Handling state space explosion in Markov chains

How lumping introduces imprecision (almost) inevitably

Alexander Erreygers
SMACS Research Group, Ghent University

Jasper De Bock
IDLab, Ghent University-imec

Markov chains

An introduction

Consider a generic continuous-time stochastic process $(X_t)_{t\in\mathbb{R}_{\geq 0}}$, where for all $t\in\mathbb{R}_{\geq 0}$ the state X_t is a random variable that takes values x in the finite state space $\mathscr{X}$. We provide $\mathscr{X}$ with some ordering, such that any real-valued function f on $\mathscr{X}$ can be identified with a row vector. We furthermore let $\mathscr{L}(\mathscr{X})$ denote the set of all real-valued functions on $\mathscr{X}$. Then any linear operator $T\colon \mathscr{L}(\mathscr{X})\to\mathscr{L}(\mathscr{X})$ can be identified with a matrix.

Precise Markov chains

The stochastic process $(X_t)_{t\in\mathbb{R}_{\geq 0}}$ is a *precise (continuous-time) Markov chain* (pMC) if it satisfies the *Markov property*: where $n\geq 0$ is an integer and $\{t_1,\dots,t_n,s,t\}$ is a strictly increasing sequence of non-negative time points. The *transition matrix* T_s^t thus defined satisfies

$$\begin{aligned}[T_s^t f](x_s) &= \mathrm{E}(f(X_t)|X_s=x_s)\\ &= \mathrm{E}(f(X_t)|X_{t_1}=x_1,\dots,X_{t_n}=x_n,X_s=x_s).\end{aligned} \tag{P1}$$

A pMC is called *stationary* if it satisfies $T_t^{t+\Delta}=T_0^\Delta =: T_\Delta$ for all $t,\Delta\in\mathbb{R}_{\geq 0}$. In this case, there is a unique *transition rate matrix* Q—a matrix with non-negative off-diagonal elements and rows that sum up to zero—such that

$$(\forall t\in\mathbb{R}_{\geq 0})\ T_\Delta = T_t^{t+\Delta}\approx I+\Delta Q \quad \text{for } \Delta \text{ suff. small.}$$

Furthermore, T_t then satisfies the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}T_t = QT_t, \qquad \text{with } T_0 = I. \tag{P2}$$

Similarly, for any non-stationary pMC there is a time-dependent transition rate matrix Q_t such that

$$(\forall t\in\mathbb{R}_{\geq 0})\ T_t^{t+\Delta}\approx I+\Delta Q_t \quad \text{for } \Delta \text{ suff. small.}$$

Imprecise Markov chains

It is often infeasible to precisely specify the transition rate matrix Q of a stationary pMC. Furthermore, assuming stationarity is not always justified. Therefore, we here consider the case where the (time-dependent) transition rate matrix Q_t of a (non-stationary) pMC is only known to be contained in some (non-empty and bounded) set $\mathscr{Q}$. In other words, we consider the set $\mathbb{P}_{\mathscr{Q}}$ of all pMCs that are consistent with $\mathscr{Q}$, in the sense that

$$(\forall t\in\mathbb{R}_{\geq 0})(\exists Q_t\in\mathscr{Q})\ T_t^{t+\Delta}\approx I+\Delta Q_t \quad \text{for } \Delta \text{ suff. small.}$$

This set $\mathbb{P}_{\mathscr{Q}}$ characterises an *imprecise (continuous-time) Markov chain* (iMC) as follows. Analogous to (P1), we define a *lower transition operator* $\underline{T}_s^t$ as

$$\begin{aligned}[\underline{T}_s^t f](x_s) &:= \underline{\mathrm{E}}(f(X_t)|X_s=x_s)\\ &= \underline{\mathrm{E}}(f(X_t)|X_{t_1}=x_1,\dots,X_{t_n}=x_n,X_s=x_s),\end{aligned} \tag{I1}$$

where $\underline{\mathrm{E}}(\cdot|\cdot)$ is the minimum of the conditional expectations that are induced by the set of consistent processes.

