Bounding Inferences for Large-Scale Continuous-Time Markov Chains: A New Approach Based on Lumping and Imprecise Markov Chains

Alexander Erreygers, Jasper De Bock[1]

FLip, ELIS Department, Ghent University, Technologiepark 125, 9052 Zwijnaarde, Belgium

Abstract

If the state space of a homogeneous continuous-time Markov chain is too large, making inferences becomes computationally infeasible. Fortunately, the state space of such a chain is usually too detailed for the inferences we are interested in, in the sense that a less detailed—smaller—state space suffices to unambiguously formalise the inference. However, in general this so-called lumped state space inhibits computing exact inferences because the corresponding dynamics are unknown and/or intractable to obtain. We address this issue by considering an imprecise continuous-time Markov chain. In this way, we are able to provide guaranteed lower and upper bounds for the inferences of interest, without suffering from the curse of dimensionality.

Keywords: lumping, imprecise Markov chain, state space explosion

1. Introduction

State space explosion, or the exponential dependency of the size of a finite state space on a system's dimensions, is a frequently encountered inconvenience when constructing mathematical models of systems. In the setting of continuous-time Markov chains this exponentially increasing number of states has as a consequence that using the model to perform inferences about large-scale systems becomes computationally intractable. Fortunately, for many of the inferences we would like to make, a higher-level state description actually suffices to formalise the inference, allowing for a reduced state space with considerably fewer states. However, unfortunately, this creates a new problem, because the low-level description and its corresponding larger state space are necessary in order to easily characterise the system's dynamics.

The procedure of going from a low-level to a higher-level state description is called *lumping*. It was—to the best of our knowledge—first proposed by Kemeny and Snell [1] in the discrete-time setting and later considered by Burke and Rosenblatt [2] in both the discrete-time and continuous-time settings. These authors exploit the relation between the original state space—corresponding to the low-level state description—and the lumped state space—corresponding to the higher-level state description—to obtain a *lumped stochastic process* from the original Markov chain. Unfortunately, this lumped stochastic process is not necessarily a homogeneous Markov chain. In fact, Burke and Rosenblatt [2] provide a very stringent (necessary and) sufficient condition on the original Markov chain under which the lumped stochastic process is again a homogeneous Markov chain. That this condition is not trivially satisfied is quite unfortunate, because if the lumped stochastic process is not a homogeneous Markov chain, then using it to make inferences about the system is not feasible in practice.

Further research on the lumping of Markov chains centred around two separate subjects. On the one hand, several authors generalised the aforementioned (necessary and) sufficient conditions to other settings—see for instance [3–6]—or devised algorithms to determine the smallest reduced state space for which the lumped process is still a homogeneous Markov chain—see for instance [7, 8]. Franceschinis and Muntz [9] and Buchholz

[1] Jasper De Bock's research was partially funded by the European Commission's H2020 programme, through the UTOPIAE Marie Curie Innovative Training Network, H2020-MSCA-ITN-2016, Grant Agreement number 722734.

[10], on the other hand, proposed methods based on the lumping procedure to bound limit expectations with respect to the original Markov chain. Furthermore, the lumping procedure has also been used by Katoen et al. [11] in the context of model checking and bisimulation.

We here follow the historical evolution of the previously mentioned research: we start with a theoretical study of the lumped stochastic process and then propose methods based on the lumping procedure to bound expectations with respect to the original Markov chain. We start off our theoretical study in Section 2 with recalling some notation and terminology regarding Markov chains. This refresher is followed by a formal introduction of the lumping procedure and the resulting lumped stochastic process in Section 3. We consider the latter to be our first—albeit minor—contribution, because previous studies—for instance that of Burke and Rosenblatt [2]—always ignored some technicalities in the construction of the lumped stochastic process. Following this, we briefly introduce imprecise continuous-time Markov chains [12–14] in Section 4. Subsequently, we look at the lumped stochastic process from the point of view of imprecise Markov chains in Section 5. Specifically, we argue that the lumping procedure induces an elegantly characterised imprecise Markov chain and explain how it yields bounds on (conditional) expectations with respect to the lumped stochastic process. Next, we use the (building blocks of) our lumped imprecise Markov chain to obtain bounds on (conditional) expectations with respect to the original Markov chain in Section 6 and limit expectations with respect to the original (ergodic) Markov chain in Section 7. Compellingly, for Markov chains with a very large original state space but a significantly smaller lumped state space, it turns out that these bounds can be tractably computed even if it is infeasible to compute the precise value of the expectation with respect to the original Markov chain! The performance of our methods is evaluated, and compared to the performance of the methods of Franceschinis and Muntz [9] and Buchholz [10], by means of some simple numerical experiments in Section 8. After a brief return to the original setting of lumping in Section 9, we report our conclusions and provide some suggestions for future research in Section 10. Proofs for the results in the main text, as well as some extra material, can be found in the Appendix.

This is by no means the first time that we approach the lumping procedure using imprecise Markov chains, but our present approach is significantly more general. In [15, 16], we limited ourselves to providing bounds on the limit expectation of a specific irreducible Markov chain that appears in the context of telecommunication; an approach we generalised—to marginal expectations of general irreducible Markov chains—and properly justified from a theoretical point of view in [17]. We here extend [17] on three fronts: (i) we define the lumped stochastic process corresponding to a general Markov chain instead of an irreducible homogeneous Markov chain; (ii) we provide bounds for (conditional) expectations of real-valued functions on the state at any finite number of time points instead of marginal expectations of real-valued functions on the state at a single time point; and (iii) we provide two methods to determine bounds on the limit expectation of ergodic—that is, including irreducible—Markov chains instead of just one method for the limit expectation of irreducible Markov chains.

2. Continuous-time Markov chains

We are interested in making inferences about a system, and more specifically about the state of this system at any finite number of time points. The complication is that we are unable to predict the temporal evolution of this state with certainty. Therefore, at all times $t \in \mathbb{R}_{\geq 0}$[2], the state X_t of the system is a random variable that takes values—generically denoted by x, y or z—in the non-empty and finite state space $\mathscr{X}$.

Because we limit ourselves to inferences that depend on the state of the system at any finite number of time points, we adopt the framework of Krak et al. [12] for stochastic processes, which is a bit different from the standard framework. The main difference is that they consider (finitely-additive) coherent conditional probability measures [18] instead of the more common σ-additive probability measures; we refer to [12] for several arguments in favour of this choice. In light of the current contribution, we here summarise three of these arguments. First and foremost, they—and we—only consider expectations of functions that depend

[2] We use $\mathbb{R}$, $\mathbb{R}_{\geq 0}$ and $\mathbb{R}_{>0}$ to denote the set of real numbers, non-negative real numbers and positive real numbers, respectively. Furthermore, we use $\mathbb{N}$ to denote the natural numbers and write $\mathbb{N}_0$ when including zero.

on the state of the system at a finite number of time points, hence eliminating the need for a σ-algebra of events. A second argument is that, in their framework, the conditional probability of an event is always defined, even if the conditioning event has probability zero, whereas in the standard framework conditional events are derived from the unconditional probabilities and therefore not defined if the conditioning event has probability zero. Finally, Krak et al. [12] mention that their framework can be extended to allow for a σ-algebra of events, and therefore also functions of the state at a (countably) infinite number of time points; however, a proper theoretical study of this extension is no small feat and still remains to be done.

2.1. Finite sequences of time points

Essential to the description of stochastic processes are finite and increasing sequences of time points $t_1, \dots, t_n$, which is why we introduce some simplifying notation. Following Krak et al. [12], we collect all such sequences—including the empty sequence $\emptyset$—in the set $\mathscr{U}$, and denote a generic element of this set by u. We denote the set of all time sequences without the empty sequence by $\mathscr{U}_\emptyset$, and for all t in $\mathbb{R}_{\geq 0}$ use $\mathscr{U}_{<t}$ to denote the set of all time sequences of which the last time point strictly precedes t. Furthermore, for any non-empty and finite set $\mathscr{Y}$ and any sequence $u = t_1, \dots, t_n$ in $\mathscr{U}$, we define $\mathscr{Y}_u \coloneqq \mathscr{Y}^n$ and then use x_u to elegantly denote a generic n-tuple $(x_{t_1}, \dots, x_{t_n})$ in $\mathscr{Y}_u$. For the empty sequence $\emptyset$, we let $\mathscr{Y}_\emptyset$ be the singleton containing the empty tuple, usually denoted by $x_\emptyset$. We will sometimes also need to concatenate two increasing sequences of finite time points, for instance u and v in $\mathscr{U}$. Since u and v can be identified with sets, we let $u \cup v$ denote their concatenation, taken to be their ordered union. Finally, for any sequence $u = t_1, \dots, t_n$ in $\mathscr{U}_\emptyset$, we let $\min u \coloneqq \min\{t_i \colon i \in \{1, \dots, n\}\} = t_1$ and $\max u \coloneqq \max\{t_i \colon i \in \{1, \dots, n\}\} = t_n$. If u is the empty sequence, then conditions of the form "$\max u < \cdot$" are taken to be trivially satisfied.

2.2. Continuous-time stochastic processes

For a formal treatment of the coherent conditional probability framework for continuous-time stochastic processes, we refer to the extensive exposition in [12, Section 4] or to the summary in Appendix A.2. For our present purposes, it suffices to think of a continuous-time stochastic process P with state space $\mathscr{X}$ as being fully defined by its initial and transition probabilities. The initial probabilities are

$$P(X_0 = x_0),$$

with x_0 a state; the transition probabilities are of the form

$$P(X_{t+\Delta} = y \mid X_u = x_u, X_t = x),$$

where t and $t + \Delta$ are time points—that is, t and Δ are non-negative real numbers—u is a sequence of time points preceding t, x and y are states and x_u is a state instantiation in $\mathscr{X}_u$. Technically speaking, these initial and transition probabilities are assumed to be part of a coherent conditional probability, as explained in Appendix A.2. From a practical point of view, however, and for the purpose of following the main text of this contribution, it suffices to understand that we essentially demand that the initial and transition probabilities are compatible with the laws of probability: we demand that P is non-negative, normed, finitely additive for disjunct events and satisfies—the multiplicative version of—Bayes' rule.

2.3. Homogeneous continuous-time Markov chains

A well-known and often-used type of stochastic processes are *homogeneous continuous-time Markov chains*. Their popularity stems largely from the fact that they are easily characterised. Because most of the terminology and notation concerning homogeneous continuous-time Markov chains is essentially well-known, we here limit ourselves to the bare necessities; for a more thorough treatment, we refer to [12, 19, 20]. Since we deal exclusively with continuous-time Markov chains, we will henceforth drop the "continuous-time" adjective for the sake of brevity.

We now call a stochastic process P a Markov chain if satisfies the Markov property. Informally, this means that the transition probabilities only depend on the last state and not on the entire state history; formally, the Markov property holds if for all t, Δ in $\mathbb{R}_{\geq 0}$, u in $\mathscr{U}_{<t}$, x, y in $\mathscr{X}$ and x_u in $\mathscr{X}_u$,

$$P(X_{t+\Delta} = y \mid X_u = x_u, X_t = x) = P(X_{t+\Delta} = y \mid X_t = x). \tag{1}$$

This Markov chain P is called *homogeneous* if furthermore

$$P(X_{t+\Delta} = y \mid X_t = x) = P(X_\Delta = y \mid X_0 = x). \tag{2}$$

It is well-known that—both in the classical σ-additive or measure-theoretic framework [20] and the full conditional framework [12]—a homogeneous Markov chain is uniquely characterised by the triplet $(\mathscr{X}, \pi_0, Q)$, where $\mathscr{X}$ is the state space, π_0 an initial distribution and Q a transition rate matrix. In practice—see for instance [9, 10, 15, 16, 21, 22]—a homogeneous Markov chain model is therefore specified by providing such a triplet.

The initial distribution π_0 models the initial state of the system. It is given by

$$\pi_0(x) \coloneqq P(X_0 = x) \qquad \text{for all } x \in \mathscr{X}, \tag{3}$$

and hence is a probability mass function on $\mathscr{X}$—that is, is a non-negative function that sums to one.

The transition rate matrix Q models the dynamics of the system. Informally, the transition rate $Q(x, y)$ is to be understood as the rate of change—that is, the derivative—of the probability of going from a state x to another state y in an infinitesimal time period. More formally, it is the given by the limit expression

$$Q(x, y) = \lim_{\Delta \to 0^+} \frac{P(X_\Delta = y \mid X_0 = x) - I(x, y)}{\Delta} \quad \text{for all } x, y \in \mathscr{X}, \tag{4}$$

where I is the identity matrix. It now follows from the laws of probability that the matrix Q is a *transition rate matrix*: it has non-negative off-diagonal entries and rows that sum up to zero. Before we continue with explaining the use of the transition rate matrix, we first need to introduce some notation and terminology concerning matrices and more general—not necessarily linear—transformations.

2.4. Functions, transformations and norms

For any non-empty and finite set $\mathscr{Y}$, we let $\mathscr{L}(\mathscr{Y})$ denote the set of all real-valued functions on $\mathscr{Y}$. An often-used type of functions in $\mathscr{L}(\mathscr{Y})$ are the probability mass functions, in this setting usually called *distributions*: a non-negative real-valued function on $\mathscr{Y}$ that sums up to one. We denote the subset of $\mathscr{L}(\mathscr{Y})$ that consists of all distributions by $\mathscr{D}(\mathscr{Y})$. A second often-used type of function in $\mathscr{L}(\mathscr{Y})$ is the *indicator* of some subset $A \subseteq \mathscr{Y}$, denoted by $\mathbb{I}_A$ and defined by $\mathbb{I}_A(x) \coloneqq 1$ if x is an element of A and $\mathbb{I}_A(x) \coloneqq 0$ otherwise. In order not to obfuscate the notation too much, for all x in $\mathscr{Y}$, we write $\mathbb{I}_x$ instead of $\mathbb{I}_{\{x\}}$. Another notational convenience that we will adopt is to implicitly identify any real number μ with the corresponding constant function. Finally, the inner product $\langle \cdot, \cdot \rangle$ on $\mathscr{L}(\mathscr{Y})$ is given by $\langle f, g \rangle = \sum_{x \in \mathscr{Y}} f(x)g(x)$ for all f, g in $\mathscr{L}(\mathscr{Y})$.

In the setting of (imprecise) Markov chains, we often use *transformations* on $\mathscr{L}(\mathscr{Y})$: maps from $\mathscr{L}(\mathscr{Y})$ to $\mathscr{L}(\mathscr{Y})$. Such a transformation M is *non-negatively homogeneous* if, for all f in $\mathscr{L}(\mathscr{Y})$ and λ in $\mathbb{R}_{\geq 0}$, $M(\lambda f) = \lambda M f$. If furthermore $M(f + g) \geq Mf + Mg$ for all f, g in $\mathscr{L}(\mathscr{Y})$, then M is super-additive; M is called linear if this relation holds with equality instead of inequality. Note that if M is linear, then $M(\lambda f) = \lambda M f$ for negative real numbers λ as well. If we fix some ordering on the set $\mathscr{Y}$, then we can identify any linear transformation with a square matrix, the (x, y)-component of which is $[M\mathbb{I}_y](x)$. Therefore, we will use the terms matrix and linear transformation interchangeably. One example of a linear transformation is the identity transformation (or matrix) I that we have already used in Eqn. (4), which maps any f in $\mathscr{L}(\mathscr{Y})$ to itself: $If \coloneqq f$. Another example are transition rate matrices such as the matrix Q defined in Eqn. (4).

We end our discussion of transformations on $\mathscr{L}(\mathscr{Y})$ with norms. We bestow $\mathscr{L}(\mathscr{Y})$ with the maximum norm:

$$\|f\| \coloneqq \max|f| = \max\{|f(x)| \colon x \in \mathscr{Y}\} \quad \text{for all } f \text{ in } \mathscr{L}(\mathscr{Y}).$$

This norm on $\mathscr{L}(\mathscr{Y})$ induces a norm for non-negatively homogeneous transformations $M \colon \mathscr{L}(\mathscr{Y}) \to \mathscr{L}(\mathscr{Y})$:

$$\|M\| \coloneqq \sup\{\|Mf\| \colon f \in \mathscr{L}(\mathscr{Y}), \|f\| = 1\}.$$

Specifically, for any transition rate matrix R we have

$$\|R\| = 2 \max\{-[R\mathbb{I}_x](x) \colon x \in \mathscr{Y}\} = 2 \max\{-R(x, x) \colon x \in \mathscr{Y}\}. \tag{5}$$

2.5. Computing expectations

It is well-known—see for instance [12, 19]—that for any t, s in $\mathbb{R}_{\geq 0}$ such that $t \leq s$, any u in $\mathscr{U}_{<t}$, any x_u in $\mathscr{X}_u$, x in $\mathscr{X}$ and f in $\mathscr{L}(\mathscr{X})$,

$$E(f(X_s) \mid X_u = x_u, X_t = x) = [T_t^s f](x), \tag{6}$$

with

$$T_t^s := e^{(s-t)Q} = \lim_{n \to +\infty} \left(I + \frac{s-t}{n} Q \right)^n, \tag{7}$$

where the n-th power is to be interpreted as n consecutive applications. It is then well-known—see for instance [19, Theorem 2.1.2]—that T_t^s is a *transition matrix*: it has non-negative entries and rows that sum up to one.

Recall that our goal is to determine the expectation of functions on the state at any finite number of time points instead of just a single time point; that is, expectations of the form $E(f(X_u, X_v) \mid X_u = x_u)$. One well-known way to achieve this is to combine Eqn. (6) with the law of iterated expectation. We refer the interested reader to [12, Section 9] or our summary in Appendix A.3. For future reference, however, we here do explicitly mention that

$$E(f(X_t)) = \langle \pi_0, T_0^t f \rangle. \tag{8}$$

3. Lumping and the lumped process

As we have just established, evaluating $T_t^s f$ is essential when computing expectations for a homogeneous Markov chain. Unsurprisingly, analytically evaluating the limit in Eqn. (7) is often infeasible. Therefore, in order to compute expectations we usually have to resort to one of the many available numerical methods—see for example [23]—that approximate $T_t^s f$. The problem is that these numerical methods turn out to be computationally intractable when the state space becomes too large. This is especially unfortunate in applications where Markov chains are used, because a system with practical relevance often results in a model with a state space that is too large. We refer to [16] for one example of an application where the size of the state space leads to tractability issues.

Fortunately, as we already mentioned in the Introduction, the state space $\mathscr{X}$ is often unnecessarily detailed. Indeed, many interesting inferences can usually still be unambiguously defined using real-valued functions on a less detailed state space that corresponds to a higher-order description of the system, denoted by $\hat{\mathscr{X}}$. However, this provides no immediate solution as the rationale for using the detailed state space $\mathscr{X}$ in the first place is that this allows one to accurately model the system using a Markov chain; see [9, 10, 16, 21] for practical examples. In contrast, the transition probabilities of the system with respect to the reduced state space $\hat{\mathscr{X}}$—that is, (the dynamics of) the induced stochastic process—are often unknown and/or intractable to obtain, which inhibits us from making exact inferences using the induced stochastic process. We will determine the initial and transition probabilities of the lumped process in Section 3.2 and address the tractability issues by allowing for imprecision in Sections 5, 6 and 7.

3.1. Notation and terminology concerning lumping

We assume that the lumped state space $\hat{\mathscr{X}}$ is obtained by *lumping*—sometimes called grouping or aggregating, see [2, 3]—states in $\mathscr{X}$, such that $1 \leq |\hat{\mathscr{X}}| \leq |\mathscr{X}|$. This lumping is formalised by the surjective *lumping map* $\Lambda\colon \mathscr{X} \to \hat{\mathscr{X}}$, which maps every state x in $\mathscr{X}$ to a state $\Lambda(x) = \hat{x}$ in $\hat{\mathscr{X}}$. In the remainder, we will implicitly use the obvious extension of the lumping map Λ to tuples of states—that is, to the domain $\mathscr{X}_u$, where u is a sequence of time points. Using this (extended) lumping map, we define the *inverse lumping map* Λ^{-1}: $\hat{x}_\emptyset$ is mapped to $\Lambda^{-1}(\hat{x}_\emptyset) := x_\emptyset$ and, for any u in $\mathscr{U}_\emptyset$, $\hat{x}_u$ in $\hat{\mathscr{X}}_u$ is mapped to

$$\Lambda^{-1}(\hat{x}_u) := \{x_u \in \mathscr{X}_u \colon \Lambda(x_u) = \hat{x}_u\} = \{x_u \in \mathscr{X}_u \colon (\forall t \in u)\ \Lambda(x_t) = \hat{x}_t\}.$$

In order to lighten our notation, we will frequently shorten "$x_u \in \Lambda^{-1}(\hat{x}_u)$" to "$x_u \in \hat{x}_u$".

As far as our results are concerned, it does not matter in which way the states are lumped. Say we are interested in $E(f(X_t))$ for a given f in $\mathscr{L}(\mathscr{X})$ and t in $\mathbb{R}_{\geq 0}$. Then a naive choice is to lump together all states that have the same image under f. However, this is not necessarily a good choice. One reason is that the resulting lumped state space can become very small, for example when f is an indicator, resulting in too much imprecision in the dynamics and/or the inference. Lumping-based methods therefore often let $\hat{\mathscr{X}}$ correspond to a natural higher-level description of the state of the system; see for example [9, 10, 16] for some positive results. An extra benefit of this approach is that the resulting model can be used to determine the expectation of multiple functions.

3.2. The lumped stochastic process

All of the necessary notation and tools have been introduced to define the lumped stochastic process. In order not to unnecessarily complicate our exposition, we here restrict ourselves to an intuitive summary; for the fully formal definition, we refer the interested reader to Appendix B. Furthermore, we will here and in the remainder limit ourselves to homogeneous Markov chains, while our formal definition of the lumped stochastic process actually concerns general Markov chains. We have two reasons for doing so: (i) it is in line with the historical interest in lumping, see for instance [2, 3, 9, 10, 21] and Section 9; and (ii) almost all of the Markov chains that arise when modelling practical problems are assumed to be homogeneous for the sake of simplicity, see for instance [9, 10, 16, 21].

