{Q}_\tau)$ with respect to $\mathcal{P}$.

We are now ready for the soundness proof.

Theorem 7 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 By Theorem 6, we have to show that $\hat{\Pi}$ is the ergodic projection of $\hat{Q}_\tau$ and that $\hat{\Pi}\hat{Q}_\lambda\hat{\Pi} = \hat{Q}$.

For the second part, using Lemma 2, we have the following derivation:

$$\begin{array}{rcl} \hat{\Pi}\hat{Q}_\lambda\hat{\Pi} & = & U\Pi V W Q_\lambda V U \Pi V \\ & = & U\Pi V W \Pi Q_\lambda \Pi V \\ & = & U\Pi\Pi Q_\lambda \Pi V \\ & = & U\Pi Q_\lambda \Pi V \\ & = & UQV \\ & = & \hat{Q}. \end{array}$$

By Theorems 1 and 2 we have that $\hat{\Pi} \geq 0$, $\hat{\Pi} \cdot \mathbf{1} = \mathbf{1}$ and that $\hat{\Pi}^2 = \hat{\Pi}$.

We also derive $\hat{\Pi}\hat{Q}_{\mathcal{T}} = U\Pi V W Q_{\mathcal{T}} V = U\Pi V W \Pi Q_{\mathcal{T}} V = \mathbf{0}$ since $\Pi Q_{\mathcal{T}} = \mathbf{0}$. Similarly, $\hat{Q}_{\mathcal{T}}\hat{\Pi} = W Q_{\mathcal{T}} V U \Pi V = W Q_{\mathcal{T}} \Pi V = \mathbf{0}$ since $Q_{\mathcal{T}}\Pi = \mathbf{0}$.

Now we prove that $\operatorname{rank}(\hat{\Pi}) + \operatorname{rank}(\hat{Q}_{\mathcal{T}}) = N$.

First, we compute $\hat{\Pi}$:

$$\hat{\Pi} = W\Pi V = \begin{pmatrix} W_1\Pi_1V_1 & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & W_2\Pi_2V_2 & \dots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & W_S\Pi_SV_S & \mathbf{0} \\ \widetilde{W}\widetilde{\Pi}_1V_1 & \widetilde{W}\widetilde{\Pi}_2V_2 & \dots & \widetilde{W}\widetilde{\Pi}_SV_S & \mathbf{0} \end{pmatrix}$$

$$\text{where} \quad W_i\Pi_iV_i = W_i \cdot \mathbf{1} \cdot \rho_i = \mathbf{1} \cdot \rho_i$$

Since $\hat{\Pi}$ is idempotent, its rank is equal to its trace and so:

$$\begin{aligned} \operatorname{rank}(\hat{\Pi}) &= \operatorname{trace}(\hat{\Pi}) \\ &= \sum_{i=1}^{S} \operatorname{trace}(W_i\Pi_iV_i) \\ &= \sum_{i=1}^{S} \operatorname{trace}(\mathbf{1} \cdot \rho_i) \\ &= S \cdot 1 \\ &= S. \end{aligned}$$

We now show that $\operatorname{rank}(\hat{Q}_{\mathcal{T}}) = N - S$.

Note that a generator is called *irreducible* if there does not exist a renumbering after which it is represented as $\left(\begin{smallmatrix} A' & A'' \\ \mathbf{0} & B \end{smallmatrix}\right)$ for some (non-empty) square matrices A' and B. It is known (cf. [11]) that, in a numbering that makes the ergodic partitioning of $(Q_\lambda, Q_{\mathcal{T}})$ explicit (and our numbering is just a more refined one), $Q_{\mathcal{T}}$ has the following form:

$$Q_{\mathcal{T}} = \begin{pmatrix} Q_1 & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & Q_2 & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & Q_S & \mathbf{0} & \mathbf{0} \\ \overline{Q}_1 & \overline{Q}_2 & \dots & \overline{Q}_S & \overline{Q} & \overline{Q}' \\ \widetilde{Q}_1 & \widetilde{Q}_2 & \dots & \widetilde{Q}_S & \widetilde{Q} & \widetilde{Q}' \end{pmatrix} \qquad Q_i = \operatorname{diag}\left(Q_{i1}, \dots, Q_{ie_i}\right),$$

where Q_{ij} are irreducible and $\left(\begin{smallmatrix} \overline{Q} & \overline{Q}' \\ \widetilde{Q} & \widetilde{Q}' \end{smallmatrix}\right)$ cannot be further reduced (after any renumbering) to $\left(\begin{smallmatrix} \overline{Q}'' & \mathbf{0} \\ \mathbf{0} & \widetilde{Q}'' \end{smallmatrix}\right)$ such that $\overline{Q}''$ is an (irreducible) generator matrix.