In case $\mathscr{Q}$ has separately specified rows, Krak et al. (2017) show that $\underline{T}_t^{t+\Delta}=\underline{T}_0^\Delta =: \underline{T}_\Delta$ for all $t,\Delta\in\mathbb{R}_{\geq 0}$. Moreover, they show that $\underline{T}_\Delta$ is the unique operator that satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}\underline{T}_t = \underline{Q}\,\underline{T}_t, \qquad \text{with } \underline{T}_0 = I. \tag{I2}$$

In (I2), $\underline{Q}$ is the so-called *lower transition rate operator* of $\mathscr{Q}$, which, for any $f\in\mathscr{L}(\mathscr{X})$ and $x\in\mathscr{X}$, is defined as

$$[\underline{Q}f](x) := \min\{[Qf](x)\colon Q\in\mathscr{Q}\}. \tag{I3}$$

Ergodicity

We are often interested in the long-term limit behaviour of stationary pMCs and iMCs. For iMCs, a special case is when

$$\lim_{t\to+\infty}[\underline{T}_t f](x) = \underline{\mathrm{E}}_\infty(f) \quad \text{for all } f\in\mathscr{L}(\mathscr{X}) \text{ and } x\in\mathscr{X}.$$

If this is the case, then the iMC is said to be *ergodic* and $\underline{\mathrm{E}}_\infty(f)$ is called the *limit lower expectation*. Similarly, a stationary pMC is ergodic if

$$\lim_{t\to+\infty}[T_t f](x) = \mathrm{E}_\infty(f) \quad \text{for all } f\in\mathscr{L}(\mathscr{X}) \text{ and } x\in\mathscr{X},$$

where E_∞ is now called the limit expectation.

State space explosion

Precise Markov chains (or pMCs, as introduced in **Markov chains: An introduction**) are used ubiquitously to model systems with uncertain dynamics. Consider a stationary pMC and assume that we are interested in making inferences of the form

$$\lim_{t\to+\infty}\mathrm{E}(f(X_t)) = \lim_{t\to+\infty}\sum_{x\in\mathscr{X}}\pi_0(x)\mathrm{E}(f(X_t)|X_0=x) = \lim_{t\to+\infty}\sum_{x\in\mathscr{X}}\pi_0(x)[T_t f](x),$$

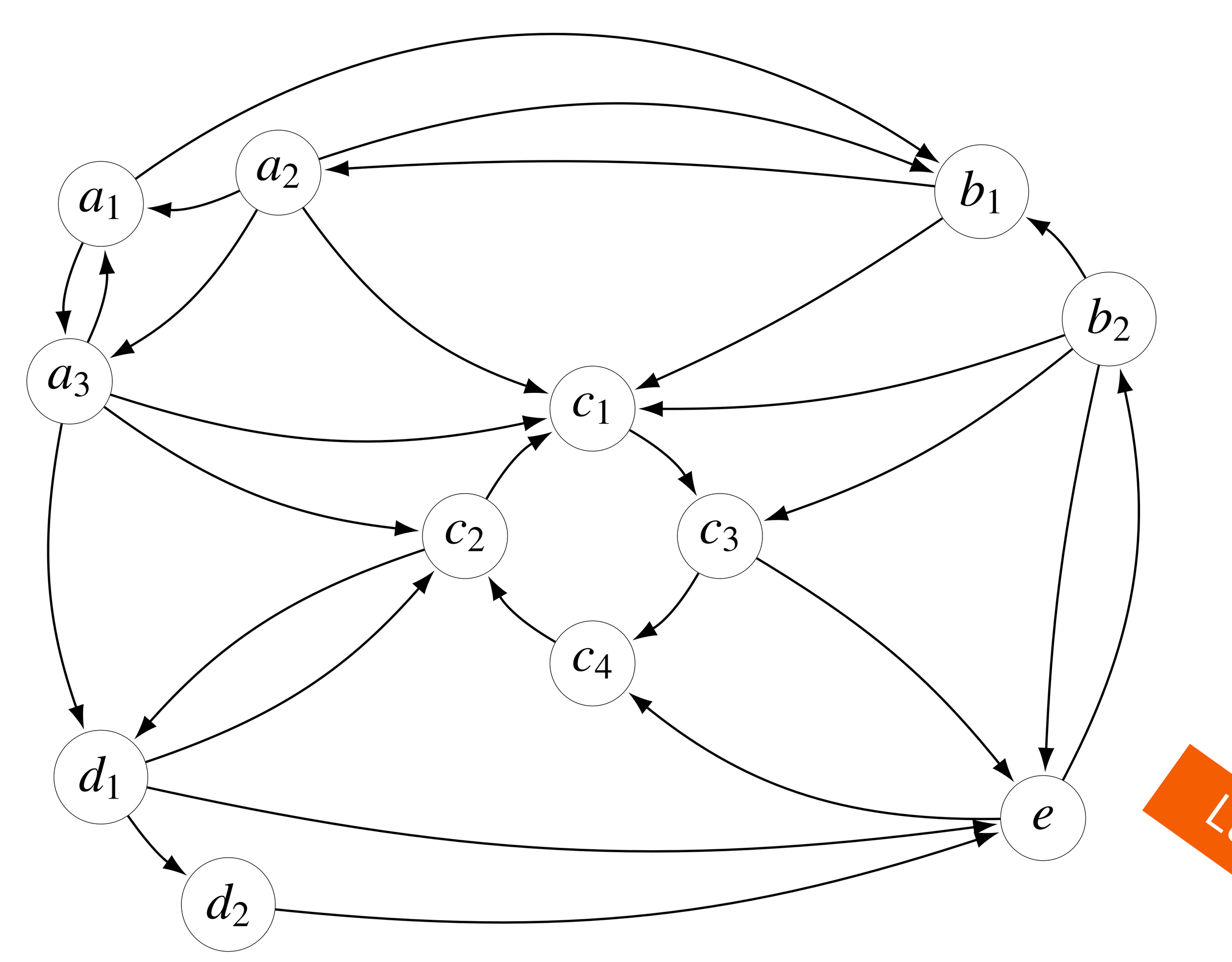
where f is a real-valued function on $\mathscr{X}$ and π_0 is an initial probability distribution. If the pMC is ergodic, then we immediately obtain that