Recall from Section 2 that a stochastic process is characterised by its initial and transition probabilities. For the lumped stochastic process, denoted by $\hat{P}$, these probabilities are determined by the homogeneous Markov chain P and the lumping map $\Lambda\colon \mathscr{X} \to \hat{\mathscr{X}}$. More specifically, they follow from the relation

$$(\hat{X}_u = \hat{x}_u) \Leftrightarrow (X_u \in \hat{x}_u) = \bigcup_{x_u \in \hat{x}_u} (X_u = x_u) \text{ for all } u \text{ in } \mathscr{U} \text{ and } \hat{x}_u \text{ in } \hat{\mathscr{X}}_u. \tag{9}$$

First, we observe that it follows from this relation that, for any u in $\mathscr{U}$ and $\hat{x}_u$ in $\hat{\mathscr{X}}_u$,

$$\hat{P}(\hat{X}_u = \hat{x}_u) = P(X_u \in \hat{x}_u) = \sum_{x_u \in \hat{x}_u} P(X_u = x_u). \tag{10}$$

The initial probabilities of the lumped stochastic process $\hat{P}$ are obtained by setting $u = 0$ in Eqn. (10):

$$\hat{P}(\hat{X}_0 = \hat{x}_0) = \sum_{x_0 \in \hat{x}_0} P(X_0 = x_0) \qquad \text{for all } \hat{x}_0 \in \hat{\mathscr{X}}. \tag{11}$$

The transition probabilities $\hat{P}(\hat{X}_{t+\Delta} = \hat{y} \mid \hat{X}_u = \hat{x}_u, \hat{X}_t = \hat{x})$—with t, Δ in $\mathbb{R}_{\geq 0}$, u in $\mathscr{U}_{<t}$, $\hat{x}, \hat{y}$ in $\hat{\mathscr{X}}$ and $\hat{x}_u$ in $\hat{\mathscr{X}}_u$—follow from Eqn. (10) and Bayes' rule, at least if the conditioning event $(\hat{X}_u = \hat{x}_u, \hat{X}_t = \hat{x})$ has non-zero probability. In case the conditioning event has zero probability—that is, if $\hat{P}(\hat{X}_u = \hat{x}_u, \hat{X}_t = \hat{x}) = 0$—the transition probability does not follow from Bayes' rule. Hence, if there are conditioning events with zero probability, "the" lumped stochastic process is not uniquely defined! Therefore, we need to speak of *a* lumped stochastic process instead of the lumped stochastic process. From a technical point of view, the non-uniqueness is a direct consequence of our formal definition of a lumped stochastic process—see Appendix B—because we use an extension of the Markov chain—or coherent conditional probability—P to a coherent conditional probability on a larger domain, which need not be unique. That said, while the transition probabilities conditional on events with probability zero need not be uniquely defined, this does not mean that they can take any arbitrary value. For example, we show in Appendix D that for all t, Δ in $\mathbb{R}_{\geq 0}$, u in $\mathscr{U}_{<t}$, x, y in $\mathscr{X}$, x_u in $\mathscr{X}_u$ and $\hat{f}$ in $\mathscr{L}(\hat{\mathscr{X}})$,

$$\min_{x \in \hat{x}} E([\hat{f} \circ \Lambda](X_\Delta) \mid X_0 = x) \leq \hat{E}(\hat{f}(\hat{X}_{t+\Delta}) \mid \hat{X}_u = \hat{x}_u, \hat{X}_t = \hat{x}) \leq \max_{x \in \hat{x}} E([\hat{f} \circ \Lambda](X_\Delta) \mid X_0 = x),^{3} \tag{12}$$

[3] Here and in the remainder we use $h \circ g$ to denote the composition of any two functions g and h, given by $h \circ g\colon \operatorname{dom} g \to \operatorname{range} h\colon x \mapsto h(g(x))$.

where $\hat{E}$ denotes the expectation with respect to a lumped stochastic process $\hat{P}$. Setting $\hat{f} = \mathbb{I}_{\hat{y}}$, it follows from this inequality and the additivity of P that

$$\min_{x\in\hat{x}} \sum_{y\in\hat{y}} P(X_\Delta = y \mid X_0 = x) \leq \hat{P}(\hat{X}_{t+\Delta} = \hat{y} \mid \hat{X}_u = \hat{x}_u, \hat{X}_t = \hat{x}) \leq \max_{x\in\hat{x}} \sum_{y\in\hat{y}} P(X_\Delta = y \mid X_0 = x). \tag{13}$$

We emphasise here that the inequalities of Eqns. (12) and (13) hold for *any* conditioning event $(\hat{X}_u = \hat{x}_u, \hat{X}_t = \hat{x})$, regardless of whether it has probability zero or not.

In [17], we circumvented the non-uniqueness due to conditioning on events with probability zero by a priori limiting ourselves to homogeneous Markov chains with an irreducible transition rate matrix Q—see Section 7 further on for a definition—and positive initial distribution π_0. In that case, as more thoroughly explained in [24, Appendix D.1]—the appendix of the extended pre-print of [17]—the lumped stochastic process is uniquely defined because any conditioning event $(\hat{X}_u = \hat{x}_u, \hat{X}_t = \hat{x})$ has non-zero probability.

Putting aside the issue of non-uniqueness for now, there is another, much more important issue that we need to address: how can we describe the lumped stochastic process directly, without resorting to the original process. It is precisely for this reason that the original interest in the lumping of a Markov chain was limited to the case that the—or, more precisely, any—lumped stochastic process is again a homogeneous Markov chain. As is essentially well-known, this is not necessarily the case. We will return to this specific case in Section 9 further on; for now, we only mention that in order for a lumped stochastic process to again be a homogeneous Markov chain, the original chain needs to satisfy a very restrictive condition.

This is a major setback because this means that in order to compute the transition probabilities of—or, more generally, expectations with respect to—a lumped stochastic process $\hat{P}$, we cannot simply determine a lumped transition rate matrix $\hat{Q}$ and use the matrix exponential $\hat{T}_t^s$ that it generates according to Eqn. (7); instead, we have to explicitly determine the transition probabilities from the initial probability distribution and transition probabilities of the original chain. In the setting of large-scale Markov chains, computing the transition probabilities of the original chain is already intractable, and so explicitly determining the transition probabilities of—or, more generally, expectations with respect to—a lumped stochastic process is intractable as well! This is rather unfortunate, as it renders the whole lumping procedure useless at first sight: since lumping does not yield more tractable computations, we might as well just stick with the original Markov chain anyway.

The essential point of our contribution is that, while—in general—we cannot tractably determine (the dynamics of) a lumped stochastic process $\hat{P}$, we can consider a *set* of stochastic processes, not necessarily homogeneous and/or Markovian but all with $\hat{\mathscr{X}}$ as state space, that is fully characterised by π_0, Q and Λ and definitely contains any lumped stochastic process $\hat{P}$. Crucially, it turns out that this set takes the form of a so-called *imprecise (continuous-time) Markov chain.*

In the upcoming section, we explain how tight lower and upper bounds on the expectations that correspond to this set of processes are relatively easy to obtain. In particular, they can be determined without having to explicitly optimise over this set of processes, thus mitigating the need to actually construct it. Besides computational tractability, another benefit of this approach is that we circumvent the uniqueness issue because we can obtain conclusions about any lumped stochastic process $\hat{P}$ without having to explicitly determine it entirely.

4. Imprecise Markov chains

For a formal definition of (continuous-time) imprecise Markov chains, and an extensive study of their properties, we refer the reader to the work of Krak et al. [12] and De Bock [13]. We here only present a brief overview of the terminology, notation and results that are relevant in our setting.

4.1. Sets of consistent processes and lower expectations

In general, the main idea behind imprecise Markov chains is to consider a set of stochastic process instead of a single stochastic process. In particular, Krak et al. [12] focus on a set of processes that is fully characterised by a non-empty set of initial distributions $\mathscr{M}$ and a non-empty bounded set of transition rate matrices $\mathscr{Q}$. More specifically, they collect in $\mathbb{P}^{\mathrm{W}}_{\mathscr{Q},\mathscr{M}}$ all stochastic processes that are: (i) well-behaved, a technical condition [12, Definition 4.4]; (ii) consistent with $\mathscr{Q}$, in the sense that at all times the "outer partial derivative of the history-dependent transition matrix" is contained in $\mathscr{Q}$ [12, Definition 6.1]; and (iii) consistent with $\mathscr{M}$, in the sense that $\mathscr{M}$ contains the initial distribution [12, Definition 6.2].

Using this set $\mathbb{P}^{\mathrm{W}}_{\mathscr{Q},\mathscr{M}}$ of well-behaved and consistent stochastic processes, Krak et al. [12] define lower and upper expectations as follows. For any non-empty set of initial distributions $\mathscr{M}$ and non-empty bounded set of transition rate matrices $\mathscr{Q}$, they let

$$\underline{E}^{\mathrm{W}}_{\mathscr{Q},\mathscr{M}}(\cdot \mid \cdot) := \inf\{E_P(\cdot \mid \cdot) \colon P \in \mathbb{P}^{\mathrm{W}}_{\mathscr{Q},\mathscr{M}}\} \quad \text{and} \quad \overline{E}^{\mathrm{W}}_{\mathscr{Q},\mathscr{M}}(\cdot \mid \cdot) := \sup\{E_P(\cdot \mid \cdot) \colon P \in \mathbb{P}^{\mathrm{W}}_{\mathscr{Q},\mathscr{M}}\}, \tag{14}$$

where E_P denotes the expectation with respect to the process P. It is obvious from Eqn. (14) that the lower expectation $\underline{E}^{\mathrm{W}}_{\mathscr{Q},\mathscr{M}}$ and the upper expectation $\overline{E}^{\mathrm{W}}_{\mathscr{Q},\mathscr{M}}$ are conjugate, in the sense that

$$\overline{E}^{\mathrm{W}}_{\mathscr{Q},\mathscr{M}}(f(X_u, X_v) \mid X_u = x_u) = -\underline{E}^{\mathrm{W}}_{\mathscr{Q},\mathscr{M}}(-f(X_u, X_v) \mid X_u = x_u),$$

with v in $\mathscr{U}_\emptyset$, u in $\mathscr{U}_{<\min v}$, f in $\mathscr{L}(\mathscr{X}_{u\cup v})$ and x_u in $\mathscr{X}_u$. Because of this conjugacy, it suffices to focus on either one of them; we here focus on the lower expectation.

At first sight, it would seem that in order to execute the minimisation in Eqn. (14), we first have to explicitly construct the set of consistent processes $\mathbb{P}^{\mathrm{W}}_{\mathscr{Q},\mathscr{M}}$. Fortunately, Krak et al. [12] show that this is *not* the case. Instead, lower expectations can be computed using a non-linear semi-group that is generated by a so-called lower transition rate operator. The analogy with the case of homogeneous Markov chains is quite striking, as in that case expectations can be computed using a linear semi-group—the matrix exponential—that is generated by the transition rate matrix.

4.2. Lower transition rate operators

Let us first focus on the generator of the non-linear semi-group that we will be using. For any non-empty bounded set of transition rate matrices $\mathscr{Q}$, this generator is given by the transformation $\underline{Q}_{\mathscr{Q}} \colon \mathscr{L}(\mathscr{X}) \to \mathscr{L}(\mathscr{X}) \colon f \mapsto \underline{Q}_{\mathscr{Q}} f$, where

$$[\underline{Q}_{\mathscr{Q}} f](x) := \inf\{[Qf](x) \colon Q \in \mathscr{Q}\} \qquad \text{for all } f \in \mathscr{L}(\mathscr{X}), x \in \mathscr{X}. \tag{15}$$

This operator $\underline{Q}_{\mathscr{Q}}$ is called the *lower envelope* of $\mathscr{Q}$. By [12, Proposition 7.5], we know that it is a *lower transition rate operator* [12, Definition 7.2]: a super-additive and non-negatively homogeneous transformation that has "non-negative off-diagonal entries"—in the sense that $[\underline{Q}_{\mathscr{Q}} \mathbb{I}_y](x) \geq 0$ if $x \neq y$—and "rows that sum up to zero"—in the sense that $\underline{Q}_{\mathscr{Q}} \mu = 0$ for any constant function μ. Note that transition rate matrices are lower transition rate operators that are furthermore linear; hence, a lower transition rate operator is a non-linear generalisation of a transition rate matrix.

Just like a set of lower transition rate matrices $\mathscr{Q}$ defines a lower transition rate operator $\underline{Q}_{\mathscr{Q}}$, any generic lower transition rate operator $\underline{R}$ defines a corresponding set of *dominating transition rate matrices*

$$\mathscr{Q}_{\underline{R}} := \{Q \in \mathscr{R}(\mathscr{X}) \colon (\forall f \in \mathscr{L}(\mathscr{X}))\ Qf \geq \underline{R} f\}, \tag{16}$$

where $\mathscr{R}(\mathscr{X})$ denotes the set of all transition rate matrices on $\mathscr{L}(\mathscr{X})$. As we will see in Proposition 1 further on, the set $\mathscr{Q}_{\underline{R}}$ has some very nice properties. One of these properties is that is has separately specified rows, which is defined as follows.

Definition 1 (Definition 7.3 in [12])**.** A non-empty set of transition rate matrices $\mathscr{Q}$ has *separately specified rows* if for any collection $\{Q_x\}_{x\in\mathscr{X}}$ in $\mathscr{Q}$, there is a $Q^\star$ in $\mathscr{Q}$ such that

$$Q^\star(x, y) = Q_x(x, y) \qquad \text{for all } x, y \in \mathscr{X}.$$

The following result establishes that the set of dominating rate matrices satisfies this as well as several other convenient properties. It is one of our motivations for introducing the specific lumped lower transition rate operator in Section 5.2 further on.

Proposition 1 (Propositions 7.6, 7.7 and 7.8 in [12])**.** *Let $\underline{R}$ be a lower transition rate operator and $\mathscr{Q}_{\underline{R}}$ be the corresponding set of dominating transition rate matrices. Then $\mathscr{Q}_{\underline{R}}$ is non-empty, bounded, closed and convex. Furthermore, $\mathscr{Q}_{\underline{R}}$ has separately specified rows and $\underline{R}$ is the lower envelope of $\mathscr{Q}_{\underline{R}}$.*

4.3. Computing lower expectations

As we already hinted at in the introduction of Section 4.2, the lower transition rate operator $\underline{Q}_{\mathscr{Q}}$ is an essential tool when computing lower (conditional) expectations for an imprecise Markov chain $\mathbb{P}^{\mathrm{W}}_{\mathscr{Q},\mathscr{M}}$, much like Q is essential when computing expectations for precise Markov chains; the following result establishes an imprecise version of Eqn. (6).

Proposition 2 (Corollary 8.3 in [12])**.** *Let $\mathscr{M}$ be a non-empty set of initial distributions and $\mathscr{Q}$ a non-empty and bounded set of transition rate matrices that has separately specified rows. Then for any t, s in $\mathbb{R}_{\geq 0}$ with $t \leq s$, u in $\mathscr{U}_{<t}$, x in $\mathscr{X}$, x_u in $\mathscr{X}_u$ and f in $\mathscr{L}(\mathscr{X})$,*

$$\underline{E}^{\mathrm{W}}_{\mathscr{Q},\mathscr{M}}(f(X_s) \mid X_u = x_u, X_t = x) = [\underline{T}_t^s f](x) \tag{17}$$

with

$$\underline{T}_t^s \coloneqq \lim_{n \to +\infty} \left(I + \frac{s-t}{n} \underline{Q}_{\mathscr{Q}} \right)^n, \tag{18}$$

where the n-th power should be interpreted as n consecutive applications.

By [12, Theorem 7.12], we know that $\underline{T}_t^s$, as defined in Eqn. (18), is a *lower transition operator* [12, Definition 7.1]: a super-additive and non-negatively homogeneous transformation that dominates the minimum, so a non-linear generalisation of a transition matrix. Important to mention here is that, at least in general, it is infeasible if not impossible to determine $\underline{T}_t^s f$ by analytically evaluating the limit in Eqn. (18), and we therefore have to resort to an approximation method. In essence, this approximation method comes down to (i) choosing a suitable sequence $\delta_1, \dots, \delta_n$ of—sufficiently small—positive real numbers such that $\sum_{k=1}^n \delta_k = s - t$; and (ii) iteratively determining $g_k \coloneqq (I + \delta_k \underline{Q}) g_{k-1} = g_{k-1} + \delta_k \underline{Q} g_{k-1}$ for k ranging from 1 to n with initial condition $g_0 \coloneqq f$. This way, we end up with an approximation g_n for $\underline{T}_t^s f$. We refer to [12, Section 8.2] and [25] for a more thorough treatment of this approximation method, including procedures for choosing $\delta_1, \dots, \delta_n$ such that the error of the approximation is guaranteed to be smaller than some desired maximal error.

It is an immediate consequence of Eqns. (17) and (18) that

$$\underline{E}^{\mathrm{W}}_{\mathscr{Q},\mathscr{M}}(f(X_{t+\Delta}) \mid X_u = x_u, X_t = x) = \underline{E}^{\mathrm{W}}_{\mathscr{Q},\mathscr{M}}(f(X_{t+\Delta}) \mid X_t = x) = \underline{E}^{\mathrm{W}}_{\mathscr{Q},\mathscr{M}}(f(X_\Delta) \mid X_0 = x).$$

The first equality is an imprecise version of the Markov property, while the second equality is an imprecise version of the homogeneity property of a (precise) Markov chain. Therefore, Proposition 2 justifies calling $\mathbb{P}^{\mathrm{W}}_{\mathscr{Q},\mathscr{M}}$ a (homogeneous) imprecise Markov chain. Even more, the imprecise Markov chain $\mathbb{P}^{\mathrm{W}}_{\mathscr{Q},\mathscr{M}}$ also satisfies an imprecise version of the law of iterated expectation.

Proposition 3 (Theorem 6.5 in [12])**.** *If $\mathscr{M}$ is a non-empty set of initial distributions and $\mathscr{Q}$ a non-empty, bounded and convex set of transition rate matrices, then for any u, v, w in $\mathscr{U}$ with $\max u < \min v$ and $\max v < \min w$, x_u in $\mathscr{X}_u$ and f in $\mathscr{L}(\mathscr{X}_{u \cup v \cup w})$,*

$$\underline{E}^{\mathrm{W}}_{\mathscr{Q},\mathscr{M}}(f(X_u, X_v, X_w) \mid X_u = x_u) = \underline{E}^{\mathrm{W}}_{\mathscr{Q},\mathscr{M}}(\underline{E}^{\mathrm{W}}_{\mathscr{Q},\mathscr{M}}(f(X_u, X_v, X_w) \mid X_u, X_v) \mid X_u = x_u).$$

Propositions 2 and 3 imply a practical method to compute lower expectations that is entirely similar to the method for precise Markov chains outlined in Appendix A.3; for a detailed explanation of how and why this works, we refer to [12, Section 9.2] or our summary in Appendix C.2. This method is tractable as long as the state space $\mathscr{X}$ is sufficiently small and either the number of time points in v is small—because the number of computations is clearly exponential in $|v|$—or the function f is of a particular type.

5. The induced imprecise Markov chain

Now that we have established what imprecise Markov chains are, we are ready to consider the lumping procedure from their point of view. More specifically, we set out to characterise an imprecise Markov chain that contains any lumped stochastic process corresponding to a given (precise) Markov chain. To that end, we need to determine a set of initial distributions that contains the lumped initial probabilities and a set of transition rate matrices that contains the "outer partial derivatives of the instantaneous transition matrix" of any lumped stochastic process. Throughout this section, we let P be some homogeneous Markov chain and $\Lambda\colon \mathscr{X} \to \hat{\mathscr{X}}$ some lumping map.

5.1. The set of lumped initial probability distributions

We start with determining a suitable set of initial distributions. Recall from Eqn. (11) that, for any lumped stochastic process $\hat{P}$,

$$\hat{P}(\hat{X}_0 = \hat{x}) = \sum_{x\in\hat{x}} P(X_0 = x) = \sum_{x\in\hat{x}} \pi_0(x) \qquad \text{for all } \hat{x} \in \hat{\mathscr{X}},$$

where the final equality holds due to the definition of the initial distribution π_0. From this, it follows immediately that the initial distribution of any lumped stochastic process $\hat{P}$ is the lumped initial distribution

$$\hat{\pi}_0\colon \hat{\mathscr{X}} \to \mathbb{R}\colon \hat{x} \mapsto \hat{\pi}_0(\hat{x}) \coloneqq \sum_{x\in\hat{x}} \pi_0(x). \tag{19}$$

Hence, if we let $\hat{\mathscr{M}} \coloneqq \{\hat{\pi}_0\}$, then any lumped stochastic process $\hat{P}$ is *consistent* with $\hat{\mathscr{M}}$; see Section 4.1.

5.2. The set of lumped transition rate matrices

For the set transition rate matrices, we are looking for a set of transition rate matrices on $\mathscr{L}(\hat{\mathscr{X}})$ that contains the "outer partial derivative of the instantaneous transition matrix" of any lumped stochastic process. We will not explicitly construct such a set of transition rate matrices; instead, we define a lower transition rate operator and then consider the set of dominating transition rate matrices; see Section 4.2.

We deliberately do not go into detail about what the "outer partial derivative of the instantaneous transition matrix" of a lumped stochastic process $\hat{P}$ exactly is. For our present purposes, it suffices to understand that we are interested in the rate of change of $\hat{E}(\hat{f}(\hat{X}_{t+\Delta}) \mid \hat{X}_u = \hat{x}_u, \hat{X}_t = \hat{x})$ for Δ going to 0; the interested reader is referred to Appendix D. Using Eqn. (12), we show in Appendix D that there is a lower bound for this rate of change that applies to *any* lumped stochastic process. Specifically, this lower bound uses the transformation $\underline{\hat{Q}}\colon \mathscr{L}(\hat{\mathscr{X}}) \to \mathscr{L}(\hat{\mathscr{X}})$, defined for all $\hat{f}$ in $\mathscr{L}(\hat{\mathscr{X}})$ and $\hat{x}$ in $\hat{\mathscr{X}}$ as

$$[\underline{\hat{Q}}\hat{f}](\hat{x}) \coloneqq \min\left\{ \sum_{\hat{y}\in\hat{\mathscr{X}}} \hat{f}(\hat{y}) \sum_{y\in\hat{y}} Q(x,y)\colon x \in \hat{x} \right\} = \min\Big\{ [Q(\hat{f}\circ\Lambda)](x)\colon x \in \hat{x} \Big\}. \tag{20}$$

We call $\underline{\hat{Q}}$ the *lumped lower transition rate operator* because of the following straightforward result.

Proposition 4. *Let Q be a transition rate matrix and $\Lambda\colon \mathscr{X} \to \hat{\mathscr{X}}$ a lumping map. Then the corresponding transformation $\underline{\hat{Q}}$ is a lower transition rate operator.*

Important to mention here is that in case the lumped state space corresponds to some higher-order state description, we often find that executing the optimisation in Eqn. (20) is fairly straightforward—that is, reduces to computing the minimum over a very small (in our examples as low as two to four) number of cases—as is for instance observed in [15, 16] and Section 8 further on. In fact, a numerical implementation of $\underline{\hat{Q}}$ usually does not require an explicit construction of the original transition rate matrix Q.

Because the lumped lower transition rate operator $\underline{\hat{Q}}$ is a lower transition rate operator, we know from Section 4.2 that it induces a set of dominating transition rate matrices. In the current setting, we call this set the set of *lumped transition rate matrices*

$$\hat{\mathscr{Q}} \coloneqq \Big\{ \hat{Q} \in \mathscr{R}(\hat{\mathscr{X}})\colon (\forall \hat{f} \in \mathscr{L}(\hat{\mathscr{X}}))\ \hat{Q}\hat{f} \geq \underline{\hat{Q}}\hat{f} \Big\}, \tag{21}$$

where $\mathscr{R}(\hat{\mathscr{X}})$ denotes the set of all transition rate matrices on $\mathscr{L}(\hat{\mathscr{X}})$. The following result is now an immediate corollary of Propositions 1 and 4.

Corollary 5. *Let Q be a transition rate matrix and $\Lambda\colon \mathscr{X} \to \hat{\mathscr{X}}$ a lumping map. The associated set $\hat{\mathscr{Q}}$ of lumped transition rate matrices is non-empty, bounded, closed and convex. Furthermore, it has separately specified rows and its lower envelope is $\underline{\hat{Q}}$, in the sense that $\underline{Q}_{\hat{\mathscr{Q}}} = \underline{\hat{Q}}$.*

5.3. The imprecise lumped Markov chain

With some more work, we can furthermore prove that any lumped stochastic process $\hat{P}$ is well-behaved; see Section 4.1. The exact statements and proofs of the relevant results are rather technical and do not immediately contribute to a better understanding of the main text, which is why we have relegated these to Appendix D. Nonetheless, we now know that any lumped stochastic process is well-behaved and consistent with $\hat{\mathscr{M}}$ and $\hat{\mathscr{Q}}$. Therefore, we immediately obtain the following result.

Theorem 6. *Consider a homogeneous Markov chain P and a lumping map $\Lambda\colon \mathscr{X} \to \hat{\mathscr{X}}$. Then any corresponding lumped stochastic process $\hat{P}$ is contained in $\mathbb{P}^{\mathrm{W}}_{\hat{\mathscr{Q}},\hat{\mathscr{M}}}$.*

We consider Theorem 6 to be one of our main contributions, as it establishes that we can use all of the results from the theory of imprecise Markov chains to determine bounds on (conditional) expectations with respect to any lumped stochastic process. At present—at least to the best of our knowledge—this is limited to lower and upper expectations of functions that depend on the state at a *finite* number of time points. However, if the theory of imprecise Markov chains were to be extended to more general inferences—for instance to lower and upper conditional expectations of functions that depend on the state at an infinite number of future time points—then this result will immediately allow us to obtain bounds for these more general inferences with respect to any lumped process. With our present knowledge, however, we are limited to inferences of the following type.