We compute $\hat{Q}_\mathcal{T}$:

$$\hat{Q}_\mathcal{T} = WQ_\mathcal{T}V = \begin{pmatrix} W_1Q_1V_1 & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & W_2Q_2V_2 & \dots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & W_SQ_SV_S & \mathbf{0} \\ \widetilde{W}\begin{pmatrix} \widetilde{Q}_1V_1 \\ + \\ \widetilde{Q}\overline{V}_1 \end{pmatrix} & \widetilde{W}\begin{pmatrix} \widetilde{Q}_2V_2 \\ + \\ \widetilde{Q}\overline{V}_2 \end{pmatrix} & \dots & \widetilde{W}\begin{pmatrix} \widetilde{Q}_SV_S \\ + \\ \widetilde{Q}\overline{V}_S \end{pmatrix} & \widetilde{W}\widetilde{Q}'\widetilde{V} \end{pmatrix}$$

and

$$W_iQ_iV_i = \sum_{j=1}^{e_i} W_{ij}Q_{ij}V_{ij}.$$

Since by Theorem 2 $\hat{Q}_\mathcal{T}$ is a generator matrix, to show that $\text{rank}(\hat{Q}_\mathcal{T}) = N - S$ it is sufficient to show that, for all $1 \leq i \leq S$, $W_iQ_iV_i$ is an irreducible generator matrix and that $\widetilde{W}\widetilde{Q}'\widetilde{V}$ cannot be represented in a form $\left(\begin{smallmatrix} G & \mathbf{0} \\ \mathbf{0} & X \end{smallmatrix}\right)$, where G is an irreducible generator matrix.

That $W_iQ_iV_i$ is a generator follows directly from the form of $\hat{Q}_\mathcal{T}$. We prove by contradiction that $W_iQ_iV_i$ is also irreducible. Note that the matrix $W_iQ_iV_i$ is a restriction of $\hat{Q}_\mathcal{T}$ on the states $C_{i1}, \dots, C_{ic_i}$. Let us assume that $W_iQ_iV_i$ is not an irreducible matrix. Then we can number the states such that $W_iQ_iV_i = \left(\begin{smallmatrix} \hat{A} & \hat{B} \\ \mathbf{0} & \hat{D} \end{smallmatrix}\right)$, for some square matrices $\hat{A}$ and $\hat{D}$.

Using the same numbering (but now the states C_{ij} are classes of states) we obtain the following forms of W_{ij}, Q_{ij} and V_{ij}:

$$W_{ij} = \begin{pmatrix} W'_{ij} & \mathbf{0} \\ \mathbf{0} & W''_{ij} \end{pmatrix} \qquad Q_{ij} = \begin{pmatrix} A_{ij} & B_{ij} \\ C_{ij} & D_{ij} \end{pmatrix} \qquad V_{ij} = \begin{pmatrix} V'_{ij} & \mathbf{0} \\ \mathbf{0} & V''_{ij} \end{pmatrix},$$

where A_{ij}, D_{ij} are square matrices for $1 \leq j \leq c_i$. We compute:

$$W_iQ_iV_i = \sum_{j=1}^{e_i} W_{ij}Q_{ij}V_{ij} = \sum_{j=1}^{e_i} \begin{pmatrix} W'_{ij}A_{ij}V'_{ij} & W'_{ij}B_{ij}V''_{ij} \\ W''_{ij}C_{ij}V'_{ij} & W''_{ij}D_{ij}V''_{ij} \end{pmatrix}.$$

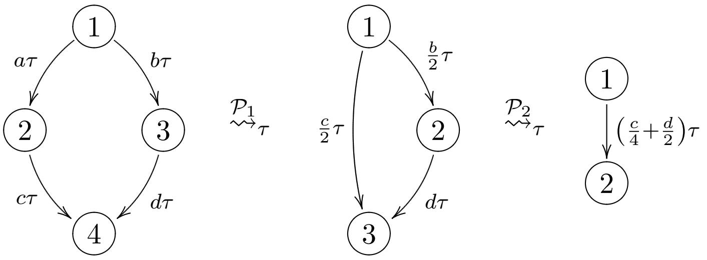
Clearly, one obtains $\sum_{j=1}^{e_i} W''_{ij}C_{ij}V'_{ij} = \mathbf{0}$. The matrix $W_iQ_iV_i$ is a generator, so $C_{ij} \geq 0$. However, W_{ij} has a positive diagonal and V_{ij} is a collector matrix, so we conclude that $C_{ij} = \mathbf{0}$. Contradiction, because Q_{ij} is an irreducible matrix for every $1 \leq j \leq c_i$. Thus, $W_iQ_iV_i$ is irreducible.