$$\lim_{t\to+\infty}\mathrm{E}(f(X_t)) = \mathrm{E}_\infty(f) = \sum_{x\in\mathscr{X}}\pi_\infty(x)f(x),$$

regardless of π_0. It is well-know that π_∞ is the unique probability distribution on $\mathscr{X}$ that satisfies the equilibrium condition

$$\pi_\infty Q = 0. \tag{1}$$

In case the number of states $|\mathscr{X}|$ is relatively small, this linear system of equations can be efficiently solved analytically or numerically. However, in many applications—see for instance **Modelling spectrum assignment in a two-service flexi-grid optical link**—the number of states grows **exponentially** with respect to the dimensions of the system! **This state space explosion makes** (1) **practically unsolvable for large systems.**



Lumping reduces the number of states

One way to reduce the number of states is to ***lump*** together states. For example, in **Modelling spectrum assignment in a two-service flexi-grid optical link** we lump together states that correspond to the same higher-order description. In any case, this lumping of states yields the *lumped state space* $\hat{\mathscr{X}}$, which is a partition of $\mathscr{X}$.

The lumped stochastic process $(\hat{X}_t)_{t\in\mathbb{R}_{\geq 0}}$, which has state space $\hat{\mathscr{X}}$, is derived from the original stochastic process $(X_t)_{t\in\mathbb{R}_{\geq 0}}$ using the relation

$$(\forall \hat{x}\in\hat{\mathscr{X}})\ \hat{X}_t = \hat{x} \Leftrightarrow X_t\in\hat{x}.$$

Throughout this poster, we only consider real-valued functions on $\mathscr{X}$ that are constant over the elements of the partition, as such a function f can be trivially identified with a real-valued function $\hat{f}$ on $\hat{\mathscr{X}}$.

Consider a stationary and ergodic pMC with state space $\mathscr{X}$ and transition rate matrix Q. Assume, for the sake of simplicity, that the pMC is *irreducible*, in the sense that $\mathrm{P}(X_t=x)>0$ for all $t>0$ and all $x\in\mathscr{X}$. Given an initial distribution π_0 for the original pMC , we find that the lumped process $(\hat{X}_t)_{t\in\mathbb{R}_{\geq 0}}$ is a pMC with (time-dependent) transition rate matrix

$$\hat{Q}_t(\hat{x},\hat{y}) = \frac{\sum_{x\in\hat{x}}\mathrm{P}(X_t=x)\sum_{y\in\hat{y}}Q(x,y)}{\sum_{x\in\hat{x}}\mathrm{P}(X_t=x)}.$$

Moreover, regardless of the initial distribution,

$$\lim_{t\to+\infty}\mathrm{E}(\hat{f}(\hat{X}_t)) = \sum_{\hat{x}\in\hat{\mathscr{X}}}\hat{\pi}_\infty(\hat{x})\hat{f}(\hat{x}).$$