Corollary 7. *Consider a homogeneous Markov chain P, a lumping map $\Lambda\colon \mathscr{X} \to \hat{\mathscr{X}}$ and a corresponding lumped stochastic process $\hat{P}$. Then for all u in $\mathscr{U}$, v in $\mathscr{U}_\emptyset$ with $\max u < \min v$, $\hat{x}_u$ in $\hat{\mathscr{X}}_u$ and $\hat{f}$ in $\mathscr{L}(\hat{\mathscr{X}}_{u\cup v})$,*

$$\underline{E}^{\mathrm{W}}_{\hat{\mathscr{Q}},\hat{\mathscr{M}}}(\hat{f}(\hat{X}_u, \hat{X}_v) \mid \hat{X}_u = \hat{x}_u) \leq \hat{E}(\hat{f}(\hat{X}_u, \hat{X}_v) \mid \hat{X}_u = \hat{x}_u) \leq \overline{E}^{\mathrm{W}}_{\hat{\mathscr{Q}},\hat{\mathscr{M}}}(\hat{f}(\hat{X}_u, \hat{X}_v) \mid \hat{X}_u = \hat{x}_u),$$

where $\hat{E}$ denotes expectation—in the usual sense—with respect to the lumped stochastic process $\hat{P}$.

The benefit of this result is that it mitigates the need to explicitly determine the lumped transition probabilities; instead—as explained in Section 4.3—we can use the semi-group $\underline{\hat{T}}_t^s$ generated by $\underline{\hat{Q}}$ instead. This is especially useful in case the size of the original state space $\mathscr{X}$ makes computing $T_t^s f$ infeasible, but the size of the lumped state space $\hat{\mathscr{X}}$ is significantly smaller so that computing $\underline{\hat{T}}_t^s \hat{f}$ *is* feasible. In this case, our results allow us to obtain guaranteed bounds on an inference that we could not compute otherwise!

6. Bounding expectations

Corollary 7 allows us to bound expectations with respect to any lumped process. However, we are actually interested in (bounding) expectations with respect to the original Markov chain, and it is not immediately clear how one can use Corollary 7 to do this. In this section, we nevertheless set out to use the lumped imprecise Markov chain to determine bounds on expectations with respect to the original Markov chain.

One problem is that the expectations with respect to the original Markov chain are for real-valued functions on $\mathscr{X}_u$, while the expectations with respect to the lumped imprecise Markov chain are for real-valued functions on $\hat{\mathscr{X}}_u$. Therefore, we need a way to reduce real-valued functions on $\mathscr{X}_u$ to real-valued functions on $\hat{\mathscr{X}}_u$. Fix some u in $\mathscr{U}_\emptyset$. The sole functions f in $\mathscr{L}(\mathscr{X}_u)$ for which this reduction is obvious are those that are constant on the lumps, in the sense that for all $\hat{x}_u$ in $\hat{\mathscr{X}}_u$, $f(x_u) = f(y_u)$ for all x_u, y_u in $\Lambda^{-1}(\hat{x}_u)$. Clearly, this is equivalent to the existence of a real-valued function $\hat{f}$ on $\hat{\mathscr{X}}_u$ such that $f = \hat{f} \circ \Lambda$. If such a function $\hat{f}$ in $\mathscr{L}(\hat{\mathscr{X}}_u)$ exists, then the real-valued function f on $\mathscr{X}_u$ is called *lumpable with respect to* Λ. The reduction of a non-lumpable function f to $\hat{\mathscr{X}}_u$ is not unequivocally defined. In the remainder, we will make extensive use of the following two reductions:

$$\hat{f}_{\mathrm{L}}\colon \hat{\mathscr{X}}_u \to \mathbb{R}\colon \hat{x}_u \mapsto \min\{f(x_u)\colon x_u \in \hat{x}_u\} \quad \text{and} \quad \hat{f}_{\mathrm{U}}\colon \hat{\mathscr{X}}_u \to \mathbb{R}\colon \hat{x}_u \mapsto \max\{f(x_u)\colon x_u \in \hat{x}_u\}.$$

These two reductions are generalised versions of the reductions defined by Franceschinis and Muntz [9, p. 232]; they clearly provide bounds on the original function:

$$\hat{f}_{\mathrm{L}} \circ \Lambda \leq f \leq \hat{f}_{\mathrm{U}} \circ \Lambda. \tag{22}$$

Using Eqn. (22) and after some additional work, we establish the following result.

Theorem 8. *Consider a homogeneous Markov chain P and a lumping map $\Lambda\colon \mathscr{X} \to \hat{\mathscr{X}}$. Then for all u in $\mathscr{U}$, v in $\mathscr{U}_\emptyset$ with $\max u < \min v$, x_u in $\mathscr{X}_u$ and f in $\mathscr{L}(\mathscr{X}_{u\cup v})$,*

$$\underline{E}^{\mathrm{W}}_{\hat{\mathscr{Q}},\hat{\mathscr{M}}}(\hat{f}_{\mathrm{L}}(\hat{X}_u, \hat{X}_v) \mid \hat{X}_u = \hat{x}_u) \leq E(f(X_u, X_v) \mid X_u = x_u) \leq \overline{E}^{\mathrm{W}}_{\hat{\mathscr{Q}},\hat{\mathscr{M}}}(\hat{f}_{\mathrm{U}}(\hat{X}_u, \hat{X}_v) \mid \hat{X}_u = \hat{x}_u), \tag{23}$$

where $\hat{x}_u \coloneqq \Lambda(x_u)$.

Theorem 8 is similar to Corollary 7, but it is more useful as it provides bounds on the (conditional) expectation with respect to the original Markov chain—instead of with respect to a lumped stochastic process—of a real-valued function that depends on the state at any finite number of time points. As the lower and upper bounds of Theorem 8 are exactly the same as those of Corollary 7, the remarks we made earlier—right after Corollary 7—about the tractability of the computations needed to determine these bounds are applicable here as well.

We repeat here that our sole reason for limiting ourselves to functions that depend on a finite number of future time points in Theorem 8 is that, for now, this is the most general type of inference that can be dealt with using the framework of imprecise (continuous-time) Markov chains [12]. A possible alternative for the reader that is interested in more general inferences is the work of Katoen et al. [11], who also use lumping but follow a different approach that enables them to compute bounds on several specific types of inferences, including some that depend on an infinite number of future time points.

7. Bounding limit expectations

In many practical applications, see for instance [15, 21, 22], the Markov chain model of the system is used to determine the limit expectation $\lim_{t\to+\infty} E(f(X_t))$ of some real-valued function f on $\mathscr{X}$. Under some conditions on the transition rate matrix Q—ergodicity, see Section 7.1 further on—this limit expectation $\lim_{t\to+\infty} E(f(X_t))$ does not depend on the initial distribution π_0 of the Markov chain. In other applications—for instance those treated in [9, 10, 16]—one is interested in the long-term temporal average of $f(X_t)$:

$$\lim_{s\to+\infty} \frac{1}{s} \int_0^s f(X_t)\,\mathrm{d}t.[4]$$

Under a slightly more stringent condition on the transition rate matrix Q—irreducibility, see for instance [19, Theorem 3.8.1]—this long-term temporal average is (almost surely) equal to the limit expectation $\lim_{t\to+\infty} E(f(X_t))$, which then again does not depend on the initial distribution π_0 of the Markov chain. Clearly, methods to efficiently determine the limit expectation are therefore of tremendous practical interest.

There are plenty of methods available to determine the limit expectation; we refer to [26, Section 10] for an overview. However, it is well-known—see [9, 10, 16]—that these methods to precisely determine the limit expectation become computationally intractable for Markov chains with large state spaces. In this section, we therefore set out to obtain bounds on the limit expectation using the lumped lower transition rate operator $\underline{\hat{Q}}$. Our hope is that, if the lumped state space is sufficiently small, these bounds *can* be tractably computed. If this is the case, then we can bound inferences that we could not tractably compute using

[4] Note that in our current framework, the expectation of this inference cannot be expressed because it depends on the state at an infinite number of time points. This is not a problem, however, because we can always extend the domain of the coherent conditional probability P so that the expectation is well-defined. Furthermore, in the classical framework this turns out to be almost surely equal to the limit expectation, which *is* well-defined in our framework.

the precise methods for the original Markov chain. We refer to Sections 7.2 and 7.3 further on for more arguments regarding the tractability of the relevant methods.

We here consider two well-known methods to determine the limit expectation precisely. More specifically, we argue how these methods can be made computationally tractable again using the lumped transition rate operator $\underline{\hat{Q}}$ at the cost of imprecision—provided of course that $\underline{\hat{Q}}\hat{f}$ can be evaluated much more efficiently than Qf. First, however, we start with some general theory concerning ergodic Markov chains.

7.1. Ergodicity and irreducibility

Essential to the study of limit expectations and long-term temporal averages are the concepts of ergodicity and irreducibility. These two terms are not always used in the same sense by all authors; we here adhere to the use of Norris [19] and Tornambè [27].

Definition 2 (Definition 4.17 in [27])**.** A transition rate matrix Q is *ergodic* if there is a distribution π_∞ on $\mathscr{X}$ such that, for all f in $\mathscr{L}(\mathscr{X})$ and π_0 in $\mathscr{D}(\mathscr{X})$, $\lim_{t\to+\infty}\langle\pi_0, T_0^t f\rangle = \langle\pi_\infty, f\rangle$.

We call the distribution π_∞ corresponding to an ergodic transition rate matrix Q the *limit distribution.* The reason for this name is that if Q is the transition rate matrix of a homogeneous Markov chain P, then

$$\lim_{t\to+\infty} E(f(X_t)) = \lim_{t\to+\infty}\langle\pi_0, T_0^t f\rangle = \langle\pi_\infty, f\rangle \qquad \text{for all } f \in \mathscr{L}(\mathscr{X}),$$

where the first equality follows from Eqn. (8). It is well-known—see for example [27, Theorem 4.12]—that if the transition rate matrix Q is ergodic, then the corresponding limit distribution π_∞ is the unique distribution that satisfies the equilibrium condition

$$(\forall y \in \mathscr{X}) \sum_{x\in\mathscr{X}} \pi_\infty(x) Q(x, y) = 0. \tag{24}$$

Many equivalent necessary and sufficient conditions for ergodicity exist; see for instance [19, Theorem 3.2.1]. The following is the one that is arguably the most easy to check for a given transition rate matrix. It is based on the accessibility relation $\cdot \rightsquigarrow \cdot$. We say that a state x is *accessible* from a state y, denoted by $y \rightsquigarrow x$, if there is a sequence $y = x_0, x_1 \dots, x_n = x$ in $\mathscr{X}$ such that $Q(x_{i-1}, x_i) > 0$ for all i in $\{1, \dots, n\}$. Note that any state x is always accessible from itself because the sequence $x = x_0 = x_n = x$ (with $n = 0$) trivially satisfies this condition.

Proposition 9. *A transition rate matrix Q is ergodic if and only if*

$$\mathscr{X}_{\text{top}} \coloneqq \{x \in \mathscr{X} : (\forall y \in \mathscr{X})\ y \rightsquigarrow x\} \neq \emptyset.$$

A transition rate matrix Q is said to be *irreducible* if $\mathscr{X}_{\text{top}} = \mathscr{X}$; see for instance [19, Sections 1.2 and 3.2]. Most authors limit themselves to irreducible instead of the more general ergodic transition rate matrices whenever they are interested in limit expectations. Their reason for doing so is the following result, which states that as far as the limit expectation is concerned, one can limit the state space to the top class $\mathscr{X}_{\text{top}}$.

Proposition 10. *Let Q be an ergodic transition rate matrix. Then the matrix Q' on $\mathscr{L}(\mathscr{X}_{\text{top}})$, defined by*

$$Q'(x, y) \coloneqq Q(x, y) \text{ for all } x, y \text{ in } \mathscr{X}_{\text{top}},$$

is an irreducible transition rate matrix. Furthermore, for all f in $\mathscr{L}(\mathscr{X})$, $\langle\pi_\infty, f\rangle = \langle\pi'_\infty, f'\rangle$, where π'_∞ is the limit distribution of Q' and f' is the restriction of f to $\mathscr{X}_{\text{top}}$.

Observe that in order to use this result, one has to explicitly determine the top class. If this top class can be easily obtained, then reducing the state space to this top class makes sense because it will speed up all methods to determine the limit expectation $\langle\pi_\infty, f\rangle$. However, this is not always the case, as it might occur that checking that the top class is non-empty is straightforward, while explicitly determining the top class is not. Therefore, in the remainder, we will not a priori limit ourselves to irreducible transition rate matrices but will always consider general ergodic ones.

7.2. Bounding limit expectations iteratively

The first method we consider is based on a link between Markov chains in discrete and continuous time. Here, we are especially interested in the fact that $(I + \delta Q)$ is a transition matrix—at least under suitable conditions on δ. It is essentially well-known that $(I + \delta Q)$ is ergodic if Q is ergodic, and that both have the same limit distribution. The following result establishes these links in the form that we will need them.

Proposition 11. *If Q is an ergodic transition rate matrix, then for all f in $\mathscr{L}(\mathscr{X})$, δ in $\mathbb{R}_{>0}$ with $\delta\|Q\| < 2$, and n in $\mathbb{N}_0$,*

$$\min(I + \delta Q)^n f \leq \langle \pi_\infty, f\rangle \leq \max(I + \delta Q)^n f.$$

Furthermore, the lower and upper bounds in this expression become monotonously tighter with increasing n, and converge to $\langle \pi_\infty, f\rangle$ as n approaches $+\infty$.

Note that the step size δ in Proposition 11 is only required to be sufficiently small such that $\delta\|Q\| < 2$. Empirically, we observe that the convergence of the bounds is faster—in the sense that we need smaller n—for larger values of δ.

Note, however, that if the size of the original state space is too large, then the bounds in Proposition 11 cannot be tractably computed. One way to make the computations tractable is to "replace" the transition rate matrix Q with the lumped lower transition rate operator $\underline{\hat{Q}}$ and its conjugate

$$\hat{\overline{Q}}\colon \mathscr{L}(\hat{\mathscr{X}}) \to \mathscr{L}(\hat{\mathscr{X}})\colon \hat{f} \mapsto \hat{\overline{Q}}\hat{f} := -\underline{\hat{Q}}(-\hat{f}).$$

The following result establishes that this replacement is allowed.

Theorem 12. *Consider an ergodic transition rate matrix Q and a lumping map $\Lambda\colon \mathscr{X} \to \hat{\mathscr{X}}$. Then for all f in $\mathscr{L}(\mathscr{X})$, δ in $\mathbb{R}_{>0}$ with $\delta\|Q\| < 2$, and n in $\mathbb{N}_0$,*

$$\min(I + \delta\underline{\hat{Q}})^n \hat{f}_{\mathrm{L}} \leq \langle \pi_\infty, f\rangle \leq \max(I + \delta\hat{\overline{Q}})^n \hat{f}_{\mathrm{U}}.$$

Moreover, for fixed δ, the lower and upper bounds in this expression become monotonously tighter with increasing n, and each converges to a—possibly different—limit as n approaches $+\infty$. If Q is furthermore irreducible, $(I + \delta\underline{\hat{Q}})^n \hat{f}_{\mathrm{L}}$ and $(I + \delta\hat{\overline{Q}})^n \hat{f}_{\mathrm{U}}$ both converge to a—possibly different—constant function as n approaches $+\infty$.

Theorem 12 naturally suggests an iterative method to determine guaranteed bounds on the limit expectation $\langle \pi_\infty, f\rangle$. For the lower bound, one simply needs to (i) choose a step size δ such that $\delta\|Q\| < 2$; (ii) iteratively compute $\hat{g}_i := (I + \delta\underline{\hat{Q}})g_{i-1}$ with initial condition $g_0 := \hat{f}_{\mathrm{L}}$; and (iii) stop after n iterations if $\min g_n$—or in case Q is irreducible, g_n—has empirically converged; the lower bound is then $\min(I + \delta\underline{\hat{Q}})^n \hat{f}_{\mathrm{L}} = \min g_n$. As a consequence of the conjugacy of $\hat{\overline{Q}}$, the upper bound can be iteratively obtained by applying the iterative scheme for the lower bound with initial condition $g_0 := -\hat{f}_{\mathrm{U}}$; the upper bound is then $\max(I + \delta\hat{\overline{Q}})^n \hat{f}_{\mathrm{U}} = -\min g_n = -\min(I + \delta\underline{\hat{Q}})^n g_0$. Empirically, we observe that larger step sizes δ result in faster convergence, in the sense that less iterations are required. The influence of the step size δ on the tightness of the bounds is something that we have not properly investigated yet. Our limited experiments suggest that a smaller step size δ results in tighter bounds, although after some threshold—that depends on the specific model being used and that can be rather large—the tightness does not seem to change any more.

7.3. Bounding limit expectations with a linear program

Another popular method to determine the limit expectation $\langle \pi_\infty, f\rangle$ is to first determine the limit distribution π_∞ and then compute the inner product directly. Recall from Section 7.1 that π_∞ is the *unique* distribution that satisfies the equilibrium condition, so—in theory—we can determine it by explicitly solving Eqn. (24), see for instance [26, Section 10.2]. Unfortunately, if the state space $\mathscr{X}$ is large then solving the resulting linear system of $|\mathscr{X}|$ equations becomes computationally infeasible. Therefore, we here combine the equilibrium condition for several different states y; subsequent manipulation of the resulting expressions then allows us to establish the following result.

Theorem 13. *Consider an ergodic transition rate matrix Q and a lumping map $\Lambda\colon \mathscr{X} \to \hat{\mathscr{X}}$. Then for all $\hat{\mathcal{A}} \subseteq \mathcal{P}(\hat{\mathscr{X}})$*[5] *and f in $\mathscr{L}(\mathscr{X})$,*

$$\min\{\langle \hat{\pi}, \hat{f}_{\mathrm{L}}\rangle \colon \hat{\pi} \in \mathscr{D}_{\hat{\mathcal{A}}}\} \leq \langle \pi_\infty, f\rangle \leq \max\{\langle \hat{\pi}, \hat{f}_U\rangle \colon \hat{\pi} \in \mathscr{D}_{\hat{\mathcal{A}}}\}$$

with

$$\mathscr{D}_{\hat{\mathcal{A}}} \coloneqq \{\hat{\pi} \in \mathscr{D}(\hat{\mathscr{X}}) \colon (\forall \hat{A} \in \hat{\mathcal{A}})\ \langle \hat{\pi}, \underline{\hat{Q}}\mathbb{I}_{\hat{A}}\rangle \leq 0\}.$$

It might not immediately look like it, but closer inspection shows us that the two optimisations in the expression above are in fact straightforward linear programs. The variables of these linear programs are the components of $\hat{\pi}$. Since $\hat{\pi}$ is a distribution on $\hat{\mathscr{X}}$, there are $|\hat{\mathscr{X}}|$ variables and already $|\hat{\mathscr{X}}| + 1$ constraints: $\hat{\pi}(\hat{x}) \geq 0$ for all $\hat{x}$ in $\hat{\mathscr{X}}$ and $\sum_{\hat{x} \in \hat{\mathscr{X}}} \hat{\pi}(\hat{x}) = 1$. Some $|\hat{\mathcal{A}}|$ additional constraints are induced by the requirement that $\hat{\pi}$ is an element of $\mathscr{D}_{\hat{\mathcal{A}}}$: for all $\hat{A}$ in $\hat{\mathcal{A}}$, $\langle \hat{\pi}, \underline{\hat{Q}}\mathbb{I}_{\hat{A}}\rangle \leq 0$.

An obvious issue when applying the method of Theorem 13 is how to choose the collection of subsets $\hat{\mathcal{A}}$. First and foremost, as $\underline{\hat{Q}}$ is a lower transition rate operator, $\underline{\hat{Q}}\mathbb{I}_\emptyset = 0$ and $\underline{\hat{Q}}\mathbb{I}_{\hat{\mathscr{X}}} = 0$. Therefore, the condition

$$\langle \hat{\pi}, \underline{\hat{Q}}\mathbb{I}_{\hat{A}}\rangle \leq 0$$

is always satisfied if $\hat{A}$ is equal to $\emptyset$ or $\hat{\mathscr{X}}$. Knowing this, one obvious choice is $\hat{\mathcal{A}} = \mathcal{P}(\hat{\mathscr{X}}) \setminus \{\emptyset, \hat{\mathscr{X}}\}$, which leads to a linear program with $|\hat{\mathscr{X}}| + 2^{|\hat{\mathscr{X}}|} - 1$ distinct constraints. As the number of constraints scales exponentially with the number of lumped states, this becomes computationally intractable if the lumped state space $\hat{\mathscr{X}}$ is large. An obvious alternative choice is to consider the collection of all singletons: $\hat{\mathcal{A}} = \{\{\hat{x}\} \colon \hat{x} \in \hat{\mathscr{X}}\}$. This choice results in $2|\hat{\mathscr{X}}| + 1$ distinct constraints for the linear program, which is certainly tractable. From an implementation point of view, it makes sense to also add the complements of the singletons to the collection $\hat{\mathcal{A}}$. The reason for this is that the functions $\mathbb{I}_{A^c}$ need not be explicitly constructed because $\underline{\hat{Q}}\mathbb{I}_{\hat{A}^c} = \underline{\hat{Q}}(-\mathbb{I}_{\hat{A}})$—as can be easily verified. In this case, the linear program has $3|\hat{\mathscr{X}}| + 1$ constraints. The tractability of this choice for a smaller collection $\hat{\mathcal{A}}$ can come at the cost of reduced tightness of the resulting bounds compared to using the power set, because the set $\mathscr{D}_{\hat{\mathcal{A}}}$ is less constrained and hence can be larger. This trade-off between tractability and tightness is something that we leave for future research.

8. Numerical assessment

We leave the numerical assessment of Theorem 8 for future work. Our main reason for this is that the performance of this method is tied to the performance of the method used to approximate $\underline{\hat{T}}_t^s f$, and this would lead us too far astray. For a preliminary study of the performance of some methods to approximate $\underline{\hat{T}}_t^s f$, we refer to [25]. We here fully focus on bounding limit expectations: we compare the methods implied by Theorem 12 and 13 with the methods of Franceschinis and Muntz [9] and Buchholz [10] in Sections 8.1–8.3, and consider the large-scale tractability of our methods in Section 8.4.

8.1. A closed queueing network

Following Franceschinis and Muntz [9] and Buchholz [10], we test our methods on the closed queueing network depicted in Figure 1. This closed queueing network is populated by N customers and consists of a single server, denoted by $\mathtt{S}_0$, in series with K parallel servers, denoted by $\mathtt{S}_1, \ldots, \mathtt{S}_K$. In order not to unnecessarily complicate our exposition, we will assume that K is even. As is clear from Figure 1, the customers alternatingly visit the server $\mathtt{S}_0$ and one of the parallel servers $\mathtt{S}_1, \ldots, \mathtt{S}_K$.

One obvious way to describe this closed queueing network is to use states of the form $(i, i_1, \ldots, i_K)$, where i is the number of customers in the single server and i_k is the number of customers in the k-th parallel server. This state description yields the state space

$$\mathscr{X} \coloneqq \left\{(i, i_1, \ldots, i_K) \in \{0, 1, \ldots, N\}^{K+1} \colon i + \sum_{k=1}^{K} i_k = N\right\}.$$

[5] Here and in the remainder, we denote the power set of the set S by $\mathcal{P}(S)$.