Assume now that there exists a numbering in which $\widetilde{W}\widetilde{Q}'\widetilde{V} = \left(\begin{smallmatrix} G & \mathbf{0} \\ \mathbf{0} & X \end{smallmatrix}\right)$ and G is an irreducible generator matrix. Similarly as in the previous proof we conclude that in this numbering $\widetilde{Q}' = \left(\begin{smallmatrix} A & \mathbf{0} \\ \mathbf{0} & B \end{smallmatrix}\right)$, for some square matrices A and B. To obtain the contradiction we need to show that A is a generator.

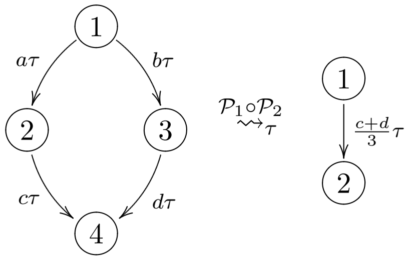
Because $\hat{Q}_\mathcal{T}$ is a generator matrix, $\widetilde{W}\widetilde{Q}_iV_i + \widetilde{W}\widetilde{Q}\overline{V}_i = \left(\begin{smallmatrix} \mathbf{0} \\ X_i \end{smallmatrix}\right)$, for all $1 \leq i \leq S$ and some matrix X_i. Note that $\widetilde{Q}_i$, $\widetilde{Q} \geq \mathbf{0}$ and $\widetilde{W}$ is a distributor matrix for the collector matrices V_i and $\overline{V}_i$. We conclude that $\widetilde{Q}$ and $\widetilde{Q}_i$ have the form $\widetilde{Q} = \left(\begin{smallmatrix} \mathbf{0} \\ Y \end{smallmatrix}\right)$ and $\widetilde{Q}_i = \left(\begin{smallmatrix} \mathbf{0} \\ Y_i \end{smallmatrix}\right)$. Now, because $Q_\mathcal{T}$ is a generator matrix we conclude that A is a generator matrix. ■

We now consider to what extent τ-lumping is transitive. In general, τ-lumping is not transitive. Consider the following example.

Example 10 Let $\mathcal{S} = \{1, \ldots, 4\}$, $\mathcal{P}_1 = \big\{\{1,2\},\{3\},\{4\}\big\}$ and $\mathcal{P}_2 = \big\{\{\{1,2\},\{3\}\},\{\{4\}\big\}$. We have


1
aτ
bτ
2
3
cτ
dτ
4
$\overset{\mathcal{P}_1}{\rightsquigarrow}_\tau$
1
$\frac{b}{2}\tau$
$\frac{c}{2}\tau$
2
dτ
3
$\overset{\mathcal{P}_2}{\rightsquigarrow}_\tau$
1
$(\frac{c}{4}+\frac{d}{2})\tau$
2

but


1
aτ
bτ
2
3
cτ
dτ
4
$\overset{\mathcal{P}_1\circ\mathcal{P}_2}{\rightsquigarrow}_\tau$
1
$\frac{c+d}{3}\tau$
2

The resulting Markov chains with fast transitions are not identical but they have the same limit. If one loosens the criterion for transitivity and only considers infinite behavior, τ-lumping becomes transitive. This is shown in the following theorem.

Theorem 8 (Transitivity of $\boldsymbol{\tau}$-lumping) Suppose $(Q_\lambda, Q_\tau) \overset{\mathcal{P}_1}{\rightsquigarrow}_\tau (\hat{Q}_\lambda, \hat{Q}_\tau)$, $(\hat{Q}_\lambda, \hat{Q}_\tau) \overset{\mathcal{P}_2}{\rightsquigarrow}_\tau (\check{Q}_\lambda, \check{Q}_\tau)$ and $(\check{Q}_\lambda, \check{Q}_\tau) \longrightarrow_\infty (\check{\Pi}, \check{Q})$. Then

$$(Q_\lambda, Q_\tau) \overset{\mathcal{P}_1\circ\mathcal{P}_2}{\rightsquigarrow}_\tau (\check{Q}'_\lambda, \check{Q}'_\tau)$$

for some Markov chain with fast transitions $(\check{Q}'_\lambda, \check{Q}'_\tau)$ such that

$$(\check{Q}'_\lambda, \check{Q}'_\tau) \longrightarrow_\infty (\check{\Pi}, \check{Q}).$$

Proof That $\mathcal{P}_1 \circ \mathcal{P}_2$ is a good τ-lumping follows directly from Theorem 5. What needs to be shown is that lumping directly with the composed partitioning results in a Markov chain with fast transitions that has the same limit as $(\check{Q}_\lambda, \check{Q}_\tau)$.