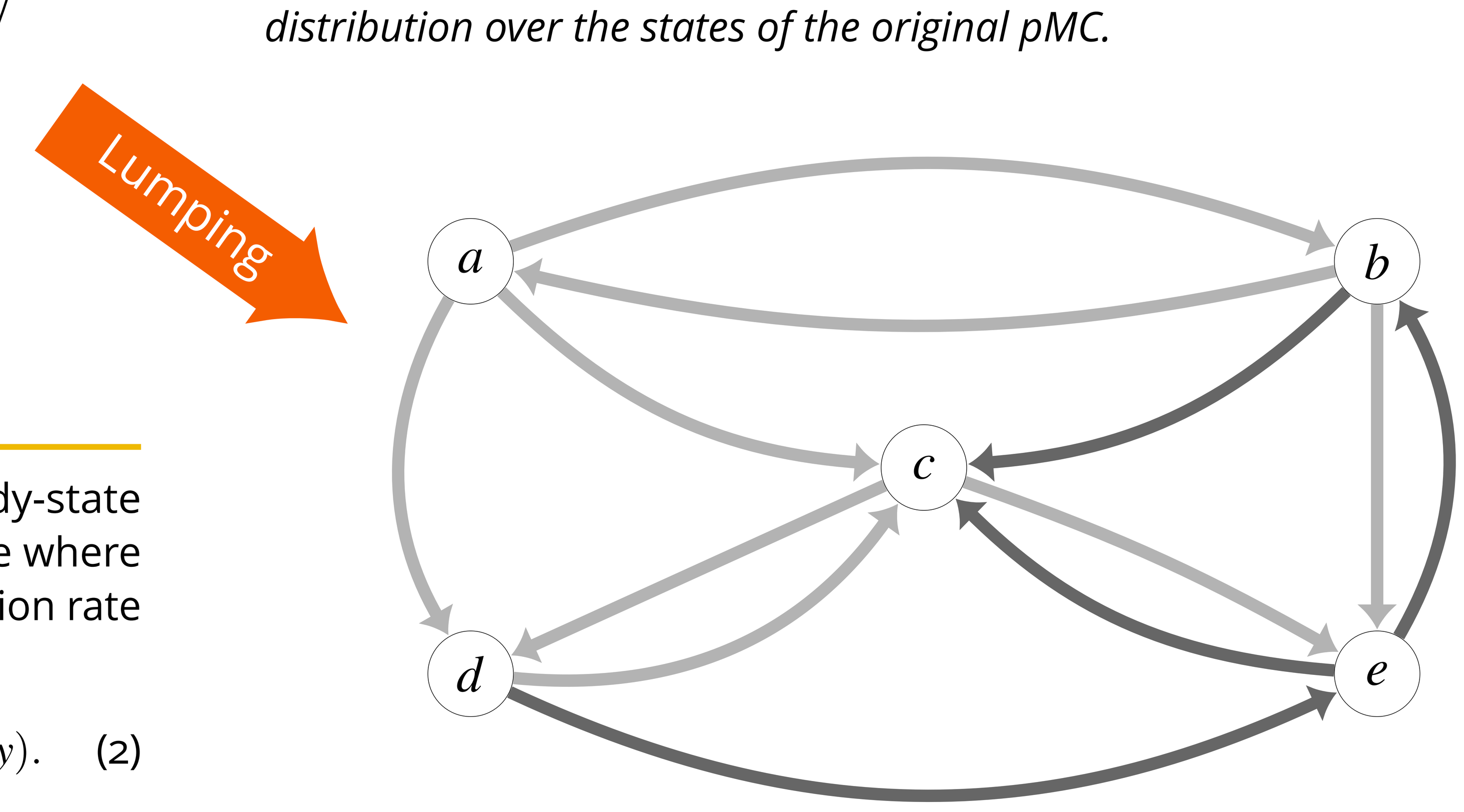
In this expression, $\hat{\pi}_\infty$ is the unique distribution on $\hat{\mathscr{X}}$ that satisfies

$$\hat{\pi}_\infty\hat{Q}_\infty = 0$$

where

$$\hat{Q}_\infty(\hat{x},\hat{y}) := \frac{\sum_{x\in\hat{x}}\pi_\infty(x)\sum_{y\in\hat{y}}Q(x,y)}{\sum_{x\in\hat{x}}\pi_\infty(x)}.$$

In general, we can only precisely determine the (long-term limit of the) temporal evolution of the probability distribution over the lumps if we first determine the (long-term limit of the) temporal evolution of the probability distribution over the states of the original pMC.