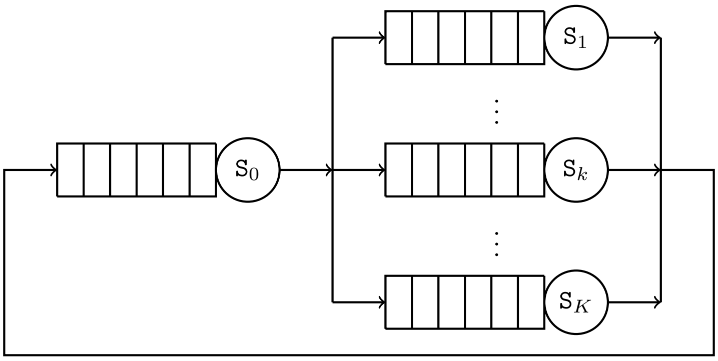

Figure 1: The closed queueing network

As a reduced, higher-order state description, Franceschinis and Muntz [9] propose to use $(i, j_0, \ldots, j_N)$, where i is again the number of customers in the single server and j_ℓ is the number of parallel servers that have ℓ customers. This yields the lumped state space

$$\hat{\mathscr{X}} := \left\{ (i, j_0, \ldots, j_N) \in \{0, 1, \ldots, N\} \times \{0, 1, \ldots, K\}^{N+1} : i + \sum_{n=1}^{N} n j_n = N, \sum_{n=0}^{N} j_n = K \right\}.$$

The lumped state space $\hat{\mathscr{X}}$ is significantly smaller than the original state space $\mathscr{X}$, as is clear from the number of states and lumps reported in [9, Table 3] and Table 2.

We are interested in two performance measures for the first server $\mathtt{S}_0$ of the queueing network: (i) the population—or, in the words of Franceschinis and Muntz, the mean queue length— of this server (POP), which is the limit expectation of the lumpable function

$$f : \mathscr{X} \to \mathbb{R} : (i, i_1, \ldots, i_K) \mapsto i, \text{ with } \hat{f} : \hat{\mathscr{X}} \to \mathbb{R} : (i, j_0, \ldots, j_N) \mapsto i$$

and (ii) the throughput at this server (TP), which is the limit expectation of the lumpable function

$$g : \mathscr{X} \to \mathbb{R} : (i, i_1, \ldots, i_K) \mapsto \begin{cases} \mu & \text{if } i < N, \\ 0 & \text{otherwise,} \end{cases} \text{ with } \hat{g} : \hat{\mathscr{X}} \to \mathbb{R} : (i, j_0, \ldots, j_N) \mapsto \begin{cases} \mu & \text{if } i < N, \\ 0 & \text{otherwise,} \end{cases}$$

where μ is the rate of the distribution of the service time of $\mathtt{S}_0$.

8.2. Exponential service time

Franceschinis and Muntz [9] assume that the single server $\mathtt{S}_0$ has a service time that is exponentially distributed with rate μ—or equivalently, mean service time $1/\mu$. After its service in server $\mathtt{S}_0$ has completed, a customer randomly joins the queue of one of the parallel servers $\mathtt{S}_1$, ..., $\mathtt{S}_K$, and for the sake of simplicity each choice is assumed to be equally probable. The parallel servers also have a service time that is exponentially distributed: half of the parallel servers have rate λ_1 and the other half have rate λ_2, with $\lambda_1 < \lambda_2$. Note that due to symmetry, without loss of generality, we may assume that $\mathtt{S}_1$, ..., $\mathtt{S}_{\frac{K}{2}}$ have rate λ_1 and the remaining servers have rate λ_2 Under these assumptions, we can model the system as a homogeneous Markov chain with state space $\mathscr{X}$, but the lumped state space $\hat{\mathscr{X}}$ is *not* sufficiently detailed to allow a homogeneous Markov chain model.

The lumping map Λ that models the relation between these two state descriptions is easily obtained. Hence, we obtain the lumped lower transition rate operator $\underline{\hat{Q}}$ according to Eqn. (20). In this case, it turns out that the minimisation reduces to

$$[\underline{\hat{Q}}\hat{f}](i, j_0, \dots, j_N) = \frac{\mu}{N} \sum_{n=1}^{N} (\hat{f}(i-1, j_0, \dots, j_{n-1}, j_n + 1, j_{n+1} \dots, j_N) - \hat{f}(i, j_0, \dots, j_N))$$
$$+ \sum_{n \in \mathcal{N}} \lambda'_n (\hat{f}(i+1, j_0, \dots, j_{n-1}, j_n - 1, j_{n+1} \dots, j_N) - \hat{f}(i, j_0, \dots, j_N)), \quad (25)$$

where the first summation is only present in case $i > 0$. In this expression, $\mathcal{N}$ is the set of all indices n such that $j_n > 0$. The rates λ'_n are equal to either λ_1 or λ_2. One has to choose their values in such a way that the resulting value of the expression in Eqn. (25) is minimised, under the condition that at most $K/2$ can have value λ_1 and at most $K/2$ can have value λ_2.

Table 1: Comparison of the bounds obtained by using Theorems 12 and 13 with those obtained by the method presented in [9, Section 3.2]. Model parameters: $K = 4$, $N = 5$, $\mu = 1$, $\lambda_1 = 1$, $\lambda_2 = 1.01$. Computation parameters: $\hat{\mathcal{A}}_1$ consists of all the singletons, $\hat{\mathcal{A}}_2$ consists of all the singletons and their complements.

				Theorem 12				Theorem 13			
		[9, Table 2]		$\delta = 1.8/\|Q\|$		$\delta = 0.9/\|Q\|$		$\hat{\mathcal{A}}_1$		$\hat{\mathcal{A}}_2$	
	Exact	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
POP	3.7262	3.6141	3.7493	3.7195	3.7336	3.7195	3.7336	3.6162	3.7487	3.7113	3.7414
TP	0.9828	0.9676	0.9835	0.9825	0.9831	0.9825	0.9831	0.9679	0.9835	0.9823	0.9834

Franceschinis and Muntz [9] compute bounds on the performance measures for the closed queueing network with $K = 4$ parallel servers and $N = 5$ customers. These parameters yield a state space $\mathscr{X}$ with 126 states and a lumped state space $\hat{\mathscr{X}}$ with 18 states. For the service time distributions, they use the parameters $\mu = 1$, $\lambda_1 = 1$ and $\lambda_2 = 1.01$. In Table 1, we report bounds on the performance measures that we obtain with the two methods based on Theorems 12 and 13. Note that Franceschinis and Muntz actually report the limit expectation of $N - i$ instead of that of i in [9, Table 2], but we have transformed their bounds to correspond to our setting.

For all cases reported in Table 1, our bounds are tighter than those of Franceschinis and Muntz [9]. Furthermore, the bounds obtained with the iterative method based on Theorem 12 are tighter than those obtained with the linear programming method based on Theorem 13. For the method based on Theorem 12, the obtain bounds are the same—at least up to the reported precision—for both choices of step sizes. We observed that in this case, halving the step size results in a doubling of the number of iterations required to reach empirical convergence. For the method based on Theorem 13, the obtained bounds are noticeably tighter if we use the collection $\hat{\mathcal{A}}_2$ that consists of all singletons and their complements instead of the collection $\hat{\mathcal{A}}_1$ that consists of only the singletons. We have chosen not to use the power set as collection because we believe that adding $2^{|\hat{\mathscr{X}}|} - 2 = 65\,534$ constraints on 18 variables is a bit excessive.

8.3. Erlang-2 serivce time

Buchholz [10] considers a slightly changed version of the closed queueing network: instead of assuming that the service time of $\mathtt{S}_0$ is exponentially distributed, he assumes an Erlang-2 distribution with mean service time $1/\mu$. This assumption still allows for a homogeneous Markov chain model, be it that the component i is replaced by two components in both the full state description and the lumped state description. The lumped lower transition rate operator $\underline{\hat{Q}}$ obtained with Eqn. (20) reduces to an expression similar to that of Eqn. (25).

Buchholz [10] considers several combinations of the number of parallel servers K and the number of customers N. For the service time distributions, he uses the parameters $\mu = 5$, $\lambda_1 = 1$ and $\lambda_2 = 1 + \epsilon$, with ϵ equal to 0.1 or 0.01. In Table 1, we report bounds on the throughput, obtained with the two methods based on Theorems 12 and 13, and compare those with the bounds obtained by Buchholz. First and foremost, we

Table 2: Comparison of the bounds obtained on the throughput by using Theorems 12 and 13 with those obtained by the method presented in [10, Section 4]. Model parameters: $\mu = 5$, $\lambda_1 = 1$, $\lambda_2 = \lambda_1 + \epsilon$. Computation parameters: $\hat{\mathcal{A}}_1$ consists of all the singletons, $\hat{\mathcal{A}}_2$ consists of all the singletons and their complements.

							Theorem 12				Theorem 13			
					[10, Figure 3]		$\delta = 1.8/\lVert Q\rVert$		$\delta = 0.9/\lVert Q\rVert$		$\hat{\mathcal{A}}_1$		$\hat{\mathcal{A}}_2$	
K	N	$\lvert\mathscr{X}\rvert$	$\lvert\hat{\mathscr{X}}\rvert$	Exact	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
$\epsilon = 0.1$														
4	6	336	45	2.611	2.509	2.730	2.529	2.713	2.531	2.713	1.928	3.181	2.401	2.820
4	8	825	91	2.892	2.784	3.028	2.813	3.003	2.814	3.003	1.982	3.633	2.582	3.171
4	10	1716	165	3.090	2.973	3.239	3.012	3.207	3.014	3.207	2.002	3.961	2.670	3.435
6	8	4719	108	3.486	3.365	3.624	3.382	3.613	3.383	3.613	2.068	4.188	3.069	3.846
6	10	13 013	215	3.802	3.675	3.984	3.707	3.930	3.708	3.930	2.083	4.515	3.191	4.244
8	10	68 068	232	4.202	4.087	4.327	4.103	4.320	4.104	4.320	2.111	4.736	3.542	4.595
$\epsilon = 0.01$														
4	6	336	45	2.520	2.509	2.532	2.511	2.530	2.511	2.530	2.428	2.591	2.500	2.541
4	8	825	91	2.793	2.780	2.806	2.784	2.803	2.784	2.803	2.655	2.894	2.764	2.821
4	10	1716	165	2.984	2.971	2.998	2.975	2.995	2.976	2.995	2.802	3.116	2.948	3.020
6	8	4719	108	3.378	3.365	3.392	3.367	3.391	3.367	3.391	3.121	3.488	3.340	3.416
6	10	13 013	215	3.689	3.675	3.704	3.678	3.702	3.678	3.702	3.336	3.821	3.639	3.738
8	10	68 068	232	4.100	4.087	4.113	4.088	4.113	4.088	4.113	3.609	4.213	4.047	4.151

observe that the bounds obtained with the iterative method based on Theorem 12 are tighter than those obtained with the method of Buchholz [10], which in turn are tighter than those obtained with the linear programming method based on Theorem 13; hence, as was also the case in Section 8.1, our iterative method outperforms the linear programming method. Second, we observe that halving the step size δ in the iterative method does increase the tightness of the bounds, be it only marginally. Adding the complements to the collection $\hat{\mathcal{A}}$ in the linear programming method clearly results in tighter bounds, as was also the case in Section 8.1.

We end with two caveats. First, we have also conducted experiments for the population. However, the exact values we obtain for the population do not lie in the intervals reported in [10, Fig. 3]. We have not managed to clear out whether this is due to an error on our part or not. Since this prevents a proper comparison of our bounds with those reported in [10, Fig. 3], we have chosen not to report any bounds on the population. Second, Buchholz mentions in [10, Section 5.3] that he assumes "service rates between 1.0 and $1.0 + \epsilon$", but it is unclear to us if he additionally assumes that half of the servers have rate 1.0 and the other half of the servers have rate $1.0 + \epsilon$. If he does not make this additional assumption, his bounds hold for the more general setting that the transition rate matrix Q is only known to belong to some set of transition rate matrices. Our methods can be adapted to this more general setting as well, in a similar fashion to the approach followed in [16], but we leave this for future work. We here only mention that this yields a lumped lower transition rate operator $\underline{\hat{Q}}$ that is very similar to that of Eqn. (25); it turns out that the λ'_n's can all be independently optimised without taking into account the restrictions on how many of them can take the value λ_1 or λ_2.

8.4. Large-scale tractability

We have chosen to limit the scenarios for our numerical experiments to those scenarios that were also considered by Franceschinis and Muntz [9] and Buchholz [10]. Our reason for this is two-fold. First, this allows us to compare our bounds with the exact result. Second, this allow us to compare our bounds with those obtained by Franceschinis and Muntz [9] and Buchholz [10] without needing to implement their methods ourselves. As we have previously argued, our experiments seem to suggest that our bounds—or at least those obtained using the iterative approximation method—are tighter than those obtained by the existing methods.

As the precise value of the relevant limit expectations can still be tractably computed, these scenarios do not correspond to the intended range of applications. We leave a thorough study of the tractability for large-scale Markov chains, as well as a comparison with the scalability of the existing methods, for future work. However, this does not mean that we are oblivious to the scalability of our methods. Quite the contrary, we have previously studied the tractability of our iterative approximation method for large-scale Markov chains in [16], where we used the iterative approximation method based on Theorem 12 to compute (bounds on) performance measures of a specific telecommunication system. The largest Markov chain model that we consider there has 1 221 759 states, which gets reduced to 35 301 states for the lumped imprecise Markov chain. For that large-scale model, the precise methods turn out to be computationally infeasible, while the computations for our iterative approximation method are still tractable.

9. Lumpability

Until now, our sole objective has been to use the lumped imprecise Markov chain to determine bounds on expectations with respect to the original Markov chain. Recall from the Introduction that this is only a secondary setting in which the lumping of Markov chains has been previously studied. In this section, we briefly return to the original setting of lumping in Markov chains, and ask ourselves the question if the lumped process is again a homogeneous Markov chain. Crucial to this question is the concept of lumpability.

Definition 3. The transition rate matrix Q is *lumpable* with respect to the lumping map $\Lambda\colon \mathscr{X} \to \hat{\mathscr{X}}$ if there is a $\hat{Q}$ in $\mathscr{R}(\hat{\mathscr{X}})$ such that

$$(\forall \hat{x}, \hat{y} \in \hat{\mathscr{X}})(\forall x \in \hat{x}) \sum_{y \in \hat{y}} Q(x,y) = \hat{Q}(\hat{x}, \hat{y}). \tag{26}$$

Burke and Rosenblatt consider homogeneous Markov chains and prove in [2, Theorem 4] that lumpability of the transition rate matrix is necessary and sufficient for the lumped stochastic process to be a Markov chain, regardless of the initial distribution. In [3, Theorem 2.4], Ball and Yeo establish that if the original homogeneous Markov chain has a denumerable state space and is irreducible, then lumpability of the transition rate matrix is also a necessary and sufficient condition for the lumped stochastic process to be a homogeneous Markov chain.

Since lumpability is clearly a strong property, we expect that our previous results can be simplified under the condition that it holds. First and foremost, we have the following obvious result.

Proposition 14. *Consider a transition rate matrix Q and lumping map $\Lambda\colon \mathscr{X} \to \hat{\mathscr{X}}$. Then Q is lumpable with respect to Λ if and only if $\underline{\hat{Q}}$ is linear, or equivalently, if and only if $\hat{\mathscr{Q}}$ is a singleton. In this case, $\hat{\mathscr{Q}} = \{\hat{Q}\}$ and $\underline{\hat{Q}} = \hat{Q}$, where $\hat{Q}$ is the unique transition rate matrix that satisfies Eqn.* (26)*.*

We combine Proposition 14 with Theorem 6 to immediately obtain the following very strong result.

Corollary 15. *Consider a homogeneous Markov chain P with transition rate matrix Q, and a lumping map $\Lambda\colon \mathscr{X} \to \hat{\mathscr{X}}$. If Q is lumpable with respect to Λ, then the lumped stochastic process $\hat{P}$ is uniquely defined and equal to the homogeneous Markov chain characterised by the lumped initial distribution $\hat{\pi}_0$ and the transition rate matrix $\hat{Q}$.*

Note that Corollary 15 is similar to [2, Theorem 4], although it only provides the sufficiency and not the necessity of the lumpability condition. However, if we follow the strategy that Burke and Rosenblatt use in the proof of [2, Theorem 4], then we also obtain the necessity of the condition, at least if—like Burke and Rosenblatt in [2, Theorem 4]—we demand that the lumped stochastic process is a homogeneous Markov chain for *any* initial distribution.

For our present purposes, what is particularly interesting about the above results, is that we can specialise the results of Sections 6 and 7 if combine them with Proposition 14 and/or Corollary 15. We limit ourselves to lumpable functions because this allows for more elegant statements. Our first result is then a specialisation of Theorem 8 that is useful in the same setting, and follows almost immediately from Corollary 15.

Corollary 16. *Consider a homogeneous Markov chain P with transition rate matrix Q, and a lumping map $\Lambda\colon \mathscr{X} \to \hat{\mathscr{X}}$. If Q is lumpable with respect to Λ, then for all u in $\mathscr{U}$, v in $\mathscr{U}_\emptyset$ with $\max u < \min v$, x_u in $\mathscr{X}_u$ and all lumpable f in $\mathscr{L}(\mathscr{X}_{u \cup v})$,*

$$E(f(X_u, X_v) \mid X_u = x_u) = \hat{E}(\hat{f}(\hat{X}_u, \hat{X}_v) \mid \hat{X}_u = \hat{x}_u),$$

with $\hat{x}_u := \Lambda(x_u)$ and where $\hat{E}$ denotes the expectation with respect to the lumped homogeneous Markov chain $\hat{P}$.

We now turn to specialising the results of Section 7. While we could simply combine Theorems 12 and 13 with Corollary 15, the following result makes more sense.

Theorem 17. *Consider an ergodic transition rate matrix Q and a lumping map Λ. If Q is lumpable with respect to Λ, then $\hat{Q}$ is ergodic. Furthermore, for any lumpable f in $\mathscr{L}(\mathscr{X})$,*

$$\langle \pi_\infty, f \rangle = \langle \hat{\pi}_\infty, \hat{f} \rangle,$$

where $\hat{\pi}_\infty$ denotes the limit distribution of $\hat{Q}$.

The reason why Theorem 17 is more sensible than simply specialising Theorems 12 and 13 is that we can use any of the standard methods to determine the limit expectation $\langle \hat{\pi}_\infty, \hat{f} \rangle$, instead of only our two (approximate) methods.

10. Conclusions

Broadly speaking, we can conclude that imprecise Markov chains are not only a robust uncertainty model—as they were originally intended to be—but also a useful computational tool for determining bounds on inferences for large-scale homogeneous (continuous-time) Markov chains. More concretely, we have shown that lumping states in a homogeneous Markov chain inevitably introduces imprecision, in the sense that—in general—we cannot exactly determine the transition probabilities of the lumped stochastic process without also explicitly determining the transition probabilities of the original Markov chain. However, we can easily characterise a set of processes that definitely contains the lumped process, in the form of an imprecise Markov chain. Using this imprecise Markov chain, we can then determine guaranteed lower and upper bounds on the (conditional) expectation—with respect to the original Markov chain—of a real-valued function on the state of the system at any finite number of time points. Furthermore, we also presented two methods to bound the limit expectations of ergodic Markov chains. From a practical point of view, these results are essential tools in cases where state space explosion occurs: they allow us to determine guaranteed lower and upper bounds on inferences that we otherwise could not determine at all!

Regarding future work, we envision the following. For starters, a more thorough numerical assessment of the methods outlined in Sections 6 and 7.3 is necessary. Furthermore, we believe that almost all of our result can be adapted to the setting of *discrete-time* Markov chains. Moreover, Theorems 12 and 13 can also be quite easily extended to the setting in which we are interested in the limit expectation associated with some ergodic homogeneous Markov chain P, but where we only know that its transition rate matrix Q is contained in some non-empty and bounded set of (ergodic) transition rate matrices. Finally, it would be of theoretical as well as practical interest to determine bounds on the (conditional) expectation of functions that depend on the state at infinitely many time points.

References

[1] J. G. Kemeny, J. L. Snell, Finite Markov Chains, Springer-Verlag, 1960.

[2] C. J. Burke, M. Rosenblatt, A Markovian function of a Markov chain, The Annals of Mathematical Statistics 29 (1958) 1112–1122. doi:10.1214/aoms/1177706444.

[3] F. Ball, G. F. Yeo, Lumpability and marginalisability for continuous-time Markov chains, Journal of Applied Probability 30 (1993) 518–528. doi:10.2307/3214762.

[4] G. Rubino, B. Sericola, A finite characterization of weak lumpable Markov processes. Part II: The continuous time case, Stochastic Processes and their Applications 45 (1993) 115–125. doi:10.1016/0304-4149(93)90063-A.

[5] J. Hachigian, Collapsed Markov chains and the Chapman-Kolmogorov equation, The Annals of Mathematical Statistics 34 (1963) 233–237.

[6] D. Hartfiel, Lumping in Markov set-chains, Stochastic Processes and their Applications 50 (1994) 275–279. doi:10.1016/0304-4149(94)90124-4.

[7] S. Derisavi, H. Hermanns, W. H. Sanders, Optimal state-space lumping in Markov chains, Information Processing Letters 87 (2003) 309–315. doi:10.1016/S0020-0190(03)00343-0.

[8] A. Valmari, G. Franceschinis, Simple $O(m \log n)$ time Markov chain lumping, in: Tools and Algorithms for the Construction and Analysis of Systems, Springer Berlin Heidelberg, 2010, pp. 38–52.

[9] G. Franceschinis, R. R. Muntz, Bounds for quasi-lumpable Markov chains, Performance Evaluation 20 (1994) 223–243. doi:10.1016/0166-5316(94)90015-9.

[10] P. Buchholz, An improved method for bounding stationary measures of finite Markov processes, Performance Evaluation 62 (2005) 349–365. doi:10.1016/j.peva.2005.07.002.

[11] J.-P. Katoen, D. Klink, M. Leucker, V. Wolf, Three-valued abstraction for probabilistic systems, The Journal of Logic and Algebraic Programming 81 (2012) 356–389. doi:10.1016/j.jlap.2012.03.007, special Issue: NWPT 2009.

[12] T. Krak, J. De Bock, A. Siebes, Imprecise continuous-time Markov chains, International Journal of Approximate Reasoning 88 (2017) 452–528. doi:10.1016/j.ijar.2017.06.012.

[13] J. De Bock, The limit behaviour of imprecise continuous-time Markov chains, Journal of Nonlinear Science 27 (2017) 159–196. doi:10.1007/s00332-016-9328-3.

[14] D. Škulj, Efficient computation of the bounds of continuous time imprecise Markov chains, Applied Mathematics and Computation 250 (2015) 165–180. doi:10.1016/j.amc.2014.10.092.

[15] C. Rottondi, A. Erreygers, G. Verticale, J. De Bock, Modelling spectrum assignment in a two-service flexi-grid optical link with imprecise continuous-time Markov chains, in: Proceedings of DRCN 2017, VDE Verlag, 2017, pp. 39–46.

[16] A. Erreygers, C. Rottondi, G. Verticale, J. De Bock, Imprecise Markov models for scalable and robust performance evaluation of flexi-grid spectrum allocation policies, IEEE Transactions on Communications 66 (2018) 5401–5414. doi:10.1109/TCOMM.2018.2846235.

[17] A. Erreygers, J. De Bock, Computing inferences for large-scale continuous-time Markov chains by combining lumping with imprecision, in: Uncertainty Modelling in Data Science (Proceedings of SMPS2018), Springer International Publishing, 2018, pp. 78–86. Extended preprint: Erreygers and De Bock [24].

[18] E. Regazzini, Finitely additive conditional probabilities, Rendiconti del Seminario Matematico e Fisico di Milano 55 (1985) 69–89. doi:10.1007/BF02924866.

[19] J. R. Norris, Markov chains, Cambridge University Press, 1997. doi:10.1017/CBO9780511810633.

[20] W. J. Anderson, Continuous-Time Markov Chains, Springer-Verlag, 1991. doi:10.1007/978-1-4612-3038-0.

[21] A. Ganguly, T. Petrov, H. Koeppl, Markov chain aggregation and its applications to combinatorial reaction networks, Journal of Mathematical Biology 69 (2014) 767–797. doi:10.1007/s00285-013-0738-7.