Let $V = V_1V_2$. Assume that $\check{Q}'_i = WQ_iV$ for $i \in \{\lambda, \tau\}$ and let $(\check{Q}'_\lambda, \check{Q}'_\tau) \longrightarrow_\infty (\check{\Pi}', \check{Q}')$. We show that $\check{\Pi}' = \check{\Pi}$ and $\check{Q}' = \check{Q}$.

Clearly $\check{\Pi}' = W\Pi V = U\Pi V = \check{\Pi}$. By Lemma 2, $\Pi VW\Pi = \Pi VW$, so we have

$$\begin{aligned}
\check{Q}' &= \check{\Pi}'\check{Q}_\lambda\check{\Pi}' \\
&= \check{\Pi}\check{Q}_\lambda\check{\Pi} \\
&= U\Pi VWQ_\lambda VU\Pi V \\
&= U\Pi VW\Pi Q_\lambda \Pi V \\
&= U\Pi^2 Q_\lambda \Pi V \\
&= U\Pi Q_\lambda \Pi V \\
&= \check{Q}.
\end{aligned}$$

■

Remark 1 It is not hard, only notationally quite cumbersome, to show that, if there are no partitioning classes that contain only transient states, the notion of τ-lumping is transitive also up to isomorphism.

5 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.

5.1 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 $1 \leq i, j \leq n$, $A[i,j] = 0$ iff $B[i,j] = 0$.

Example 11 For $a, b, c \neq 0$, 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.

5.2 $\tau_\sim$-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 12 a. Consider the Markov chain with silent steps depicted in Fig. 7a. The Example 7b shows that the partitioning $\mathcal{P} = \big\{\{1,2\},\{3\}\big\}$ 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. 7b. The Example 8c shows, that although for some speeds the partitioning $\big\{\{1,2\},\{3\},\{4\}\big\}$ is a τ-lumping, it need not be so for some other speeds.

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

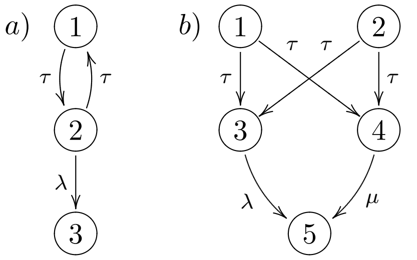


Figure 7: Markov chains with silent steps – Example 12

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 $i \in \{1, \ldots, n\}$, $\mathrm{erg}(i) = \{j \in \bigcup_{1\le k\le M} E_k \mid i \twoheadrightarrow j\}$ be the set of all ergodic states reachable from the state i. Let for all $C \in \mathcal{P}$, $\mathrm{erg}(C)$ denote $\bigcup_{i\in C} \mathrm{erg}(i)$. 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) $\mathrm{erg}(C) \subseteq D$, for some $D \in \mathcal{P}$.

   (b) $\mathrm{erg}(C) = E_i$, 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 $C \neq D$,

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

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 9 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 We first show that $\mathcal{P}$ is a τ-lumping for all (Q_λ, Q'_τ), where $Q'_\tau \sim Q_\tau$. Let $(Q_\lambda, Q'_\tau) \to_\infty (\Pi, Q)$. We have to show that $VU\Pi V = \Pi V$ and $VUQV = QV$. Recall that $VU\Pi V = \Pi V$ is equivalent to the condition that the rows of ΠV that correspond to the states that belong to the same partitioning class are equal. Note that this is the same as saying that, for all $C, D \in \mathcal{P}$, the vector $\Pi^{(C,D)} \cdot \mathbf{1}$ has all elements equal ($\Pi^{(C,D)}$ denotes the restriction of Π to the elements of C row-wise and the elements of D column-wise). By E we denote the set of ergodic states, i.e. $E = \bigcup_{i=1}^{M} E_i$, where E_i are the ergodic classes. We refer to [12] for the property that $j \in \mathrm{erg}(i)$ iff $\Pi[i, j] > 0$.

Consider $\Pi^{(C,D)}$ for some $C, D \in \mathcal{P}$. We distinguish three cases according to the conditions of Definition 9.