Lumped pMC is stationary

So far, lumping states in order to make determining the steady-state distribution feasible was limited to the well-known special case where the *lumped process is a stationary pMC*. This occurs if the transition rate matrix Q of the original (stationary and ergodic) pMC satisfies

$$(\forall \hat{x},\hat{y}\in\hat{\mathscr{X}})\ \min_{x\in\hat{x}}\sum_{y\in\hat{y}}Q(x,y) = \hat{Q}_t(\hat{x},\hat{y}) = \hat{Q}_\infty(\hat{x},\hat{y}) = \max_{x\in\hat{x}}\sum_{y\in\hat{y}}Q(x,y). \tag{2}$$

Lumped pMC is not stationary

We now consider an original stationary pMC of which the transition rate matrix Q does not satisfy (2). In this case, all we can say for sure about the lumped pMC—without determining the actual distribution $\mathrm{P}(X_t=x)$ of the original pMC—is that, for all $t\in\mathbb{R}_{\geq 0}$,

$$(\forall \hat{x},\hat{y}\in\hat{\mathscr{X}},\hat{x}\neq\hat{y})\ \min_{x\in\hat{x}}\sum_{y\in\hat{y}}Q(x,y)\leq\hat{Q}_t(\hat{x},\hat{y})\leq\max_{x\in\hat{x}}\sum_{y\in\hat{y}}Q(x,y). \tag{3}$$

We collect all transition rate matrices that satisfy (3) in the set $\hat{\mathscr{Q}}$, and let $\underline{\hat{Q}}$ denote the associated lower transition rate operator. By the theory of **imprecise Markov chains** (or iMCs), the lumped pMC is then contained in the set $\mathbb{P}_{\hat{\mathscr{Q}}}$. Consequently, we are guaranteed that

$$\underline{\mathrm{E}}(\hat{f}(\hat{X}_t)|\hat{X}_s=\hat{x})\leq\mathrm{E}(f(X_t)|X_s\in\hat{x})\leq\overline{\mathrm{E}}(\hat{f}(\hat{X}_t)|\hat{X}_s=\hat{x}) := -\underline{\mathrm{E}}(-\hat{f}(\hat{X}_t)|\hat{X}_s=\hat{x}).$$

It can be moreover shown that the obtained iMC is ergodic, whence

$$\underline{\mathrm{E}}_\infty(\hat{f})\leq\mathrm{E}_\infty(f)\leq\overline{\mathrm{E}}_\infty(\hat{f}) := -\underline{\mathrm{E}}_\infty(-\hat{f}). \tag{4}$$

Want to know how to efficiently approximate $\underline{\mathrm{E}}(\hat{f}(\hat{X}_t)|\hat{X}_s=\hat{x})$ or $\underline{\mathrm{E}}_\infty(\hat{f})$ up to some guaranteed maximal error? See **iMCs: Efficient computational methods with guaranteed error bounds**.

Alternative bounds on $\mathrm{E}_\infty(f)$

Assume we are only interested in determining (guaranteed bounds for) the limit expectation $\mathrm{E}_\infty(f)$ of some real-valued function f on $\mathscr{X}$. An alternative to computing the limit lower expectation of the induced iMC is the following.

For any $A\subseteq\hat{\mathscr{X}}$ and $\hat{x}\in\hat{\mathscr{X}}$, we define

$$\hat{Q}_L(\hat{x},A) := \min_{x\in\hat{x}}\sum_{\hat{y}\in A}\sum_{y\in\hat{y}}Q(x,y) \quad\text{and}\quad \hat{Q}_U(\hat{x},A) := \max_{x\in\hat{x}}\sum_{\hat{y}\in A}\sum_{y\in\hat{y}}Q(x,y).$$

Let $\mathscr{A}$ be a collection of subsets of $\hat{\mathscr{X}}$. If we let

$$\hat{\Pi}_{\mathscr{A}} := \Big\{\hat{\pi} \text{ a probability distribution on } \hat{\mathscr{X}}\colon (\forall A\in\mathscr{A})\ \sum_{\hat{x}\in\hat{\mathscr{X}}}\hat{\pi}(\hat{x})\hat{Q}_L(\hat{x},A)\leq 0\leq\sum_{\hat{x}\in\hat{\mathscr{X}}}\hat{\pi}(\hat{x})\hat{Q}_U(\hat{x},A)\Big\},$$

then

$$\min_{\hat{\pi}\in\hat{\Pi}_{\mathscr{A}}}\sum_{\hat{x}\in\hat{\mathscr{X}}}\hat{\pi}(\hat{x})\hat{f}(\hat{x})\leq\mathrm{E}_\infty(f)\leq\max_{\hat{\pi}\in\hat{\Pi}_{\mathscr{A}}}\sum_{\hat{x}\in\hat{\mathscr{X}}}\hat{\pi}(\hat{x})\hat{f}(\hat{x}), \tag{5}$$

where the optimisations can be solved using a linear program.

How to pick $\mathscr{A}$ and the tightness of the bounds of (5) *compared to the bounds* (4) *of the iMC is the subject of ongoing research.*