[22] M. C. Troffaes, J. Gledhill, D. Škulj, S. Blake, Using imprecise continuous time Markov chains for assessing the reliability of power networks with common cause failure and non-immediate repair, in: Proceedings of ISIPTA'15, 2015, pp. 287–294.

[23] C. Moler, C. Van Loan, Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later, SIAM Review 45 (2003) 3–49. doi:10.1137/S00361445024180.

[24] A. Erreygers, J. De Bock, Computing inferences for large-scale continuous-time Markov chains by combining lumping with imprecision, 2018. Extended preprint of [17]. arXiv:1804.01020 [math.PR].

[25] A. Erreygers, J. De Bock, Imprecise continuous-time Markov chains: Efficient computational methods with guaranteed error bounds, in: Proceedings of ISIPTA'17, PMLR, 2017, pp. 145–156. Extended pre-print: arXiv:1702.07150 [math.PR].

[26] W. J. Stewart, Probability, Markov chains, queues, and simulation: the mathematical basis of performance modeling, Princeton University Press, 2009.

[27] A. Tornambè, Discrete-event system theory: An introduction, World Scientific, 1995.

[28] F. Hermans, G. de Cooman, Characterisation of ergodic upper transition operators, International Journal of Approximate Reasoning 53 (2012) 573–583. doi:10.1016/j.ijar.2011.12.008.

[29] D. Škulj, R. Hable, Coefficients of ergodicity for Markov chains with uncertain parameters, Metrika 76 (2013) 107–133. doi:10.1007/s00184-011-0378-0.

Appendix A. Extra material for Section 2

Recall from Section 2.3 that we can consider stochastic processes in two frameworks: the classical framework of measure-theoretic probability and the slightly less standard framework of coherent conditional probability. For the former, we refer to [20, Section 1.1] and references therein. Since the latter is the approach that is introduced and followed by Krak et al. [12], it will also be the approach that we will follow here. We therefore briefly recall the notation, terminology and results from [12, Sections 4 and 5] that we will need in the remainder: we discuss coherent conditional probabilities in Appendix A.1, explain how stochastic processes are coherent conditional probabilities with a specific domain in Appendix A.2 and treat the special case of (homogeneous) Markov chains in Appendix A.3.

Appendix A.1. Coherent conditional probabilities

Fix some non-empty set S called the *outcome space*. For this outcome space S, we let $\mathscr{E}(S)$ denote the set of all events—that is, the set of all subsets of S—and furthermore let $\mathscr{E}_\emptyset(S) := \mathscr{E}(S) \setminus \{\emptyset\}$. Coherent conditional probabilities are maps from a subset $\mathscr{C}$ of $\mathscr{E}(S) \times \mathscr{E}_\emptyset(S)$ to the real numbers that, in essence, satisfy the laws of probability. More specifically, Regazzini [18] put forward the following definition.

Definition 4 (Definition 1 in [18] or Definition 4.2 in [12])**.** Let S be a non-empty set and P a real-valued map from $\mathscr{C} \subseteq \mathscr{E}(S) \times \mathscr{E}_\emptyset(S)$ to $\mathbb{R}$. Then P is a *coherent conditional probability* if, for all n in $\mathbb{N}$, (A_1, C_1), ..., (A_n, C_n) in $\mathscr{C}$ and $\lambda_1, \dots, \lambda_n$ in $\mathbb{R}$,

$$\max\left\{\sum_{i=1}^{n} \lambda_i \mathbb{I}_{C_i}(s)(P(A_i \,|\, C_i) - \mathbb{I}_{A_i}(s))\colon s \in \cup_{i=1}^{n} C_i\right\} \geq 0. \tag{A.1}$$

Lemma 18 ((5)–(8) in [18])**.** *Let S be a non-empty set. If P is a coherent conditional probability on $\mathscr{C} \subseteq \mathscr{E}(S) \times \mathscr{E}_\emptyset(S)$, then,*

P1. *$P(A \,|\, C) \geq 0$ for all (A, C) in $\mathscr{C}$;*

P2. *$P(A \,|\, C) = 1$ for all (A, C) in $\mathscr{C}$ with $C \subseteq A$;*

P3. *$P(A \cup B \,|\, C) = P(A \,|\, C) + P(B \,|\, C)$ for all (A, C), (B, C) and $(A \cup B, C)$ in $\mathscr{C}$ such that $A \cap B = \emptyset$;*

P4. *$P(A \cap B \,|\, C) = P(A \,|\, B \cap C)P(B \,|\, C)$ for all $(A \cap B, C)$, $(A, B \cap C)$ and (B, C) in $\mathscr{C}$.*

Lemma 18 states that a coherent conditional probability satisfies the standard laws of (conditional) probability on its domain: properties (P1)–(P3) state that $P(\cdot \,|\, C)$ is a (finitely-additive) probability measure, while (P4) is Bayes' rule. Important to mention here is that the conditions (P1)–(P4) are, at least in general, *not* sufficient for P to be a coherent conditional probability! However, as is established in the following result, the conditions (P1)–(P4) are necessary and sufficient for P to be a coherent conditional probability if the domain $\mathscr{C}$ has some special structure.

Proposition 19 (Theorem 3 in [18])**.** *Let S be a non-empty set and P a real-valued map from $\mathscr{C} \subseteq \mathscr{E}(S) \times \mathscr{E}_\emptyset(S)$ to $\mathbb{R}$. If there are algebras $\mathscr{A} \subseteq \mathscr{E}(S)$ and $\mathscr{H} \subseteq \mathscr{E}(S)$ such that $\mathscr{H} \subseteq \mathscr{A}$ and $\mathscr{C} = \mathscr{A} \times (\mathscr{H} \setminus \{\emptyset\})$, then P is a coherent conditional probability if and only if it satisfies* (P1)–(P4).

An additional reason for using the more abstract condition of Definition 4 are the two following results, which are essential to our formal definition of the lumped stochastic process in Appendix B further on.

Lemma 20 (Theorem 4 in [18])**.** *Let S be a non-empty set. If P is a coherent conditional probability on $\mathscr{C} \subseteq \mathscr{E}(S) \times \mathscr{E}_\emptyset(S)$, then for any $\mathscr{C}^\star$ such that $\mathscr{C} \subseteq \mathscr{C}^\star \subseteq \mathscr{E}(S) \times \mathscr{E}_\emptyset(S)$, P can be extended to a coherent conditional probability $P^\star$ on $\mathscr{C}^\star$, in the sense that $P^\star(A \,|\, C) = P(A \,|\, C)$ for all $(A, C) \in \mathscr{C}$.*

Lemma 21 (Corollary 4.3 in [12])**.** *Let S be a non-empty set. Then P is a coherent conditional probability on $\mathscr{C} \subseteq \mathscr{E}(S) \times \mathscr{E}_\emptyset(S)$ if and only if it can be extended to a coherent conditional probability on $\mathscr{E}(S) \times \mathscr{E}_\emptyset(S)$.*

Appendix A.2. Stochastic processes

We here briefly introduce the coherent conditional framework for stochastic processes; we refer to [12, Section 4.2] for a more extensive introduction. The outcome space is now the set of paths Ω, where a path ω basically is the state of the system over time, so a map from the non-negative real numbers to the (non-empty and finite) state space $\mathscr{X}$. In the current setting, the only thing that is required of this set Ω is that

$$(\forall u \in \mathscr{U}_\emptyset)(\forall x_u \in \mathscr{X}_u)(\exists \omega \in \Omega)(\forall t \in u)\ \omega(t) = x_t. \tag{A.2}$$

For all t in $\mathbb{R}_{\geq 0}$ and x in $\mathscr{X}$, we then define the elementary event

$$(X_t = x) := \{\omega \in \Omega \colon \omega(t) = x\}.$$

Similarly, for all u in $\mathscr{U}$ and x_u in $\mathscr{X}_u$, we let

$$(X_u = x_u) \coloneqq \bigcap_{t \in u} (X_t = x_t).$$

We follow the convention that an empty intersection in expressions similar to the one above corresponds to Ω; hence $(X_\emptyset = x_\emptyset) = \Omega$. For all u in $\mathscr{U}$, the set of elementary events

$$\mathscr{E}_u \coloneqq \begin{cases} \{(X_t = x) \colon t \in \mathbb{R}_{\geq 0}, x \in \mathscr{X}\} & \text{if } u = \emptyset, \\ \{(X_t = x) \colon t \in u \cup [\max u, +\infty), x \in \mathscr{X}\} & \text{otherwise,} \end{cases} \tag{A.3}$$

induces an algebra of sets $\mathscr{A}_u \coloneqq \langle \mathscr{E}_u \rangle$.

Definition 5 (Definition 4.3 in [12])**.** A *stochastic process* P is a coherent conditional probability P with domain

$$\mathscr{C}^{\mathrm{SP}} \coloneqq \{(A_u, X_u = x_u) \colon u \in \mathscr{U}, x_u \in \mathscr{X}_u, A_u \in \mathscr{A}_u\}.$$

In order not to unnecessarily clutter our notation, we leave out the conditioning event if it is $(X_\emptyset = x_\emptyset) = \Omega$:

$$P(A) \coloneqq P(A \mid X_\emptyset = x_\emptyset) = P(A \mid \Omega) \quad \text{for any } A \text{ in } \mathscr{A}_\emptyset.$$

It immediately follows from Lemma 18 that a stochastic process P satisfies the laws of (conditional) probability. Because these laws are so well-known, we will frequently use them without explicitly referring to Lemma 18.

Appendix A.3. Precise (homogeneous) continuous-time Markov chains

The following is a more formal definition of the terms introduced in Section 2.3.

Definition 6. A stochastic process $P \colon \mathscr{C}^{\mathrm{SP}} \to \mathbb{R}$ is a *continuous-time Markov chain* if, for all t, Δ in $\mathbb{R}_{\geq 0}$, u in $\mathscr{U}_{<t}$, x, y in $\mathscr{X}$ and x_u in $\mathscr{X}_u$,

$$P(X_{t+\Delta} = y \mid X_u = x_u, X_t = x) = P(X_{t+\Delta} = y \mid X_t = x).$$

This Markov chain P is *homogeneous* if furthermore

$$P(X_{t+\Delta} = y \mid X_t = x) = P(X_\Delta = y \mid X_0 = x).$$

To ensure that the process behaves sufficiently nice, Krak et al. [12] always assume that the Markov chain P is *well-behaved* [12, Definition 4.4]; throughout the remainder, we implicitly assume that the Markov chains we consider are well-behaved. Our statement in Section 2.3 that a (well-behaved) homogeneous Markov chain is uniquely characterised by the triplet $(\mathscr{X}, \pi_0, Q)$ is justified by [12, Corollary 5.3 and Theorem 5.4].

Let P be a homogeneous Markov chain. To compute expectations that depend on the state at multiple future time points, we make use of Eqn. (6) and the *law of iterated expectation*, which states that for all u in $\mathscr{U}$, v and w in $\mathscr{U}_\emptyset$ with $\max u < \min v$ and $\max v < \min w$, x_u in $\mathscr{X}_u$ and f in $\mathscr{L}(\mathscr{X}_{u \cup v \cup w})$,

$$E(f(X_u, X_v, X_w) \mid X_u = x_u) = E(E(f(X_u, X_v, X_w) \mid X_u, X_v) \mid X_u = x_u). \tag{A.4}$$

In this expression, we use the notational convention that $E(f(X_u, X_v, X_w) \mid X_u, X_v)$ is the real-valued function on $\mathscr{X}_{u \cup v}$ that maps any (x_u, x_v) in $\mathscr{X}_{u \cup v}$ to $E(f(X_u, X_v, X_w) \mid X_u = x_u, X_v = x_v)$.

We now follow the exposition in [12, Section 9]. First, we consider a non-empty sequence of time points $u = t_1, \ldots, t_n$ in $\mathscr{U}_\emptyset$ and a single future time point s in $\mathbb{R}_{\geq 0}$ such that $s > \max u = t_n$. Fix some real-valued function f in $\mathscr{L}(\mathscr{X}_{u \cup s})$ that depends on the state at these time points. It is then well-known that, for any x_u in $\mathscr{X}_u$,

$$E(f(X_u, X_s) \mid X_u = x_u) = E(f(x_u, X_s) \mid X_u = x_u) = E(f_{x_u}(X_s) \mid X_u = x_u), \tag{A.5}$$

where we let $f_{x_u}\colon \mathscr{X} \to \mathbb{R}\colon x \mapsto f_{x_u}(x) \coloneqq f(x_u, x)$. It now follows from Eqns. (6) and (A.5) that

$$E(f(X_u, X_s) \mid X_u = x_u) = [T_{t_n}^{s} f_{x_u}](x_{t_n}), \tag{A.6}$$

where $[T_{t_n}^{s} f_{x_u}](x_{t_n})$ depends on the entire state history x_u, and not just on x_{t_n}. Inspired by this equation, for any s in $\mathbb{R}_{\geq 0}$, $u = t_1, \ldots, t_n$ in $\mathscr{U}_\emptyset$ such that $t_n < s$ and f in $\mathscr{L}(\mathscr{X}_{u\cup s})$, we let $T_{t_n}^{s} f$ be the real-valued function on $\mathscr{X}_u$ defined by

$$[T_{t_n}^{s} f](x_u) \coloneqq [T_{t_n}^{s} f_{x_u}](x_{t_n}) = [T_{t_n}^{s} f(x_u, X_s)](x_{t_n}) \qquad \text{for all } x_u \in \mathscr{X}_u. \tag{A.7}$$

We are now ready to move on to functions that depend on multiple future time points. To that end, we fix some $u = t_1, \ldots, t_n$ and $v = s_1, \ldots, s_m$ in $\mathscr{U}_\emptyset$ with $t_n = \max u < \min v = s_1$, and some f in $\mathscr{L}(\mathscr{X}_{u\cup v})$. With some tedious but straightforward work—essentially repeatedly applying the law of iterated expectation and Eqn. (A.6), see for instance [12, Section 9.2]—we obtain that

$$E(f(X_u, X_v) \mid X_u = x_u) = [T_{t_n}^{s_1} T_{s_1}^{s_2} \cdots T_{s_{m-1}}^{s_m} f](x_u) \qquad \text{for all } x_u \in \mathscr{X}_u, \tag{A.8}$$

where we use the notational convention defined in Eqn. (A.7).

Finally, we are ready to consider marginal expectations. From the law of iterated expectation and Eqn. (A.8), it now follows that for all $u = t_0, \ldots, t_n$ in $\mathscr{U}_\emptyset$ such that $t_0 = 0$ and all f in $\mathscr{X}_u$,

$$E(f(X_u)) = E(f(X_u) \mid X_\emptyset = x_\emptyset) = E(E(f(X_u) \mid X_0 = x_0)) = E_{\pi_0}(T_{t_0}^{t_1} T_{t_1}^{t_2} \cdots T_{t_{n-1}}^{t_n} f), \tag{A.9}$$

where E_{π_0} is the expectation operator defined by $E_{\pi_0}(g) \coloneqq \langle \pi_0, g\rangle$ for all g in $\mathscr{L}(\mathscr{X})$. The requirement $t_0 = 0$ is a purely formal one and does not impose any restrictions. Indeed, if $\min u \neq 0$, then—as argued in [12, right after Proposition 9.5]—one can simply consider the extension $f^\star$ of f to $\mathscr{X}_{\{0\}\cup u}$, defined by $f^\star(x, x_u) \coloneqq f(x_u)$ for all (x, x_u) in $\mathscr{X}_{\{0\}\cup u}$.

Appendix B. Extra material for and proofs of the results in Section 3

In order to define the lumped process rigorously, we need a more formal construction than that given in Section 3. To that end, we now consider an arbitrary Markov chain P and an arbitrary lumping map $\Lambda\colon \mathscr{X} \to \hat{\mathscr{X}}$.

For starters, we observe that the lumping map naturally induces a set of lumped paths $\hat{\Omega}$:

$$\hat{\Omega} \coloneqq \{\Lambda \circ \omega \colon \omega \in \Omega\}. \tag{B.1}$$

Note that because Ω satisfies Eqn. (A.2), $\hat{\Omega}$ clearly satisfies a lumped version of Eqn. (A.2):

$$(\forall u \in \mathscr{U}_\emptyset)(\forall \hat{x}_u \in \hat{\mathscr{X}}_u)(\exists \hat{\omega} \in \hat{\Omega})(\forall t \in u)\ \hat{\omega}(t) = \hat{x}_t.$$

The elementary events are now of the form

$$(\hat{X}_t = \hat{x}) \coloneqq \{\hat{\omega} \in \hat{\Omega} \colon \hat{\omega}(t) = \hat{x}\},$$

with t in $\mathbb{R}_{\geq 0}$ and $\hat{x}$ in $\hat{\mathscr{X}}$. As before, for any u in $\mathscr{U}$ and $\hat{x}_u$ in $\hat{\mathscr{X}}_u$, we also let

$$(\hat{X}_u = \hat{x}_u) \coloneqq \bigcap_{t\in u} (\hat{X}_t = \hat{x}_t),$$

where $(\hat{X}_\emptyset = \hat{x}_\emptyset) = \hat{\Omega}$. For any u in $\mathscr{U}$, the set of elementary elements

$$\hat{\mathscr{E}}_u \coloneqq \begin{cases} \{(\hat{X}_t = \hat{x}) \colon t \in \mathbb{R}_{\geq 0}, \hat{x} \in \hat{\mathscr{X}}\} & \text{if } u = \emptyset, \\ \{(\hat{X}_t = \hat{x}) \colon t \in u \cup [\max u, +\infty), \hat{x} \in \hat{\mathscr{X}}\} & \text{otherwise,} \end{cases}$$

induces the algebra of sets $\hat{\mathscr{A}}_u \coloneqq \langle \hat{\mathscr{E}}_u \rangle$. The domain of the lumped stochastic process $\hat{P}$ should hence be

$$\hat{\mathscr{C}}^{\mathrm{SP}} \coloneqq \{(\hat{A}_u, \hat{X}_u = \hat{x}_u) \colon u \in \mathscr{U}, \hat{x}_u \in \hat{\mathscr{X}}_u, \hat{A}_u \in \hat{\mathscr{A}}_u\}.$$

We have now introduced almost all concepts needed to formally define the lumped stochastic process $\hat{P}$. The sole remaining concept that we need is another inverse derived from Λ, this time from $\hat{\Omega}$ to Ω. To that end, we consider the map $\Lambda_\Omega^{-1} \colon \mathscr{E}(\hat{\Omega}) \to \mathscr{E}(\Omega)$ that maps any subset $\hat{A}$ of $\hat{\Omega}$ to

$$\Lambda_\Omega^{-1}(\hat{A}) \coloneqq \{\omega \in \Omega \colon \Lambda \circ \omega \in \hat{A}\}, \tag{B.2}$$

which is a subset of Ω. Note that Λ_Ω^{-1} is indeed an inverse, as clearly

$$\{\Lambda \circ \omega \colon \omega \in \Lambda_\Omega^{-1}(\hat{A})\} = \hat{A}. \tag{B.3}$$

The following result establishes that Λ_Ω^{-1} maps events in the algebra $\hat{\mathscr{A}}_u$ to events in the algebra $\mathscr{A}_u$.

Lemma 22. *Consider a lumping map $\Lambda \colon \mathscr{X} \to \hat{\mathscr{X}}$. Then for all u in $\mathscr{U}$ and $\hat{x}_u$ in $\hat{\mathscr{X}}_u$,*

$$\Lambda_\Omega^{-1}(\hat{X}_u = \hat{x}_u) = \bigcup_{x_u \in \hat{x}_u} (X_u = x_u), \tag{B.4}$$

More generally, for all u in $\mathscr{U}$ and $\hat{A}_u$ in $\hat{\mathscr{A}}_u$, $\Lambda_\Omega^{-1}(\hat{A}_u)$ belongs to $\mathscr{A}_u$.

Proof. We start with proving the first part of the statement. To that end, we distinguish two cases: $u = \emptyset$ and $u \neq \emptyset$. If $u = \emptyset$, then $(\hat{X}_u = \hat{x}_u) = \hat{\Omega}$. From this and the definitions of $\hat{\Omega}$ and Λ_Ω^{-1}, it then follows immediately that

$$\Lambda_\Omega^{-1}(\hat{X}_u = \hat{x}_u) = \Lambda_\Omega^{-1}(\hat{\Omega}) = \Omega = (X_u = x_u),$$

which agrees with the stated.

Next, we assume that $u \neq \emptyset$. Then

$$\begin{aligned}\Lambda_\Omega^{-1}(\hat{X}_u = \hat{x}_u) &= \{\omega \in \Omega \colon \Lambda \circ \omega \in (\hat{X}_u = \hat{x}_u)\} = \{\omega \in \Omega \colon (\forall t \in u)\ [\Lambda \circ \omega](t) = \hat{x}_t\} \\ &= \bigcap_{t \in u} \{\omega \in \Omega \colon [\Lambda \circ \omega](t) = \hat{x}_t\} = \bigcap_{t \in u} \bigcup_{x_t \in \hat{x}_t} \{\omega \in \Omega \colon \omega(t) = x_t\} \\ &= \bigcap_{t \in u} \bigcup_{x_t \in \hat{x}_t} (X_t = x_t) = \bigcup_{x_{t_1} \in \hat{x}_{t_1}} \cdots \bigcup_{x_{t_n} \in \hat{x}_{t_n}} (X_{t_1} = x_{t_1}) \cap \cdots \cap (X_{t_n} = x_{t_n}) \\ &= \bigcup_{x_u \in \hat{x}_u} \bigcap_{t \in u} (X_t = x_t) = \bigcup_{x_u \in \hat{x}_u} (X_u = x_u),\end{aligned}$$

where we let $u = t_1, \ldots, t_n$.

We now move on to the second part of the stated. If $\hat{A}_u = \emptyset$, we infer from Eqn. (B.2) that $\Lambda_\Omega^{-1}(\hat{A}) = \emptyset$, so $\Lambda_\Omega^{-1}(\hat{A}_u)$ belongs to $\mathscr{A}_u$. Fix some u in $\mathscr{U}$ and some $\hat{A}_u$ in $\hat{\mathscr{A}}_u$. It remains to consider the case $\hat{A}_u \neq \emptyset$. Because $\hat{\mathscr{A}}_u$ is the algebra generated by the elementary events in $\hat{\mathscr{E}}_u$ there is—see for instance also [12, Proof of Lemma C.3]—some time sequence v in $\mathscr{U}$ with $\max u < \min v$ and a non-empty set of tuples $\hat{S} \subseteq \hat{\mathscr{X}}_{u \cup v}$ such that

$$\hat{A}_u = \bigcup_{\hat{z}_w \in \hat{S}} (\hat{X}_w = \hat{z}_w),$$

where we let $w \coloneqq u \cup v$. By Eqn. (B.2),

$$\Lambda_\Omega^{-1}(\hat{A}_u) = \left\{\omega \in \Omega \colon \Lambda \circ \omega \in \bigcup_{\hat{z}_w \in \hat{S}} (\hat{X}_w = \hat{z}_w)\right\} = \bigcup_{\hat{z}_w \in \hat{S}} \{\omega \in \Omega \colon \Lambda \circ \omega \in (\hat{X}_w = \hat{z}_w)\}.$$

Using the definition of $(\hat{X}_w = \hat{z}_w)$ and Eqn. (B.1), we write this as

$$\Lambda_\Omega^{-1}(\hat{A}_u) = \bigcup_{\hat{z}_w \in \hat{S}} \{\omega \in \Omega \colon (\forall t \in w)\ \Lambda(\omega(t)) = \hat{z}_t\} = \bigcup_{\hat{z}_w \in \hat{S}} \left(\bigcap_{t \in w} \{\omega \in \Omega \colon \Lambda(\omega(t)) = \hat{z}_t\} \right)$$
$$= \bigcup_{\hat{z}_w \in \hat{S}} \left(\bigcap_{t \in w} \left[\bigcup_{z_t \in \hat{z}_t} (X_t = z_t) \right] \right).$$

It is now immediately clear that $\Lambda_\Omega^{-1}(\hat{A}_u)$ is an element of $\mathscr{A}_u$. □

The inverse Λ_Ω^{-1} naturally suggests a sensible formal definition of *a* lumped stochastic process $\hat{P} \colon \hat{\mathscr{C}}^{\mathrm{SP}} \to \mathbb{R}$. Recall from Definition 6, Definition 5 and Lemma 20 that the Markov chain P can be extended to a coherent conditional probability $P^\star$ on $\mathscr{E}(\Omega) \times \mathscr{E}_\emptyset(\Omega)$. Then the *lumped stochastic process* $\hat{P} \colon \hat{\mathscr{C}}^{\mathrm{SP}} \to \mathbb{R}$ corresponding to this extension $P^\star$ is defined for all $(\hat{A}_u, \hat{X}_u = \hat{x}_u)$ in $\hat{\mathscr{C}}^{\mathrm{SP}}$ as

$$\hat{P}(\hat{A}_u \mid \hat{X}_u = \hat{x}_u) \coloneqq P^\star(\Lambda_\Omega^{-1}(\hat{A}_u) \mid \Lambda_\Omega^{-1}(\hat{X}_u = \hat{x}_u)). \tag{B.5}$$

This is a proper definition because $\Lambda_\Omega^{-1}(\hat{A}_u)$ belongs to $\mathscr{E}(\Omega)$ due to the definition of Λ_Ω^{-1} and $\Lambda_\Omega^{-1}(\hat{X}_u = \hat{x}_u)$ belongs to $\mathscr{E}_\emptyset(\Omega)$ due to Lemma 22 and Eqn. (A.2). Unfortunately, this definition is—at least in general—not unique, since the extension $P^\star$ of the Markov chain P need not be unique. However, it does yield a stochastic process.