Assume that Condition 1a holds. Suppose $i \in C \cap E$. Since $\mathrm{erg}(i) \subseteq D$ and $i \in \mathrm{erg}(i)$, we obtain that $i \in D$. Because $\mathcal{P}$ is a partitioning, it follows that $C = D$ and $\mathrm{erg}(i) \subseteq C$. From $\mathrm{erg}(i) \subseteq C$, it follows that $\Pi^{(\{i\},F)} = \mathbf{0}$ for all $F \neq C$. Recall that Π is a stochastic matrix, so $\Pi^{(\{i\},C)} \cdot \mathbf{1} = \mathbf{1}$. Now, assume that $i \in C \cap T$. From $\mathrm{erg}(i) \subseteq D$ we conclude that $\Pi^{(\{i\},F)} = \mathbf{0}$ for all $F \neq D$. Let $K \subseteq \{1, \ldots, M\}$ be such that $E_k \subseteq D$ for all $k \in K$. Then, $\sum_{k \in K} \delta_k[i] = 1$, because i is trapped only in the ergodic classes contained in D. One calculates $\Pi^{(\{i\},D)} \cdot \mathbf{1} = \sum_{k \in K} \delta_k[i] \cdot \mu_k \cdot \mathbf{1} = \sum_{k \in K} \delta_k[i] \cdot 1 = 1$. We conclude that if Condition 1a of Definition 9 holds then either $\Pi^{(C,F)} \cdot \mathbf{1} = \mathbf{1}$ or $\Pi^{(C,F)} \cdot \mathbf{1} = \mathbf{0}$, for every $F \in \mathcal{P}$, so $\Pi^{(C,F)} \cdot \mathbf{1}$ always has equal elements.

Next, assume that Condition 1b of Definition 9 holds. Assume that $i \in C \cap E$ and $\mathrm{erg}(C) = E_j$, for some $1 \leq j \leq M$. Then one concludes that $i \in E_j$, so $C \subseteq E_j$ and $C = D$. Thus, the restriction $\Pi^{(\{i\},D)}$ presents a part of an ergodic vector for all $i \in C$, so $\Pi^{(C,D)} = \Pi^{(C,C)} = \mathbf{1} \cdot \mu_j^{(D)}$. Note that $\Pi^{(C,F)} = \mathbf{0}$ for $F \cap E_j = \emptyset$. Now, assume that $i \in C \cap T$. Since $\mathrm{erg}(i) = E_j$, one concludes that $\delta_k[i] = 0$, for $k \neq j$ because the transient state i can only be trapped the ergodic class E_j. As Π is a stochastic matrix, $\delta_j[i] = 1$. Thus, $\Pi^{(C,D)} = \mathbf{1} \cdot (1 \cdot \mu_j^{(D)}) = \Pi^{(D,D)}$. Note that $\Pi^{(C,F)} = \mathbf{0}$ for $F \cap E_j = \emptyset$. We conclude that if Condition 1b of Definition 9 holds then either $\Pi^{(C,F)} = \Pi^{(F,F)} = \mathbf{1} \cdot \mu_j^{(F)}$ or $\Pi^{(C,F)} = \mathbf{0}$. In the first case $\Pi^{(C,F)} \cdot \mathbf{1} = \mathbf{1} \cdot (\mu_j^{(F)} \cdot \mathbf{1})$ and in the second $\Pi^{(C,F)} \cdot \mathbf{1} = \mathbf{0}$, so all elements of $\Pi^{(C,F)} \cdot \mathbf{1}$ are equal.

Finally, assume that Condition 1c of Definition 9 holds. Since there is only one state $i \in C$ such that $i \to j$ and $j \notin C$ and all the states in C are transient, we conclude that the trapping probabilities of i are equal to the trapping probabilities of all other states in C. More precisely, a transient state must be trapped in an ergodic class, so for all states $k \in C$ it must hold that $k \twoheadrightarrow \ell$, for some $\ell \notin C$. The only way to do this is by $k \twoheadrightarrow i \to j' \twoheadrightarrow \ell$, where $j' \notin C$ and $j' \twoheadrightarrow \ell$. Thus, all states have the same trapping probabilities, so $\Pi^{(C,D)} = \mathbf{1} \cdot x$ for some row vector $x \neq \mathbf{0}$. We conclude that if Condition 1c of Definition 9 holds then $\Pi^{(C,F)} \cdot \mathbf{1} = \mathbf{1} \cdot (x \cdot \mathbf{1})$, for every $F \in \mathcal{P}$, so again all elements of $\Pi^{(C,D)} \cdot \mathbf{1}$ are equal.

We conclude that $VU\Pi V = \Pi V$ holds.

To show that $VUQV = QV$, let numbering be such that it makes the division between ergodic and transient states explicit. Note that Condition 2 of Definition 9 imposes the lumping condition only on ergodic states. In order to use matrix manipulation, we rewrite it in matrix form using the following form of Q_λ and V:

$$V = \begin{pmatrix} V_E & \mathbf{0} \\ V_{TE} & V_T \end{pmatrix} \qquad Q_\lambda = \begin{pmatrix} Q_E & Q_{ET} \\ Q_{TE} & Q_T \end{pmatrix}.$$

Now, Condition 2 of Definition 9 can be rewritten in matrix form as:

$$V_E U_E \left(Q_E \; Q_{ET} \right) V = \left(Q_E \; Q_{ET} \right) V,$$

where U_E is a distributor matrix corresponding to (the collector matrix) V_E.

Note that

$$Q = \Pi Q_\lambda \Pi = \Pi \left(\begin{smallmatrix} Q_E & Q_{ET} \\ \mathbf{0} & \mathbf{0} \end{smallmatrix}\right) \Pi \qquad \text{and} \qquad \Pi V = \Pi \left(\begin{smallmatrix} V_E & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{smallmatrix}\right).$$

Then using Condition 2 of Definition 9 and $VU\Pi V = \Pi V$ we compute:

$$\begin{array}{rcl}
VUQV & = & VU\Pi Q_\lambda \Pi V \\
& = & VU\Pi \left(\begin{smallmatrix} Q_E & Q_{ET} \\ \mathbf{0} & \mathbf{0} \end{smallmatrix}\right) \Pi V \\
& = & VU\Pi \left(\begin{smallmatrix} Q_E & Q_{ET} \\ \mathbf{0} & \mathbf{0} \end{smallmatrix}\right) VU\Pi V \\
& = & VU\Pi \left(\begin{smallmatrix} V_E U_E Q_E & V_E U_E Q_{ET} \\ \mathbf{0} & \mathbf{0} \end{smallmatrix}\right) VU\Pi V \\
& = & VU\Pi \left(\begin{smallmatrix} V_E & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{smallmatrix}\right) \left(\begin{smallmatrix} U_E Q_E & U_E Q_{ET} \\ \mathbf{0} & \mathbf{0} \end{smallmatrix}\right) VU\Pi V \\
& = & VU\Pi V \left(\begin{smallmatrix} U_E Q_E & U_E Q_{ET} \\ \mathbf{0} & \mathbf{0} \end{smallmatrix}\right) VU\Pi V \\
& = & \Pi V \left(\begin{smallmatrix} U_E Q_E & U_E Q_{ET} \\ \mathbf{0} & \mathbf{0} \end{smallmatrix}\right) VU\Pi V \\
& = & \Pi Q_\lambda \Pi V \\
& = & QV.
\end{array}$$

Next, we show that $\hat{Q}_\lambda = \hat{Q}'_\lambda$. We assume that the distributor matrices are W, W' such that $\hat{Q}_\lambda = WQ_\lambda V$, $\hat{Q}_\mathcal{T} = WQ_\mathcal{T}V$, $\hat{Q}'_\lambda = W'Q_\lambda V$ and $\hat{Q}'_\mathcal{T} = W'Q'_\mathcal{T}V$, where V is the collector implied by $\mathcal{P}$. The matrices W and W' have the following form:

$$W = \left(\begin{smallmatrix} W_E & W_T \end{smallmatrix}\right) \qquad W' = \left(\begin{smallmatrix} W'_E & W'_T \end{smallmatrix}\right)$$

By Definitions 6 and 9,

$$W'_T = W_T \quad \text{and} \quad W \left(\begin{smallmatrix} Q_E & Q_{ET} \\ \mathbf{0} & \mathbf{0} \end{smallmatrix}\right) V = W' \left(\begin{smallmatrix} Q_E & Q_{ET} \\ \mathbf{0} & \mathbf{0} \end{smallmatrix}\right) V.$$