Theorem 23. *Consider a Markov chain P and a lumping map $\Lambda \colon \mathscr{X} \to \hat{\mathscr{X}}$. Let $P^\star$ be a coherent extension of P to $\mathscr{E}(\Omega) \times \mathscr{E}_\emptyset(\mathscr{X})$. Then $\hat{P} \colon \hat{\mathscr{C}}^{\mathrm{SP}} \to \mathbb{R}$, as defined by Eqn.* (B.5)*, is a stochastic process.*

Proof. We first construct the real-valued map $\hat{P}^\star$ on $\mathscr{E}(\hat{\Omega}) \times \mathscr{E}_\emptyset(\hat{\Omega})$, defined by

$$\hat{P}^\star(\hat{A} \mid \hat{C}) \coloneqq P^\star(\Lambda_\Omega^{-1}(\hat{A}) \mid \Lambda_\Omega^{-1}(\hat{C})) \quad \text{for all } (\hat{A}, \hat{C}) \in \mathscr{E}(\hat{\Omega}) \times \mathscr{E}_\emptyset(\hat{\Omega}). \tag{B.6}$$

It is clear that $\hat{P}$ is the restriction of $\hat{P}^\star$ to $\hat{\mathscr{C}}^{\mathrm{SP}}$. Hence, it follows from Lemma 21 that $\hat{P}$ is a stochastic process if $\hat{P}^\star$ is a coherent conditional probability.

We now verify that $\hat{P}^\star$ is indeed a coherent conditional probability. To that end, we fix any n in $\mathbb{N}$, $(\hat{A}_1, \hat{C}_1)$, …, $(\hat{A}_n, \hat{C}_n)$ in $\mathscr{E}(\hat{\Omega}) \times \mathscr{E}_\emptyset(\hat{\Omega})$ and λ_1, …, λ_n in $\mathbb{R}$ and show that $\max S \geq 0$, where

$$S \coloneqq \left\{ \sum_{i=1}^n \lambda_i \mathbb{I}_{\hat{C}_i}(\hat{\omega}) \Big(\hat{P}^\star(\hat{A}_i \mid \hat{C}_i) - \mathbb{I}_{\hat{A}_i}(\hat{\omega}) \Big) \colon \hat{\omega} \in \bigcup_{i=1}^n \hat{C}_i \right\}.$$

Substituting Eqn. (B.6) yields

$$S = \left\{ \sum_{i=1}^n \lambda_i \mathbb{I}_{\hat{C}_i}(\hat{\omega}) \Big(P^\star(\Lambda_\Omega^{-1}(\hat{A}_i) \mid \Lambda_\Omega^{-1}(\hat{C}_i)) - \mathbb{I}_{\hat{A}_i}(\hat{\omega}) \Big) \colon \hat{\omega} \in \bigcup_{i=1}^n \hat{C}_i \right\}.$$

Furthermore, using Eqns. (B.2) and (B.3) yields

$$S = \left\{ \sum_{i=1}^n \lambda_i \mathbb{I}_{\hat{C}_i}(\Lambda \circ \omega) \Big(P^\star(\Lambda_\Omega^{-1}(\hat{A}_i) \mid \Lambda_\Omega^{-1}(\hat{C}_i)) - \mathbb{I}_{\hat{A}_i}(\Lambda \circ \omega) \Big) \colon \omega \in \bigcup_{i=1}^n \Lambda_\Omega^{-1}(\hat{C}_i) \right\}.$$

Observe that for all ω in Ω and $\hat{A} \subseteq \hat{\Omega}$,

$$\mathbb{I}_{\hat{A}}(\Lambda \circ \omega) = \begin{cases} 1 & \text{if } \Lambda \circ \omega \in \hat{A} \\ 0 & \text{otherwise} \end{cases} = \begin{cases} 1 & \text{if } \omega \in \Lambda_\Omega^{-1}(\hat{A}) \\ 0 & \text{otherwise} \end{cases} = \mathbb{I}_{\Lambda_\Omega^{-1}(\hat{A})}(\omega), \tag{B.7}$$

where the second equality follows from Eqn. (B.2). We substitute Eqn. (B.7) in our expression for S, to yield

$$S = \left\{ \sum_{i=1}^{n} \lambda_i \mathbb{I}_{C_i}(\omega)(P^\star(A_i \mid C_i) - \mathbb{I}_{A_i}(\omega)) \colon \omega \in \bigcup_{i=1}^{n} C_i \right\},$$

where, for all i in $\{1, \dots, n\}$, we let $A_i \coloneqq \Lambda_\Omega^{-1}(\hat{A}_i)$ and $C_i \coloneqq \Lambda_\Omega^{-1}(\hat{C}_i)$. Because $P^\star$ is a coherent conditional probability on $\mathscr{E}(\Omega) \times \mathscr{E}_\emptyset(\Omega)$, it follows from Definition 4 that $\max S \geq 0$. □

The previous theorem validates our use of the term "lumped stochastic process" for $\hat{P}$. However, at first sight, the sensibility of our definition might still not seem entirely obvious. That it nevertheless is, follows from the following two results. The first result provides a justification for Eqn. (10); the second one also makes clear that the lack of uniqueness in our definition is a consequence of conditioning on events with zero probability.

Corollary 24. *Consider a Markov chain P, a coherent extension $P^\star$ of P to $\mathscr{E}(\Omega) \times \mathscr{E}_\emptyset(\mathscr{X})$, a lumping map $\Lambda \colon \mathscr{X} \to \hat{\mathscr{X}}$ and the corresponding lumped stochastic process $\hat{P}$. For all u in $\mathscr{U}$ and $\hat{x}_u$ in $\hat{\mathscr{X}}_u$,*

$$\hat{P}(\hat{X}_u = \hat{x}_u) = P^\star(\Lambda_\Omega^{-1}(\hat{X}_u = \hat{x}_u)) = \sum_{x_u \in \hat{x}_u} P^\star(X_u = x_u) = \sum_{x_u \in \hat{x}_u} P(X_u = x_u) = P(X_u \in \hat{x}_u).$$

Proof. It follows immediately from Eqn. (B.5) and Lemma 22 that

$$\hat{P}(\hat{X}_u = \hat{x}_u) = P^\star(\Lambda_\Omega^{-1}(\hat{X}_u = \hat{x}_u)) = P^\star(\bigcup_{x_u \in \hat{x}_u} (X_u = x_u)) = \sum_{x_u \in \hat{x}_u} P^\star(X_u = x_u) = \sum_{x_u \in \hat{x}_u} P(X_u = x_u),$$

where the third equality follows from the finite additivity of $P^\star$ and the final equality holds because $P^\star$ is an extension of P and $(X_u = x_u, \Omega)$ belongs to the domain $\mathscr{C}^{\text{SP}}$ of P for all x_u in $\Lambda^{-1}(\hat{x}_u)$. Finally, it follows immediately from the finite additivity of P and Eqn. (9) that

$$\sum_{x_u \in \hat{x}_u} P(X_u = x_u) = P(X_u \in \hat{x}_u).$$

□

Proposition 25. *Consider a Markov chain P and a lumping map $\Lambda \colon \mathscr{X} \to \hat{\mathscr{X}}$, and let $\hat{P}$ be a lumped stochastic process. Then for any $(\hat{A}_u, \hat{X}_u = \hat{x}_u)$ in $\hat{\mathscr{C}}^{\text{SP}}$ with $\hat{P}(\hat{X}_u = \hat{x}_u) = \sum_{x_u \in \hat{x}_u} P(X_u = x_u) > 0$,*

$$\hat{P}(\hat{A}_u \mid \hat{X}_u = \hat{x}_u) = \frac{\sum_{x_u \in \hat{x}_u} P(\Lambda_\Omega^{-1}(\hat{A}_u) \mid X_u = x_u) P(X_u = x_u)}{\sum_{z_u \in \hat{x}_u} P(X_u = z_u)}.$$

Proof. Observe that by Eqn. (B.5),

$$\hat{P}(\hat{A}_u \mid \hat{X}_u = \hat{x}_u) = P^\star(\Lambda_\Omega^{-1}(\hat{A}_u) \mid \Lambda_\Omega^{-1}(\hat{X}_u = \hat{x}_u)),$$

where $P^\star$ is a coherent extension of P to $\mathscr{E}(\Omega) \times \mathscr{E}_\emptyset(\Omega)$. As $P^\star$ is a coherent conditional probability, it follows from Lemma 18 (P4) that

$$P^\star(\Lambda_\Omega^{-1}(\hat{A}_u) \mid \Lambda_\Omega^{-1}(\hat{X}_u = \hat{x}_u)) P^\star(\Lambda_\Omega^{-1}(\hat{X}_u = \hat{x}_u)) = P^\star(\Lambda_\Omega^{-1}(\hat{A}_u) \cap \Lambda_\Omega^{-1}(\hat{X}_u = \hat{x}_u)), \tag{B.8}$$

where $P^\star(\Lambda_\Omega^{-1}(\hat{X}_u = \hat{x}_u)) \coloneqq P^\star(\Lambda_\Omega^{-1}(\hat{X}_u = \hat{x}_u) \mid \Omega)$.

It follows immediately from Corollary 24 that

$$P^\star(\Lambda_\Omega^{-1}(\hat{X}_u = \hat{x}_u)) = \hat{P}(\hat{X}_u = \hat{x}_u) = \sum_{z_u \in \hat{x}_u} P(X_u = z_u). \tag{B.9}$$

Recall from Lemma 22 that $\Lambda_\Omega^{-1}(\hat{A}_u)$ is an element of $\mathscr{A}_u$. Since $\Lambda_\Omega^{-1}(\hat{X}_u = \hat{x}_u)$ is clearly also an element of $\mathscr{A}_u$, this implies that $\Lambda_\Omega^{-1}(\hat{A}_u) \cap \Lambda_\Omega^{-1}(\hat{X}_u = \hat{x}_u)$ is an element of $\mathscr{A}_u$ as well. Consequently, we find that

$$\begin{aligned} P^\star(\Lambda_\Omega^{-1}(\hat{A}_u) \cap \Lambda_\Omega^{-1}(\hat{X}_u = \hat{x}_u)) = P(\Lambda_\Omega^{-1}(\hat{A}_u) \cap \Lambda_\Omega^{-1}(\hat{X}_u = \hat{x}_u)) &= P(\Lambda_\Omega^{-1}(\hat{A}_u) \cap (\cup_{x_u \in \hat{x}_u}(X_u = x_u))) \\ &= \sum_{x_u \in \hat{x}_u} P(\Lambda_\Omega^{-1}(\hat{A}_u) \cap (X_u = x_u)) \\ &= \sum_{x_u \in \hat{x}_u} P(\Lambda_\Omega^{-1}(\hat{A}_u) \mid X_u = x_u) P(X_u = x_u). \end{aligned} \tag{B.10}$$

Since $\sum_{z_u \in \hat{x}_u} P(X_u = z_u) > 0$, substituting Eqn. (B.9) and (B.10) in Eqn. (B.8) yields

$$\hat{P}(\hat{A}_u \mid \hat{X}_u = \hat{x}_u) = P^\star(\Lambda_\Omega^{-1}(\hat{A}_u) \mid \Lambda_\Omega^{-1}(\hat{X}_u = \hat{x}_u)) = \frac{\sum_{x_u \in \hat{x}_u} P(\Lambda_\Omega^{-1}(\hat{A}_u) \mid X_u = x_u) P(X_u = x_u)}{\sum_{z_u \in \hat{x}_u} P(X_u = z_u)}. \qquad \square$$

Appendix C. Extra material regarding Section 4

In this section of the appendix, we provide some more background information on (imprecise) Markov chains. More specifically, we focus on four relevant types of transformations in Appendix C.1, and explain how to compute lower expectations using the law of iterated expectation in Appendix C.2.

Appendix C.1. Lower transition (rate) operators

A lower transition rate operator—see for instance [12, Definition 7.2]—is a transformation $\underline{R}\colon \mathscr{L}(\mathscr{X}) \to \mathscr{L}(\mathscr{X})$ such that

LTR1. $\underline{R}(f+g) \geq \underline{R}f + \underline{R}g$ for all f, g in $\mathscr{L}(\mathscr{X})$;

LTR2. $\underline{R}(\lambda f) = \lambda \underline{R} f$ for all f in $\mathscr{L}(\mathscr{X})$ and λ in $\mathbb{R}_{\geq 0}$;

LTR3. $[\underline{R}\mathbb{I}_y](x) \geq 0$ for all x, y in $\mathscr{X}$ with $x \neq y$;

LTR4. $\underline{R}\mu = 0$ for all μ in $\mathbb{R}$.

We have already seen in Section 4.2 that lower transition rate operators generate a lower transition rate operator, which can be thought of as the non-linear matrix exponential. The following lemma establishes a second connection between lower transition rate operators and lower transition operators.

Lemma 26 (Proposition 3 in [25])**.** *Consider a transition rate operator $\underline{Q}$ on $\mathscr{L}(\mathscr{X})$ and some Δ in $\mathbb{R}_{\geq 0}$. Then $(I + \Delta\underline{Q})$ is a lower transition operator if and only if $\Delta\|\underline{Q}\| \leq 2$.*

In the remainder, we will need the following interesting properties of lower transition operators. For their proofs, we refer to [28] and [12].

Lemma 27. *Let $\underline{T}$, $\underline{T}_1$ and $\underline{T}_2$ be lower transition operators on $\mathscr{L}(\mathscr{X})$. Then*

LT4. *the composition $\underline{T}_1\underline{T}_2$ is a lower transition operator;*

LT5. $\min f \leq \underline{T}f \leq \overline{T}f \leq \max f$ *for all f in $\mathscr{L}(\mathscr{X})$;*

LT6. $\underline{T}f \leq \underline{T}g$ *for all f, g in $\mathscr{L}(\mathscr{X})$ such that $f \leq g$.*

Because transition (rate) matrices are simply lower transition (rate) operators that are linear, Lemmas 26 and 27 specialise to transition (rate) matrices as follows.

Corollary 28. *Consider a transition rate matrix Q on $\mathscr{L}(\mathscr{X})$ and some Δ in $\mathbb{R}_{\geq 0}$. Then $(I + \Delta Q)$ is a transition matrix if and only if $\Delta\|Q\| \leq 2$.*

Corollary 29. *Let T, T_1 and T_2 be transition matrices on $\mathscr{L}(\mathscr{X})$. Then*

T4. *the composition T_1T_2 is a transition matrix;*

T5. $\min f \leq Tf \leq \max f$ *for all f in $\mathscr{L}(\mathscr{X})$;*

T6. $Tf \leq Tg$ *for all f, g in $\mathscr{L}(\mathscr{X})$ such that $f \leq g$.*

Appendix C.2. Computing lower (conditional) expectations

In order to compute lower (conditional) expectations of functions that depend on the state at a finite number of future time points, we follow exactly the same approach as the one for precise Markov chains that we have previously outlined in Appendix A.3; for a more detailed treatment, we refer to [12, Section 9]. We let $\mathscr{M}$ be a non-empty set of initial distributions and Q a non-empty bounded and convex set of transition rate matrices that has separately specified rows, and fix some $u = t_1, \ldots, t_n$ in $\mathscr{U}_\emptyset$. We first generalise the notational convention of Eqn. (A.7): for any s in $\mathbb{R}_{\geq 0}$ with $s > \max u = t_n$ and any f in $\mathscr{L}(\mathscr{X}_{u \cup s})$, we let $\underline{T}_{t_n}^{s} f$ be the real-valued function on $\mathscr{X}_u$ defined by

$$[\underline{T}_{t_n}^{s} f](x_u) \coloneqq [\underline{T}_{t_n}^{s} f_{x_u}](x_{t_n}) = [\underline{T}_{t_n}^{s} f(x_u, X_s)](x_{t_n}) \qquad \text{for all } x_u \in \mathscr{X}_u. \tag{C.1}$$

It now follows from [12, Corollary 9.2] that, for any $v = s_1, \ldots, s_m$ in $\mathscr{U}_\emptyset$ such that $\min v > \max u$ and any f in $\mathscr{L}(\mathscr{X}_{u \cup v})$,

$$\underline{E}^{\mathrm{W}}_{\mathscr{Q},\mathscr{M}}(f(X_u, X_v) \mid X_u = x_u) = [\underline{T}_{t_n}^{s_1} \underline{T}_{s_1}^{s_2} \cdots \underline{T}_{s_{m-1}}^{s_m} f](x_u) \qquad \text{for all } x_u \in \mathscr{X}_u. \tag{C.2}$$

Next, we move from conditional lower expectations to marginal ones. In [12, Proposition 9.5], Krak et al. establish that for all $u = t_0, \ldots, t_n$ in $\mathscr{U}_\emptyset$ such that $t_0 = 0$ and all f in $\mathscr{L}(\mathscr{X}_u)$,

$$\underline{E}^{\mathrm{W}}_{\mathscr{Q},\mathscr{M}}(f(X_u)) = \underline{E}^{\mathrm{W}}_{\mathscr{Q},\mathscr{M}}(f(X_u) \mid X_\emptyset = x_\emptyset) = \underline{E}_{\mathscr{M}}(\underline{T}_{t_0}^{t_1} \underline{T}_{t_1}^{t_2} \cdots \underline{T}_{t_{n-1}}^{t_n} f), \tag{C.3}$$

where $\underline{E}_{\mathscr{M}}$ is the lower expectation operator defined by

$$\underline{E}_{\mathscr{M}}(g) \coloneqq \inf\{\langle \pi, g \rangle \colon \pi \in \mathscr{M}\} \qquad \text{for all } g \in \mathscr{L}(\mathscr{X}).$$

Recall from Appendix A.3 that the requirement $t_0 = 0$ is a purely formal one: if $\min u \neq 0$, then one considers the extension $f^\star$ of f to $\mathscr{X}_{\{0\} \cup u}$.

Appendix D. Extra material for and proofs of the results in Section 5

Proposition 4. *Let Q be a transition rate matrix and $\Lambda \colon \mathscr{X} \to \hat{\mathscr{X}}$ a lumping map. Then the corresponding transformation $\underline{\hat{Q}}$ is a lower transition rate operator.*

Proof. We simply need to verify (LTR1)–(LTR4).

(LTR1). Fix some $\hat{x}$ in $\hat{\mathscr{X}}$, and observe that for all $\hat{f}$ and $\hat{g}$ in $\mathscr{L}(\hat{\mathscr{X}})$,

$$\begin{aligned}
[\underline{\hat{Q}}(\hat{f} + \hat{g})](\hat{x}) &= \min\left\{ \sum_{\hat{y} \in \hat{\mathscr{X}}} (\hat{f}(\hat{y}) + \hat{g}(\hat{y})) \sum_{y \in \hat{y}} Q(x, y) \colon x \in \hat{x} \right\} \\
&\geq \min\left\{ \sum_{\hat{y} \in \hat{\mathscr{X}}} \hat{f}(\hat{y}) \sum_{y \in \hat{y}} Q(x, y) \colon x \in \hat{x} \right\} + \min\left\{ \sum_{\hat{y} \in \hat{\mathscr{X}}} \hat{g}(\hat{y}) \sum_{y \in \hat{y}} Q(x, y) \colon x \in \hat{x} \right\} \\
&= [\underline{\hat{Q}}\hat{f}](\hat{x}) + [\underline{\hat{Q}}\hat{g}](\hat{x}).
\end{aligned}$$

(LTR2). Fix some $\hat{x}$ in $\hat{\mathscr{X}}$, and observe that for all $\hat{f}$ in $\mathscr{L}(\hat{\mathscr{X}})$ and λ in $\mathbb{R}_{\geq 0}$,

$$[\underline{\hat{Q}}(\lambda\hat{f})](\hat{x}) = \min\left\{ \sum_{\hat{y}\in\hat{\mathscr{X}}} \lambda\hat{f}(\hat{y}) \sum_{y\in\hat{y}} Q(x,y) \colon x \in \hat{x} \right\} = \lambda \min\left\{ \sum_{\hat{y}\in\hat{\mathscr{X}}} \hat{f}(\hat{y}) \sum_{y\in\hat{y}} Q(x,y) \colon x \in \hat{x} \right\} = \lambda[\underline{\hat{Q}}\hat{f}](\hat{x}).$$

(LTR3). Fix some $\hat{x}, \hat{y}$ in $\hat{\mathscr{X}}$ such that $\hat{x} \neq \hat{y}$, and observe that

$$[\underline{\hat{Q}}\mathbb{I}_{\hat{y}}](\hat{x}) = \min\left\{ \sum_{\hat{z}\in\hat{\mathscr{X}}} \mathbb{I}_{\hat{y}}(\hat{z}) \sum_{z\in\hat{z}} Q(x,z) \colon x \in \hat{x} \right\} = \min\left\{ \sum_{y\in\hat{y}} Q(x,y) \colon x \in \hat{x} \right\} \geq 0,$$

where the inequality holds because $Q(x,y) = [Q\mathbb{I}_y](x) \geq 0$ for all x in $\hat{x}$ and y in $\hat{y}$ as Q is a transition rate matrix.