We compute:

$$\begin{array}{rcl}
\hat{Q}_\lambda & = & WQ_\lambda V \\
& = & W \begin{pmatrix} Q_E & Q_{ET} \\ Q_{TE} & Q_T \end{pmatrix} V \\
& = & W \left(\begin{smallmatrix} Q_E & Q_{ET} \\ \mathbf{0} & \mathbf{0} \end{smallmatrix}\right) V + W \left(\begin{smallmatrix} \mathbf{0} & \mathbf{0} \\ Q_{TE} & Q_T \end{smallmatrix}\right) V \\
& = & W' \left(\begin{smallmatrix} Q_E & Q_{ET} \\ \mathbf{0} & \mathbf{0} \end{smallmatrix}\right) V + \left(\begin{smallmatrix} W_E & W_T \end{smallmatrix}\right) \left(\begin{smallmatrix} \mathbf{0} & \mathbf{0} \\ Q_{TE} & Q_T \end{smallmatrix}\right) V \\
& = & W' \left(\begin{smallmatrix} Q_E & Q_{ET} \\ \mathbf{0} & \mathbf{0} \end{smallmatrix}\right) V + \left(\begin{smallmatrix} W'_E & W'_T \end{smallmatrix}\right) \left(\begin{smallmatrix} \mathbf{0} & \mathbf{0} \\ Q_{TE} & Q_T \end{smallmatrix}\right) V \\
& = & W' \left(\begin{smallmatrix} Q_E & Q_{ET} \\ \mathbf{0} & \mathbf{0} \end{smallmatrix}\right) V + W' \left(\begin{smallmatrix} \mathbf{0} & \mathbf{0} \\ Q_{TE} & Q_T \end{smallmatrix}\right) V \\
& = & W'Q_\lambda V \\
& = & \hat{Q}'_\lambda.
\end{array}$$

Finally, we show that $\hat{Q}'_\mathcal{T} \sim \hat{Q}_\mathcal{T}$. We observe $\hat{Q}_\mathcal{T}[k, \ell]$. It is computed as:

$$\hat{Q}_\mathcal{T}[k, \ell] = \sum_{i \in C_k, j \in C_\ell} W[k, i] Q_\mathcal{T}[i, j] V[j, \ell] = \sum_{i \in C_k, j \in C_\ell} W[k, i] Q_\mathcal{T}[i, j]$$

Thus, $\hat{Q}_\mathcal{T}[i, j] = 0$, iff $\sum_{i \in C_k, j \in C_\ell} W[k, i] Q_\mathcal{T}[i, j] = 0$. If $k \neq \ell$ then $Q_\mathcal{T}[i, j] \geq 0$ and $W[k, i] > 0$, for all $i \in C_k, j \in C_\ell$. We conclude that $\hat{Q}_\mathcal{T}[i, j] = 0$ iff $Q_\mathcal{T}[i, j] = 0$, for all $i \in C_k, j \in C_\ell$. If $k = \ell$ then $Q_\mathcal{T}[i, i] = -\sum_{j \in C, C \in \mathcal{P}} Q_\mathcal{T}[i, j]$, so the sum is equal to zero iff $Q_\mathcal{T}[i, j] = 0$ for $j \notin C_k$. Since, in both cases, the conditions for $\hat{Q}_\mathcal{T}[i, j] = 0$ depend only on the grammar of $Q_\mathcal{T}$ and $Q_\mathcal{T} \sim Q'_\mathcal{T}$, we conclude that $\hat{Q}_\mathcal{T} \sim \hat{Q}'_\mathcal{T}$, which completes the proof. ■

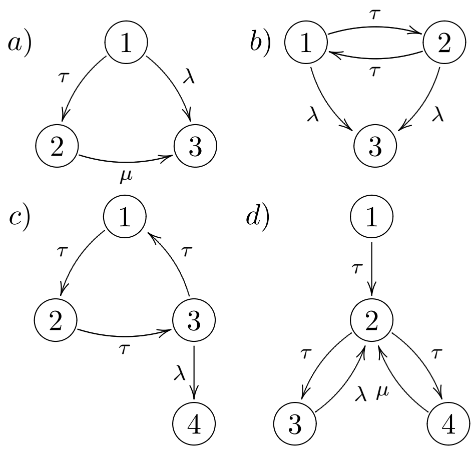


Figure 8: Markov chains with silent steps with non-trivial $\tau_\sim$-lumpings – Example 13

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 13 Consider the Markov chains with silent steps depicted in Fig. 8. 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. 9.

a. For the Markov chain with silent steps depicted in Fig. 8a the partitioning $\mathcal{P} = \big\{\{1,2\},\{3\}\big\}$ 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.

b. For the Markov chain with silent steps in Fig. 8b $\mathcal{P} = \big\{\{1,2\},\{3\}\big\}$ 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. 8c $\mathcal{P} = \big\{\{1,2\},\{3\},\{4\}\big\}$ 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. 8d $\mathcal{P} = \big\{\{1,2\},\{3\},\{4\}\big\}$ 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.

To finalize the section, we prove that $\tau_\sim$-lumping is also transitive.