(LTR4). Fix some $\hat{x}$ in $\hat{\mathscr{X}}$ and μ in $\mathbb{R}$, and observe that

$$[\underline{\hat{Q}}\mu](\hat{x}) = \min\left\{ \sum_{\hat{y}\in\hat{\mathscr{X}}} \mu \sum_{y\in\hat{y}} Q(x,y) \colon x \in \hat{x} \right\} = \min\left\{ \sum_{y\in\mathscr{X}} \mu Q(x,y) \colon x \in \hat{x} \right\} = 0,$$

where the final equality is immediate because the transition rate matrix Q has zero row-sums. □

Lemma 30. *Let Q be a transition rate matrix and $\Lambda\colon \mathscr{X} \to \hat{\mathscr{X}}$ a lumping map. Then for any $\hat{f}$ in $\mathscr{L}(\hat{\mathscr{X}})$,*

$$(\underline{\hat{Q}}\hat{f}) \circ \Lambda \leq Q(\hat{f} \circ \Lambda).$$

Proof. Fix an arbitrary x in $\mathscr{X}$. Then some straightforward manipulations yield

$$\begin{aligned}
[(\underline{\hat{Q}}\hat{f}) \circ \Lambda](x) = (\underline{\hat{Q}}\hat{f})(\Lambda(x)) &= \min\left\{ \sum_{\hat{y}\in\hat{\mathscr{X}}} \hat{f}(\hat{y}) \sum_{y\in\hat{y}} Q(x',y) \colon x' \in \Lambda(x) \right\} \\
&= \min\left\{ \sum_{\hat{y}\in\hat{\mathscr{X}}} \sum_{y\in\hat{y}} \hat{f}(\hat{y}) Q(x',y) \colon x' \in \Lambda(x) \right\} \\
&= \min\left\{ \sum_{y\in\mathscr{X}} [\hat{f} \circ \Lambda](y) Q(x',y) \colon x' \in \Lambda(x) \right\} \\
&= \min\Big\{ [Q(\hat{f} \circ \Lambda)](x') \colon x' \in \Lambda(x) \Big\} \\
&\leq [Q(\hat{f} \circ \Lambda)](x),
\end{aligned}$$

where the inequality follows from the fact that x is an element of $\Lambda^{-1}(\Lambda(x))$. □

Lemma 31. *Consider a Markov chain P and a lumping map $\Lambda\colon \mathscr{X} \to \hat{\mathscr{X}}$. Let $P^\star$ be a coherent extension of P to $\mathscr{E}(\Omega) \times \mathscr{E}_\emptyset(\mathscr{X})$ and $\hat{P}$ the corresponding lumped stochastic process. Fix some t in $\mathbb{R}_{\geq 0}$, u in $\mathscr{U}_{<t}$, v in $\mathscr{U}_\emptyset$ with $\min v > t$, $\hat{x}$ in $\hat{\mathscr{X}}$ and $\hat{x}_u$ in $\hat{\mathscr{X}}_u$. Then for any real-valued function $\hat{f}$ on $\hat{\mathscr{X}}_v$,*

$$\min\{E([\hat{f} \circ \Lambda](X_v) \mid X_t = x) \colon x \in \hat{x}\} \leq \hat{E}(\hat{f}(\hat{X}_v) \mid \hat{X}_u = \hat{x}_u, \hat{X}_t = \hat{x}),$$

where $\hat{E}$ denotes the expectation with respect to the lumped stochastic process $\hat{P}$.

Proof. By definition of the lumped stochastic process, it holds that

$$\hat{E}(\hat{f}(\hat{X}_v) \mid \hat{X}_u = \hat{x}_u, \hat{X}_t = \hat{x}) = E_{P^*}([\hat{f} \circ \Lambda](X_v) \mid C),$$

where we let $C \coloneqq \Lambda_\Omega^{-1}(\hat{X}_u = \hat{x}_u, \hat{X}_t = \hat{x})$ and where P^* is the coherent extension of P to $\mathscr{E}(\Omega) \times \mathscr{E}_\emptyset(\Omega)$ used to define the lumped stochastic process $\hat{P}$. Recall from Lemma 22—more specifically, from Eqn. (B.4)—that

$$C = \Lambda_\Omega^{-1}(\hat{X}_u = \hat{x}_u, \hat{X}_t = \hat{x}) = \bigcup_{x_u \in \hat{x}_u} \bigcup_{x \in \hat{x}} (X_u = x_u, X_t = x).$$

We now use this in combination with the law of total probability, to yield

$$E_{P^*}([\hat{f} \circ \Lambda](X_v) \mid C) = \sum_{x_u \in \hat{x}_u} \sum_{x \in \hat{x}} E_{P^*}([\hat{f} \circ \Lambda](X_v) \mid X_u = x_u, X_t = x) P^\star(X_u = x_u, X_t = x \mid C).$$

Observe that the right-hand side of this equality is a convex combination of terms of the form

$$E_{P^*}([\hat{f} \circ \Lambda](X_v) \mid X_u = x_u, X_t = x) = E([\hat{f} \circ \Lambda](X_v) \mid X_u = x_u, X_t = x) = E([\hat{f} \circ \Lambda](X_v) \mid X_t = x),$$

where the first equality holds because $P^\star$ is an extension of P and where the second equality follows from the Markov property. The stated now follows because a convex combination of terms is always bounded below by the minimum of these terms. □

In the remainder, we will need the following corollary.

Corollary 32. *Consider a Markov chain P, a lumping map $\Lambda\colon \mathscr{X} \to \hat{\mathscr{X}}$ and a corresponding lumped stochastic process $\hat{P}$. Fix some t, Δ in $\mathbb{R}_{\geq 0}$, u in $\mathscr{U}_{<t}$, $\hat{x}$ and $\hat{y}$ in $\hat{\mathscr{X}}$ and $\hat{x}_u$ in $\hat{\mathscr{X}}_u$. Then for any real-valued function $\hat{f}$ on $\hat{\mathscr{X}}$,*

$$\min\{[T_t^{t+\Delta}(\hat{f} \circ \Lambda)](x) \colon x \in \hat{x}\} \leq \hat{E}(\hat{f}(\hat{X}_{t+\Delta}) \mid \hat{X}_u = \hat{x}_u, \hat{X}_t = \hat{x}),$$

where $\hat{E}$ denotes the expectation with respect to the lumped stochastic process $\hat{P}$.

Proof. This is an immediate consequence of Lemma 31 and Eqn. (6). □

A similar corollary of Lemma 31 provides a justification for Eqn. (12) in the main text.

Corollary 33. *Consider a homogeneous Markov chain P, a lumping map $\Lambda\colon \mathscr{X} \to \hat{\mathscr{X}}$ and a corresponding lumped stochastic process $\hat{P}$. Fix some t, Δ in $\mathbb{R}_{\geq 0}$, u in $\mathscr{U}_{<t}$, $\hat{x}$ and $\hat{y}$ in $\hat{\mathscr{X}}$ and $\hat{x}_u$ in $\hat{\mathscr{X}}_u$. Then for any real-valued function $\hat{f}$ on $\hat{\mathscr{X}}$,*

$$\begin{aligned}\min\{E([\hat{f} \circ \Lambda](X_\Delta) \mid X_0 = x) \colon x \in \hat{x}\} &\leq \hat{E}(\hat{f}(\hat{X}_{t+\Delta}) \mid \hat{X}_u = \hat{x}_u, \hat{X}_t = \hat{x}) \\ &\leq \max\{E([\hat{f} \circ \Lambda](X_\Delta) \mid X_0 = x) \colon x \in \hat{x}\},\end{aligned}$$

where $\hat{E}$ denotes the expectation with respect to the lumped stochastic process $\hat{P}$.

Proof. The left inequality is an immediate consequence of Lemma 31 and the homogeneity of P; the right inequality follows almost immediately from applying the the left one to $-\hat{f}$. □

The following technical result allows us to use the results from [12].

Lemma 34. *Consider a well-behaved Markov chain P and a lumping map $\Lambda\colon \mathscr{X} \to \hat{\mathscr{X}}$, and let $\hat{P}$ be a lumped process. Then $\hat{P}$ is* well-behaved *[12, Definition 4.4], in the sense that, for all t in $\mathbb{R}_{\geq 0}$, u in $\mathscr{U}_{<t}$, $\hat{x}, \hat{y}$ in $\hat{\mathscr{X}}$ and $\hat{x}_u$ in $\hat{\mathscr{X}}_u$,*

$$\limsup_{\Delta \to 0^+} \frac{1}{\Delta} \left| \hat{P}(\hat{X}_{t+\Delta} = \hat{y} \mid \hat{X}_u = \hat{x}_u, \hat{X}_t = \hat{x}) - \mathbb{I}_{\hat{x}}(\hat{y}) \right| < +\infty$$

and, if $t \neq 0$,

$$\limsup_{\Delta \to 0^+} \frac{1}{\Delta} \left| \hat{P}(\hat{X}_t = \hat{y} \mid \hat{X}_u = \hat{x}_u, \hat{X}_{t-\Delta} = \hat{x}) - \mathbb{I}_{\hat{x}}(\hat{y}) \right| < +\infty.$$

Proof. We only prove the first inequality, the proof of the second inequality is entirely similar. To that end, we fix any arbitrary $\Delta \in \mathbb{R}_{>0}$ and observe that

$$\begin{aligned}\hat{P}(\hat{X}_{t+\Delta} = \hat{y} \mid \hat{X}_u = \hat{x}_u, \hat{X}_t = \hat{x}) &= \hat{E}(\mathbb{I}_{\hat{y}}(\hat{X}_{t+\Delta}) \mid \hat{X}_u = \hat{x}_u, \hat{X}_t = \hat{x}) \\ &\geq \min\{E([\mathbb{I}_{\hat{y}} \circ \Lambda](X_{t+\Delta}) \mid X_t = x) \colon x \in \hat{x}\} \\ &= \min\big\{E(\mathbb{I}_{\Lambda^{-1}(\hat{y})}(X_{t+\Delta}) \mid X_t = x) \colon x \in \hat{x}\big\} \\ &= \min\Bigg\{\sum_{y \in \hat{y}} P(X_{t+\Delta} = y \mid X_t = x) \colon x \in \hat{x}\Bigg\},\end{aligned}$$

where the inequality follows from Lemma 31 with $v = t + \Delta$ and $\hat{f} = \mathbb{I}_{\hat{y}}$. Similarly, it follows from Lemma 31 with $\hat{f} = -\mathbb{I}_{\hat{y}}$ that

$$\hat{P}(\hat{X}_{t+\Delta} = \hat{y} \mid \hat{X}_u = \hat{x}_u, \hat{X}_t = \hat{x}) \leq \max\Bigg\{\sum_{y \in \hat{y}} P(X_{t+\Delta} = y \mid X_t = x) \colon x \in \hat{x}\Bigg\}.$$

We combine these two inequalities, to yield

$$\begin{aligned}\min\Bigg\{\sum_{y \in \hat{y}} P(X_{t+\Delta} = y \mid X_t = x) \colon x \in \hat{x}\Bigg\} - \mathbb{I}_{\hat{x}}(\hat{y}) &\leq \hat{P}(\hat{X}_{t+\Delta} = \hat{y} \mid \hat{X}_u = \hat{x}_u, \hat{X}_t = \hat{x}) - \mathbb{I}_{\hat{x}}(\hat{y}) \\ &\leq \max\Bigg\{\sum_{y \in \hat{y}} P(X_{t+\Delta} = y \mid X_t = x) \colon x \in \hat{x}\Bigg\} - \mathbb{I}_{\hat{x}}(\hat{y}),\end{aligned}$$

where we have subtracted $\mathbb{I}_{\hat{x}}(\hat{y})$ on both sides of the inequalities. To continue, we move $\mathbb{I}_{\hat{x}}(\hat{y})$ inside the optimisations and use that $\mathbb{I}_{\hat{x}}(\hat{y}) = \sum_{y \in \hat{y}} \mathbb{I}_x$ for all x in $\hat{x}$, to yield

$$\begin{aligned}\min\Bigg\{\sum_{y \in \hat{y}} P(X_{t+\Delta} = y \mid X_t = x) - \mathbb{I}_x(y) \colon x \in \hat{x}\Bigg\} &\leq \hat{P}(\hat{X}_{t+\Delta} = \hat{y} \mid \hat{X}_u = \hat{x}_u, \hat{X}_t = \hat{x}) - \mathbb{I}_{\hat{x}}(\hat{y}) \\ &\leq \max\Bigg\{\sum_{y \in \hat{y}} P(X_{t+\Delta} = y \mid X_t = x) - \mathbb{I}_x(y) \colon x \in \hat{x}\Bigg\}.\end{aligned}$$

Next, we take the absolute value and use that both the absolute value of a minimum and the absolute value of a maximum are lower than the maximum of the absolute values, to yield

$$|\hat{P}(\hat{X}_{t+\Delta} = \hat{y} \mid \hat{X}_u = \hat{x}, \hat{X}_t = \hat{x}) - \mathbb{I}_{\hat{x}}(\hat{y})| \leq \max\Bigg\{\Bigg|\sum_{y \in \hat{y}} P(X_{t+\Delta} = y \mid X_t = x) - \mathbb{I}_x(y)\Bigg| \colon x \in \hat{x}\Bigg\}.$$

From this, it follows that

$$\begin{aligned}\frac{1}{\Delta}|\hat{P}(\hat{X}_{t+\Delta} = \hat{y} \mid \hat{X}_u = \hat{x}, \hat{X}_t = \hat{x}) - \mathbb{I}_{\hat{x}}(\hat{y})| &\leq \max\Bigg\{\frac{1}{\Delta}\Bigg|\sum_{y \in \hat{y}} P(X_{t+\Delta} = y \mid X_t = x) - \mathbb{I}_x(y)\Bigg| \colon x \in \hat{x}\Bigg\} \\ &\leq \max\Bigg\{\sum_{y \in \hat{y}} \frac{1}{\Delta}|P(X_{t+\Delta} = y \mid X_t = x) - \mathbb{I}_x(y)| \colon x \in \hat{x}\Bigg\}.\end{aligned}$$

The inequality that we set out to prove follows almost immediately from the above inequality:

$$\begin{aligned}&\limsup_{\Delta \to 0^+} \frac{1}{\Delta}|\hat{P}(\hat{X}_{t+\Delta} = \hat{y} \mid \hat{X}_u = \hat{x}, \hat{X}_t = \hat{x}) - \mathbb{I}_{\hat{x}}(\hat{y})| \\ &\quad\leq \limsup_{\Delta \to 0^+} \max\Bigg\{\sum_{y \in \hat{y}} \frac{1}{\Delta}|P(X_{t+\Delta} = y \mid X_t = x) - \mathbb{I}_x(y)| \colon x \in \hat{x}\Bigg\} \\ &\quad= \max\Bigg\{\limsup_{\Delta \to 0^+} \sum_{y \in \hat{y}} \frac{1}{\Delta}|P(X_{t+\Delta} = y \mid X_t = x) - \mathbb{I}_x(y)| \colon x \in \hat{x}\Bigg\},\end{aligned}$$

where the equality holds because we may change the order of the supremum and the maximum and then subsequently—since the maximum is taken over a finite set—of the limit and the maximum. Finally, we use that the limit superior is sub-additive, to yield

$$\begin{aligned}\limsup_{\Delta\to 0^+} &\frac{1}{\Delta}|\hat{P}(\hat{X}_{t+\Delta}=\hat{y}\mid \hat{X}_u=\hat{x},\hat{X}_t=\hat{x})-\mathbb{I}_{\hat{x}}(\hat{y})| \\ &\leq \max\left\{\sum_{y\in\hat{y}}\limsup_{\Delta\to 0^+}\frac{1}{\Delta}|P(X_{t+\Delta}=y\mid X_t=x)-\mathbb{I}_x(y)|\colon x\in\hat{x}\right\}<+\infty,\end{aligned}$$

where the final inequality holds because P is well-behaved and the maximum is taken over a finite set. □

Proposition 35. *Consider a homogeneous Markov chain P and a lumping map $\Lambda\colon \mathscr{X}\to\hat{\mathscr{X}}$, and let $\hat{P}$ be a lumped stochastic process. Then for any t in $\mathbb{R}_{\geq 0}$, $u\in\mathscr{U}_{<t}$ and $\hat{x}_u$ in $\hat{\mathscr{X}}_u$,*

$$\overline{\partial}\hat{T}^s_{t,\hat{x}_u}\subseteq\hat{\mathscr{Q}},$$

where $\overline{\partial}\hat{T}^s_{t,\hat{x}_u}$ denotes the outer partial derivative—as defined in [12, Definition 4.8]—of the history-dependent transition matrix $\hat{T}^s_{t,\hat{x}_u}$—see [12, Definition 4.6]—that, for all s in $\mathbb{R}_{\geq 0}$ with $s\geq t$, is defined by

$$\hat{T}^s_{t,\hat{x}_u}(\hat{x},\hat{y}) := \hat{P}(\hat{X}_s=\hat{y}\mid\hat{X}_u=\hat{x}_u,\hat{X}_t=\hat{x}) \qquad \text{for all } \hat{x},\hat{y}\in\hat{\mathscr{X}}.$$

In our proof for Proposition 35, we need the following result.

Lemma 36 (Theorem 2.1.1 in [19]). *Let Q be a transition rate matrix. Then for all t,s in $\mathbb{R}_{\geq 0}$ with $t\leq s$ and all x,y in $\mathscr{X}$,*

$$\lim_{\Delta\to 0^+}\frac{T_t^{s+\Delta}(x,y)-T_t^s(x,y)}{\Delta}=[QT_t^s](x,y)$$

and, if $t\neq 0$,

$$\lim_{\Delta\to 0^+}\frac{T_t^s(x,y)-T_{t-\Delta}^s(x,y)}{\Delta}=[QT_t^s](x,y).$$

Proof. First, recall from Lemma 34 that $\hat{P}$ is well-behaved. Hence, it follows from [12, Proposition 4.6] that $\overline{\partial}\hat{T}^s_{t,\hat{x}_u}$ is a non-empty, bounded and closed subset of $\mathscr{R}(\hat{\mathscr{X}})$. Therefore, we can fix an arbitrary element $\hat{Q}^\star$ of $\overline{\partial}\hat{T}^s_{t,\hat{x}_u}$. By definition of the outer partial derivative $\overline{\partial}$, this implies that there is a monotonously decreasing sequence $\{\Delta_n\}_{n\in\mathbb{N}}$ in $\mathbb{R}_{>0}$ with $\lim_{n\to+\infty}\Delta_i=0$ such that either

$$\lim_{n\to+\infty}\frac{\hat{T}^{t+\Delta_n}_{t,\hat{x}_u}-I}{\Delta_n}=\hat{Q}^\star \quad\text{or}\quad \lim_{n\to+\infty}\frac{\hat{T}^{t}_{t-\Delta_n,\hat{x}_u}-I}{\Delta_n}=\hat{Q}^\star. \tag{D.1}$$

We now proceed our argument under the assumption that it is the left equality that holds. The argument for the alternate case is entirely similar due to the homogeneity of P.

Fix any arbitrary $\hat{f}$ in $\mathscr{L}(\hat{\mathscr{X}})$. Observe that as a consequence of Eqn. (D.1) and the equality $[I\hat{f}](\hat{x})=\hat{f}(\hat{x})$,

$$\lim_{n\to+\infty}\frac{[\hat{T}^{t+\Delta_n}_{t,\hat{x}_u}\hat{f}](\hat{x})-\hat{f}(\hat{x})}{\Delta_n}=[\hat{Q}^\star\hat{f}](\hat{x}). \tag{D.2}$$

Observe that, for all n in $\mathbb{N}$,

$$\begin{aligned}\hat{E}(\hat{f}(\hat{X}_{t+\Delta_n})\mid\hat{X}_u=\hat{x}_u,\hat{X}_t=\hat{x}) &= \sum_{\hat{y}\in\hat{\mathscr{X}}}\hat{f}(\hat{y})\hat{P}(\hat{X}_{t+\Delta_n}=\hat{y}\mid\hat{X}_u=\hat{x}_u,\hat{X}_t=\hat{x}) \\ &= \sum_{\hat{y}\in\hat{\mathscr{X}}}\hat{f}(\hat{y})\hat{T}^{t+\Delta_n}_{t,\hat{x}_u}(\hat{x},\hat{y})=[\hat{T}^{t+\Delta_n}_{t,\hat{x}_u}\hat{f}](\hat{x}),\end{aligned}$$

where the final equality follows from the linearity of $\hat{T}_{t,\hat{x}_u}^{t+\Delta_n}$. It follows from this equality, Corollary 32 and the fact that $[\hat{f}\circ\Lambda](x)=\hat{f}(\Lambda(x))=\hat{f}(\hat{x})$ for all $x\in\hat{x}$ that, for all n in $\mathbb{N}$,

$$\min\left\{\frac{[T_t^{t+\Delta_n}(\hat{f}\circ\Lambda)](x)-[\hat{f}\circ\Lambda](x)}{\Delta_n}\colon x\in\hat{x}\right\}\leq\frac{[\hat{T}_{t,\hat{x}_u}^{t+\Delta_n}\hat{f}](\hat{x})-\hat{f}(\hat{x})}{\Delta_n}. \tag{D.3}$$

Observe now that, for all n in $\mathbb{N}$ and $x\in\hat{x}$,

$$[T_t^{t+\Delta_n}(\hat{f}\circ\Lambda)](x)-[\hat{f}\circ\Lambda](x)=\sum_{y\in\mathscr{X}}[\hat{f}\circ\Lambda](y)\Big(T_t^{t+\Delta_n}(x,y)-T_t^t(x,y)\Big),$$

where we have used that $T_t^t=I$, and

$$[Q(\hat{f}\circ\Lambda)](x)=\sum_{y\in\mathscr{X}}[\hat{f}\circ\Lambda](y)Q(x,y)=\sum_{y\in\mathscr{X}}[\hat{f}\circ\Lambda](y)[QT_t^t](x,y).$$

As the sequence $\{\Delta_n\}_{n\in\mathbb{N}}$ converges to 0, it follows from Lemma 36 and the previous two equalities that

$$\begin{aligned}\lim_{n\to+\infty}\frac{[T_t^{t+\Delta_n}(\hat{f}\circ\Lambda)](x)-[\hat{f}\circ\Lambda](x)}{\Delta_n}&=\sum_{y\in\hat{y}}[\hat{f}\circ\Lambda](y)\lim_{n\to+\infty}\frac{T_t^{t+\Delta_n}(x,y)-T_t^t(x,y)}{\Delta_n}\\&=\sum_{y\in\hat{y}}[\hat{f}\circ\Lambda](y)[QT_t^t](x,y)=[Q(\hat{f}\circ\Lambda)](x).\end{aligned}$$

Fix some ϵ in $\mathbb{R}_{>0}$. Due the previous equality, and also because $\Lambda^{-1}(\hat{x})$ is finite, there is some N in $\mathbb{N}$ such that

$$(\forall n\in\mathbb{N},n\geq N)(\forall x\in\hat{x})\ \left|\frac{[T_t^{t+\Delta_n}(\hat{f}\circ\Lambda)](x)-[\hat{f}\circ\Lambda](x)}{\Delta_n}-[Q(\hat{f}\circ\Lambda)](x)\right|\leq\epsilon$$

We combine this inequality with Eqn. (D.3), to yield

$$(\forall n\in\mathbb{N},n\geq N)\ \min\big\{[Q(\hat{f}\circ\Lambda)](x)\colon x\in\hat{x}\big\}-\epsilon\leq\frac{[\hat{T}_{t,\hat{x}_u}^{t+\Delta_n}\hat{f}](\hat{x})-\hat{f}(\hat{x})}{\Delta_n}.$$

We taken the limit for n going to $+\infty$ on both sides of this inequality and use Eqn. (D.2), to yield

$$\min\big\{[Q(\hat{f}\circ\Lambda)](x)\colon x\in\hat{x}\big\}-\epsilon\leq[\hat{Q}^\star\hat{f}](\hat{x}),$$

which, since ϵ is an arbitrary positive real number, implies that $\min\{[Q(\hat{f}\circ\Lambda)](x)\colon x\in\hat{x}\}\leq[\hat{Q}^\star\hat{f}](\hat{x})$. Recall from Lemma 30 that

$$[Q(\hat{f}\circ\Lambda)](x)=\sum_{y\in\mathscr{X}}Q(x,y)[\hat{f}\circ\Lambda](y)=\sum_{\hat{y}\in\hat{\mathscr{X}}}\hat{f}(\hat{y})\sum_{y\in\hat{y}}Q(x,y)\geq[\underline{\hat{Q}}\hat{f}](\hat{x})$$

for any $x\in\hat{x}$, such that $[\underline{\hat{Q}}\hat{f}](\hat{x})\leq[\hat{Q}^\star\hat{f}](\hat{x})$. Because $\hat{f}$ was an arbitrary real-valued function on $\hat{\mathscr{X}}$, it follows from this inequality and Eqn. (21) that $\hat{Q}^\star$ is contained in $\hat{\mathscr{Q}}$. □

Theorem 6. *Consider a homogeneous Markov chain P and a lumping map $\Lambda\colon\mathscr{X}\to\hat{\mathscr{X}}$. Then any corresponding lumped stochastic process $\hat{P}$ is contained in $\mathbb{P}^{\mathrm{W}}_{\hat{\mathscr{Q}},\hat{\mathscr{M}}}$.*

Proof. Recall that $\hat{P}$ is well-behaved by Lemma 34. Furthermore, $\hat{P}$ is clearly consistent with $\hat{\mathscr{M}}$ and, by Proposition 35, also consistent with $\hat{\mathscr{Q}}$. The stated now follows because, by definition, $\mathbb{P}^{\mathrm{W}}_{\hat{\mathscr{Q}},\hat{\mathscr{M}}}$ contains all well-behaved stochastic processes that are consistent with $\hat{\mathscr{Q}}$ and $\hat{\mathscr{M}}$. □

Appendix E. Extra material for and proofs of the results in Section 6

Our proof of Theorem 8 is split up into several intermediary results. First, we link the matrix exponential T_t^s generated by the transition rate matrix Q with the non-linear transformation $\hat{\underline{T}}_t^s$ generated by the lumped lower transition rate operator $\hat{\underline{Q}}$.