Theorem 10 (Transitivity of $\tau_\sim$-lumping) Suppose $(Q_\lambda, [Q_\tau]_\sim) \overset{\mathcal{P}_1}{\rightsquigarrow}_{\tau_\sim} (\hat{Q}_\lambda, [\hat{Q}_\tau]_\sim)$ and $(\hat{Q}_\lambda, [\hat{Q}_\tau]_\sim) \overset{\mathcal{P}_2}{\rightsquigarrow}_{\tau_\sim} (\check{Q}_\lambda, [\check{Q}_\tau]_\sim)$. Then $(Q_\lambda, [Q_\tau]_\sim) \overset{\mathcal{P}_1 \circ \mathcal{P}_2}{\rightsquigarrow}_{\tau_\sim} (\check{Q}_\lambda, [\check{Q}_\tau]_\sim)$.

Proof From $(Q_\lambda, [Q_\tau]_\sim) \overset{\mathcal{P}_1}{\rightsquigarrow}_{\tau_\sim} (\hat{Q}_\lambda, [\hat{Q}_\tau]_\sim)$, we have $(Q_\lambda, Q_\tau) \overset{\mathcal{P}_1}{\rightsquigarrow}_\tau (\hat{Q}_\lambda, \hat{Q}'_\tau)$ where $\hat{Q}'_\tau \sim \hat{Q}_\tau$. From $(\hat{Q}_\lambda, [\hat{Q}_\tau]_\sim) \overset{\mathcal{P}_2}{\rightsquigarrow}_{\tau_\sim} (\check{Q}_\lambda, [\check{Q}_\tau]_\sim)$ and $\hat{Q}'_\tau \sim \hat{Q}_\tau$, we have $(\hat{Q}_\lambda, \hat{Q}'_\tau) \overset{\mathcal{P}_2}{\rightsquigarrow}_\tau (\check{Q}_\lambda, \check{Q}'_\tau)$ where $\check{Q}'_\tau \sim \check{Q}_\tau$. From

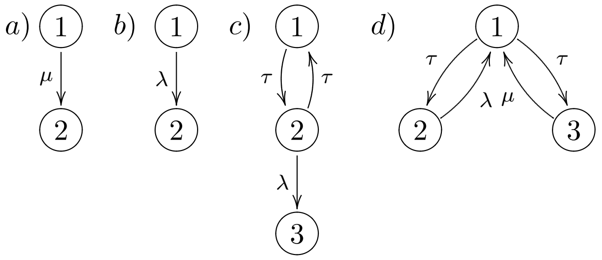


Figure 9: $\tau_{\sim}$-lumped Markov chains with silent steps – Example 13

the transitivity of τ-lumping and Theorem 9, $(Q_\lambda, Q_\tau) \overset{\mathcal{P}_1 \circ \mathcal{P}_2}{\rightsquigarrow}_\tau (\check{Q}_\lambda, \check{Q}''_\tau)$ for some $\check{Q}''_\tau \sim \check{Q}'_\tau \sim \check{Q}_\tau$. Therefore, $(Q_\lambda, [Q_\tau]_\sim) \overset{\mathcal{P}_1 \circ \mathcal{P}_2}{\rightsquigarrow}_{\tau_\sim} (\check{Q}_\lambda, [\check{Q}_\tau]_\sim)$.

6 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.

Related work We discuss how our reduction technique is different to that of IMC's. 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 13a) and not in ergodic states (see Example 12a). 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 13b). 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 12b) unless these ergodic classes are also inside some lumping class. This means that we only disallow certain intermediate lumping steps.

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 [11]. Later, this method was extended to all time scales [10, 8] leading to a hierarchy of simplified models. In [8], 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 the Petri Nets community. An algorithm for removal of fast transitions in generalized stochastic Petri Nets is given in [7]. In [13] an algorithm for finding equilibrium probabilities in the presence of immediate transitions with known speed is developed.

References

[1] R. P. Agaev and P. Yu. Chebotarev. On determining the eigenprojection and components of a matrix. *Automated Remote Control*, 63:1537–1545, 2002.

[2] M. Bravetti. Real time and stochastic time. In M. Bernardo and F. Corradini, editors, *Formal Methods for the Design of Real-Time Systems*, volume 3185 of *Lecture Notes of Computer Science*, pages 132–180. Springer, 2004.

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

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

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

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

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

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

[9] P. R. D'Argenio. *Algebras and Automata for Timed and Stochastic Systems.* PhD thesis, University of Twente, 1999.

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

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

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

[13] W. K. Grassmann and Y. Wang. Immediate events in Markov chains. In W. J. Stewart, editor, *Computations with Markov chains*, pages 163–176. Kluwer, 1995.

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

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

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

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

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

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

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