Lemma 37. *Consider a transition rate matrix Q and a lumping map $\Lambda\colon \mathscr{X} \to \hat{\mathscr{X}}$. Then for any δ in $\mathbb{R}_{\geq 0}$ such that $\delta\|Q\| \leq 2$, n in $\mathbb{N}_0$ and $\hat{f}$ in $\mathscr{L}(\hat{\mathscr{X}})$,*

$$[(I + \delta\hat{\underline{Q}})^n \hat{f}] \circ \Lambda \leq (I + \delta Q)^n (\hat{f} \circ \Lambda).$$

Proof. Observe that the stated holds trivially if $n = 0$. Hence, without loss of generality, we assume that $n > 0$. Fix some arbitrary $\hat{g}$ in $\mathscr{L}(\hat{\mathscr{X}})$. Recall from Lemma 30 that

$$(\hat{\underline{Q}}\hat{g}) \circ \Lambda \leq Q(\hat{g} \circ \Lambda).$$

From this, it follows that

$$[(I + \delta\hat{\underline{Q}})\hat{g}] \circ \Lambda = \hat{g} \circ \Lambda + \delta(\hat{\underline{Q}}\hat{g}) \circ \Lambda \leq \hat{g} \circ \Lambda + \delta Q(\hat{g} \circ \Lambda) = (I + \delta Q)(\hat{g} \circ \Lambda). \tag{E.1}$$

As $\delta\|Q\| \leq 2$, it follows from Corollary 28 that $(I + \delta Q)$ is a transition matrix. Consequently, we find that

$$(I + \delta Q)^n (\hat{f} \circ \Lambda) = (I + \delta Q)^{n-1} (I + \delta Q)(\hat{f} \circ \Lambda) \geq (I + \delta Q)^{n-1} [(I + \delta\hat{\underline{Q}})\hat{f}] \circ \Lambda,$$

where the inequality follows from (T6) for the transition matrix $T = (I + \delta Q)^{n-1}$—that $(I + \delta Q)^{n-1}$ is a transition matrix follows from (T4)—and the functions $f = [(I + \delta\hat{\underline{Q}})\hat{f}] \circ \Lambda$ and $g = (I + \delta Q)(\hat{f} \circ \Lambda)$, which satisfy $f \leq g$ due to Eqn. (E.1). Repeated application of the same trick yields

$$\begin{aligned}(I + \delta Q)^n (\hat{f} \circ \Lambda) \geq (I + \delta Q)^{n-1} [(I + \delta\hat{\underline{Q}})\hat{f}] \circ \Lambda \geq (I + \delta Q)^{n-2} [(I + \delta\hat{\underline{Q}})^2 \hat{f}] \circ \Lambda \\ \geq \cdots \geq [(I + \delta\hat{\underline{Q}})^n \hat{f}] \circ \Lambda,\end{aligned}$$

as required. □

Lemma 38. *Consider a transition rate matrix Q and a lumping map $\Lambda\colon \mathscr{X} \to \hat{\mathscr{X}}$. Then for any t, s in $\mathbb{R}_{\geq 0}$ with $t \leq s$ and $\hat{f}$ in $\mathscr{L}(\hat{\mathscr{X}})$,*

$$(\hat{\underline{T}}_t^s \hat{f}) \circ \Lambda \leq T_t^s (\hat{f} \circ \Lambda),$$

where $\hat{\underline{T}}_t^s$ is the lower transition operator generated by $\hat{\underline{Q}}$ according to Eqn. (18).

Proof. For any n in $\mathbb{N}$ such that $n \geq (s - t)\|Q\|/2$, it follows from Lemma 37 that

$$\left[\left(I + \frac{s-t}{n}\hat{\underline{Q}}\right)^n \hat{f}\right] \circ \Lambda \leq \left(I + \frac{s-t}{n} Q\right)^n (\hat{f} \circ \Lambda).$$

The stated is obtained by taking the limit for n going to $+\infty$ on both sides of the inequality and substituting Eqns. (7) and (18). □

We can easily extend Lemma 38 to functions that depend on a finite number of time points. For this result, we use the notational convention introduced in Appendix A.3 and Appendix C.2.

Lemma 39. *Consider a transition rate matrix Q and a lumping map $\Lambda\colon \mathscr{X} \to \hat{\mathscr{X}}$. Then for any $u = t_1, \dots, t_n$ and $v = s_1, \dots, s_m$ in $\mathscr{U}_\emptyset$ with $\max u < \min v$ and any $\hat{f}$ in $\mathscr{L}(\hat{\mathscr{X}}_{u \cup v})$,*

$$\left(\hat{\underline{T}}_{t_n}^{s_1} \hat{\underline{T}}_{s_1}^{s_2} \cdots \hat{\underline{T}}_{s_{m-1}}^{s_m} \hat{f}\right) \circ \Lambda \leq T_{t_n}^{s_1} T_{s_1}^{s_2} \cdots T_{s_{m-1}}^{s_m} (\hat{f} \circ \Lambda).$$

Proof. Our proof is one by induction. If $m = 1$, then $s_1 = s_m = s$ and $v = s$. Fix any x_u in $\mathscr{X}_u$ and let $\hat{x}_u \coloneqq \Lambda(x_u)$. By Eqn. (A.7),

$$[T_{t_n}^{s}(\hat{f} \circ \Lambda)](x_u) = [T_{t_n}^{s}(\hat{f} \circ \Lambda)_{x_u}](x_{t_n}), \tag{E.2}$$

where $(\hat{f} \circ \Lambda)_{x_u}$ maps any x in $\mathscr{X}$ to $(\hat{f} \circ \Lambda)(x_u, x) = \hat{f}(\hat{x}_u, \Lambda(x))$. It follows from Lemma 38 that

$$T_{t_n}^{s}(\hat{f} \circ \Lambda)_{x_u} \geq (\hat{\underline{T}}_{t_n}^{s}\hat{f}_{\hat{x}_u}) \circ \Lambda, \tag{E.3}$$

where $\hat{f}_{\hat{x}_u}$ maps any $\hat{x}$ in $\hat{\mathscr{X}}$ to $\hat{f}(\hat{x}_u, \hat{x})$. Observe now that

$$\begin{aligned}[T_{t_n}^{s}(\hat{f} \circ \Lambda)](x_u) = [T_{t_n}^{s}(\hat{f} \circ \Lambda)_{x_u}](x_{t_n}) \geq [(\hat{\underline{T}}_{t_n}^{s}\hat{f}_{\hat{x}_u}) \circ \Lambda](x_{t_n}) &= [\hat{\underline{T}}_{t_n}^{s}\hat{f}_{\hat{x}_u}](\hat{x}_{t_n}) \\ &= [\hat{\underline{T}}_{t_n}^{s}\hat{f}](\hat{x}_u) = [(\hat{\underline{T}}_{t_n}^{s}\hat{f}) \circ \Lambda](x_u),\end{aligned}$$

where the first equality follows from Eqn. (E.2), the inequality follows from (E.3) and the third equality follows from Eqn. (C.1). Because x_u was an arbitrary state instantiation in $\mathscr{X}_u$, this inequality implies the stated for $m = 1$.

Next, we fix some m in $\mathbb{N}$ such that $m \geq 2$, and assume that the stated holds for any sequence v that satisfies the conditions of the statement and has length $m' < m$. We now show that this then implies the stated for any sequence v with length m. We apply the induction hypothesis with the sequences $u' \coloneqq t_1, \dots, t_n, s_1$ and $v' \coloneqq s_2, \dots, s_m$, to yield

$$\left(\hat{\underline{T}}_{s_1}^{s_2}\hat{\underline{T}}_{s_2}^{s_3}\cdots\hat{\underline{T}}_{s_{m-1}}^{s_m}\hat{f}\right) \circ \Lambda \leq T_{s_1}^{s_2}T_{s_2}^{s_3}\cdots T_{s_{m-1}}^{s_m}(\hat{f} \circ \Lambda). \tag{E.4}$$

Observe that both sides of this inequality are functions on $\mathscr{X}_{u'} = \mathscr{X}_{u \cup \{s_1\}}$, and that the one on the left is clearly lumpable. We write this inequality as $\hat{g} \circ \Lambda \leq h$, where we let

$$\hat{g} \coloneqq \hat{\underline{T}}_{s_1}^{s_2}\hat{\underline{T}}_{s_2}^{s_3}\cdots\hat{\underline{T}}_{s_{m-1}}^{s_m}\hat{f} \qquad \text{and} \qquad h \coloneqq T_{s_1}^{s_2}T_{s_2}^{s_3}\cdots T_{s_{m-1}}^{s_m}(\hat{f} \circ \Lambda).$$

Fix some x_u in $\mathscr{X}_u$, and let $\hat{x}_u \coloneqq \Lambda(x_u)$. From $\hat{g} \circ \Lambda \leq h$, we infer that $(\hat{g} \circ \Lambda)_{x_u} \leq h_{x_u}$. Therefore,

$$[T_{t_n}^{s_1}(\hat{g} \circ \Lambda)](x_u) = [T_{t_n}^{s_1}(\hat{g} \circ \Lambda)_{x_u}](x_{t_n}) \leq [T_{t_n}^{s_1}h_{x_u}](x_{t_n}) = [T_{t_n}^{s_1}h](x_u), \tag{E.5}$$

where the equalities follow from Eqn. (A.7) and the inequality holds due to (T6) and $(\hat{g} \circ \Lambda)_{x_u} \leq h_{x_u}$. Furthermore, we observe that $(\hat{g} \circ \Lambda)_{x_u} = \hat{g}_{\hat{x}_u} \circ \Lambda$. Therefore, it follows from Lemma 38 that

$$\begin{aligned}[T_{t_n}^{s_1}(\hat{g} \circ \Lambda)](x_u) = [T_{t_n}^{s_1}(\hat{g} \circ \Lambda)_{x_u}](x_{t_n}) &= [T_{t_n}^{s_1}(\hat{g}_{\hat{x}_u} \circ \Lambda)](x_{t_n}) \\ &\geq [\hat{\underline{T}}_{t_n}^{s_1}\hat{g}_{\hat{x}_u}](\hat{x}_{t_n}) = [\hat{\underline{T}}_{t_n}^{s_1}\hat{g}](\hat{x}_u) = [(\hat{\underline{T}}_{t_n}^{s_1}\hat{g}) \circ \Lambda](x_u),\end{aligned} \tag{E.6}$$

where the first equality follows from Eqn. (A.7) and the third equality follows from Eqn. (C.1). We now combine Eqns. (E.5) and (E.6) and substitute the definitions of $\hat{g}$ and h, to yield

$$[(\hat{\underline{T}}_{t_n}^{s_1}\hat{\underline{T}}_{s_1}^{s_2}\hat{\underline{T}}_{s_2}^{s_3}\cdots\hat{\underline{T}}_{s_{m-1}}^{s_m}\hat{f}) \circ \Lambda](x_u) \leq [T_{t_n}^{s_1}T_{s_1}^{s_2}T_{s_2}^{s_3}\cdots T_{s_{m-1}}^{s_m}(\hat{f} \circ \Lambda)](x_u).$$

The induction step now follows from this inequality because x_u was an arbitrary state instantiation in $\mathscr{X}_u$. □

The last intermediary result that we need is a formalisation of Eqn. (22).

Lemma 40. *Consider a lumping map $\Lambda \colon \mathscr{X} \to \hat{\mathscr{X}}$. Then for any u in $\mathscr{U}_\emptyset$ and f in $\mathscr{L}(\mathscr{X}_u)$,*

$$\hat{f}_{\mathrm{L}} \circ \Lambda \leq f \leq \hat{f}_{\mathrm{U}} \circ \Lambda.$$

If f is furthermore lumpable with respect to Λ, then $\hat{f}_{\mathrm{L}} = \hat{f}_{\mathrm{U}} = \hat{f}$ with $f = \hat{f} \circ \Lambda$.

Everything is now set up to prove the following main result.

Theorem 8. *Consider a homogeneous Markov chain P and a lumping map $\Lambda\colon \mathscr{X} \to \hat{\mathscr{X}}$. Then for all u in $\mathscr{U}$, v in $\mathscr{U}_\emptyset$ with $\max u < \min v$, x_u in $\mathscr{X}_u$ and f in $\mathscr{L}(\mathscr{X}_{u\cup v})$,*

$$\underline{E}^{\mathrm{W}}_{\hat{\mathscr{Q}},\hat{\mathscr{M}}}(\hat{f}_{\mathrm{L}}(\hat{X}_u,\hat{X}_v) \mid \hat{X}_u = \hat{x}_u) \leq E(f(X_u,X_v) \mid X_u = x_u) \leq \overline{E}^{\mathrm{W}}_{\hat{\mathscr{Q}},\hat{\mathscr{M}}}(\hat{f}_{\mathrm{U}}(\hat{X}_u,\hat{X}_v) \mid \hat{X}_u = \hat{x}_u), \tag{23}$$

where $\hat{x}_u \coloneqq \Lambda(x_u)$.

Proof. We only need to prove the left inequality, as the right inequality is obtained by applying the left inequality to $g \coloneqq -f$ because $E(g(X_u,X_v) \mid X_u = x_u) = -E(f(X_u,X_v) \mid X_u = x_u)$ and $\hat{g}_{\mathrm{L}} = -\hat{f}_{\mathrm{U}}$.

Observe that, as E is monotonous, it follows from Lemma 40 that

$$E(f(X_u,X_v) \mid X_u = x_u) \geq E((\hat{f}_{\mathrm{L}} \circ \Lambda)(X_u,X_v) \mid X_u = x_u). \tag{E.7}$$

We now distinguish between two cases.

First, we assume that $u = t_1,\dots,t_n$ is not equal to the empty sequence. In this case, it follows from Eqn. (A.8) that

$$E((\hat{f}_{\mathrm{L}} \circ \Lambda)(X_u,X_v) \mid X_u = x_u) = [T_{t_n}^{s_1} T_{s_1}^{s_2} \cdots T_{s_{m-1}}^{s_m} (\hat{f}_{\mathrm{L}} \circ \Lambda)](x_u),$$

where $v = s_1,\dots,s_m$. We now use Lemma 39, to yield

$$\begin{aligned} E((\hat{f}_{\mathrm{L}} \circ \Lambda)(X_u,X_v) \mid X_u = x_u) &= [T_{t_n}^{s_1} T_{s_1}^{s_2} \cdots T_{s_{m-1}}^{s_m} (\hat{f}_{\mathrm{L}} \circ \Lambda)](x_u) \\ &\geq [\underline{\hat{T}}_{t_n}^{s_1} \underline{\hat{T}}_{s_1}^{s_2} \cdots \underline{\hat{T}}_{s_{m-1}}^{s_m} \hat{f}_{\mathrm{L}}](\hat{x}_u), \end{aligned}$$

where $\hat{x}_u \coloneqq \Lambda(x_u)$. Recall from Proposition 4 that $\hat{\mathscr{Q}}$ is non-empty, bounded and convex and has separately specified rows. Therefore, it follows from Eqn. (C.2) and the above inequality that

$$\begin{aligned} E((\hat{f}_{\mathrm{L}} \circ \Lambda)(X_u,X_v) \mid X_u = x_u) &\geq [\underline{\hat{T}}_{t_n}^{s_1} \underline{\hat{T}}_{s_1}^{s_2} \cdots \underline{\hat{T}}_{s_{m-1}}^{s_m} \hat{f}_{\mathrm{L}}](\hat{x}_u) \\ &= \underline{E}^{\mathrm{W}}_{\hat{\mathscr{Q}},\hat{\mathscr{M}}}(\hat{f}_{\mathrm{L}}(\hat{X}_u,\hat{X}_v) \mid \hat{X}_u = \hat{x}_u). \end{aligned}$$

We now combine this inequality with Eqn. (E.7) to obtain the stated.

Next, we assume that $u = \emptyset$. Then

$$E((\hat{f}_{\mathrm{L}} \circ \Lambda)(X_v) \mid X_u = x_u) = E((\hat{f}_{\mathrm{L}} \circ \Lambda)(X_v) \mid X_\emptyset = x_\emptyset) = E((\hat{f}_{\mathrm{L}} \circ \Lambda)(X_v)),$$

where without loss of generality we have dropped the first argument X_u of $\hat{f}_{\mathrm{L}} \circ \Lambda$. Let $v = s_1,\dots,s_m$. If $\min v = s_1 > 0$, then we let $v^\star \coloneqq 0, s_1,\dots,s_m = \{0\} \cup v$ and let $f^\star$ be the extension of f to $\mathscr{X}_{v^\star}$, as explained in Appendix A.3. If $\min v = s_1 = 0$, then we simply let $v^\star \coloneqq v$ and $f^\star \coloneqq f$. This way, we have

$$E((\hat{f}_{\mathrm{L}} \circ \Lambda)(X_v)) = E((\hat{f}^\star_{\mathrm{L}} \circ \Lambda)(X_{v^\star})) = E_{\pi_0}(T_0^{s_1} \cdots T_{s_{m-1}}^{s_m} (\hat{f}^\star_{\mathrm{L}} \circ \Lambda)),$$

where the second equality follows from Eqn. (A.9). Note that in this expression, $T_0^{s_1}$ is interpreted differently depending on s_1. If $s_1 > 0$, then $T_0^{s_1}$ is the operator on $\mathscr{X}_{\{0,s_1\}}$ as defined by Eqn. (A.7). If $s_1 = 0$, then $T_0^{s_1}$ is not needed in the expression, but we can leave it anyway as $T_0^{s_1} = T_0^0 = I$. Observe now that

$$\begin{aligned} E((\hat{f}_{\mathrm{L}} \circ \Lambda)(X_v) \mid X_u = x_u) &= E_{\pi_0}(T_0^{s_1} \cdots T_{s_{m-1}}^{s_m} (\hat{f}^\star_{\mathrm{L}} \circ \Lambda)) \\ &= \langle \pi_0, T_0^{s_1} \cdots T_{s_{m-1}}^{s_m} (\hat{f}^\star_{\mathrm{L}} \circ \Lambda) \rangle = \sum_{x \in \mathscr{X}} \pi_0(x) [T_0^{s_1} \cdots T_{s_{m-1}}^{s_m} (\hat{f}^\star_{\mathrm{L}} \circ \Lambda)](x) \\ &\geq \sum_{x \in \mathscr{X}} \pi_0(x) [\underline{\hat{T}}_0^{s_1} \cdots \underline{\hat{T}}_{s_{m-1}}^{s_m} \hat{f}^\star_{\mathrm{L}}](\Lambda(x)) = \sum_{\hat{x} \in \hat{\mathscr{X}}} [\underline{\hat{T}}_0^{s_1} \cdots \underline{\hat{T}}_{s_{m-1}}^{s_m} \hat{f}^\star_{\mathrm{L}}](\hat{x}) \sum_{x \in \hat{x}} \pi_0(x) \\ &= \sum_{\hat{x} \in \hat{\mathscr{X}}} [\underline{\hat{T}}_0^{s_1} \cdots \underline{\hat{T}}_{s_{m-1}}^{s_m} \hat{f}^\star_{\mathrm{L}}](\hat{x}) \hat{\pi}_0(\hat{x}) = \langle \hat{\pi}_0, \underline{\hat{T}}_0^{s_1} \cdots \underline{\hat{T}}_{s_{m-1}}^{s_m} \hat{f}^\star_{\mathrm{L}} \rangle \\ &= E_{\hat{\pi}_0}(\underline{\hat{T}}_0^{s_1} \cdots \underline{\hat{T}}_{s_{m-1}}^{s_m} \hat{f}^\star_{\mathrm{L}}) = \underline{E}_{\hat{\mathscr{M}}}(\underline{\hat{T}}_0^{s_1} \cdots \underline{\hat{T}}_{s_{m-1}}^{s_m} \hat{f}^\star_{\mathrm{L}}) \\ &= \underline{E}^{\mathrm{W}}_{\hat{\mathscr{Q}},\hat{\mathscr{M}}}(\hat{f}^\star_{\mathrm{L}}(\hat{X}_{v^\star})) = \underline{E}^{\mathrm{W}}_{\hat{\mathscr{Q}},\hat{\mathscr{M}}}(\hat{f}_{\mathrm{L}}(\hat{X}_v)) = \underline{E}^{\mathrm{W}}_{\hat{\mathscr{Q}},\hat{\mathscr{M}}}(\hat{f}_{\mathrm{L}}(\hat{X}_v) \mid \hat{X}_u = \hat{x}_u), \end{aligned}$$

where the inequality follows from Lemma 38 and, if $s_1 = 0$, the fact that $\underline{\hat{T}}_0^{s_1} = \underline{\hat{T}}_0^0 = I$, and where the penultimate equality follows from Eqn. (C.3). The stated is now immediately obtained by combining (E.7) with the above inequality. □

Appendix F. Extra material for and proofs of the results in Section 7

Appendix F.1. Ergodicity and irreducibility

The following three results are essentially well-known, but we did not find a good reference for them.

Proposition 9. *A transition rate matrix Q is ergodic if and only if*

$$\mathscr{X}_{\mathrm{top}} \coloneqq \{x \in \mathscr{X} \colon (\forall y \in \mathscr{X})\ y \rightsquigarrow x\} \neq \emptyset.$$

Proof. First, we observe that Definition 2 is just a specialisation of Definition 7 further on—the definition of ergodicity for a lower transition rate operator—to transition rate matrices. Next, we observe that for the transition rate matrix Q, the upper reachability relation $\cdot \rightarrowtail \cdot$—see Appendix F.2 further on—is equivalent to our accessibility relation $\cdot \rightsquigarrow \cdot$. This upper reachability relation $\cdot \rightarrowtail \cdot$ is essential to the necessary and sufficient condition for ergodicity of a lower transition rate operator established in [13, Theorem 19]. The condition exists of two parts: top class regularity and top class absorption. The transition rate matrix Q satisfies top class regularity if and only $\mathscr{X}_{\mathrm{top}} \neq \emptyset$. We now claim that in this case, Q also satisfies top class absorption. More specifically, the top class $\mathscr{X}_{\mathrm{top}}$ is lower reachable—in the sense of [13, Definition 8]—from any state x not in this top class. If this claim is true, then it follows from [13, Theorem 19] that Q is ergodic if and only if

$$\mathscr{X}_{\mathrm{top}} = \{x \in \mathscr{X} \colon (\forall y \in \mathscr{X})\ y \rightsquigarrow x\} \neq \emptyset,$$

as we had to prove.

We now set out to verify our claim about top class absorption. By [13, Definition 8], this claim is verified if we can prove that $\mathscr{X} \setminus \mathscr{X}_{\mathrm{top}} \subseteq A_n$, where the non-decreasing sequence $A_0, A_1, \dots$ is defined by the initial condition $A_0 \coloneqq \mathscr{X}_{\mathrm{top}}$ and, for all k in $\mathbb{N}_0$, the recursive relation

$$A_{k+1} \coloneqq A_k \cup \{z \in \mathscr{X} \setminus A_k \colon [Q\mathbb{I}_{A_k}](z) > 0\},$$

and where n is the first index such that $A_n = A_{n+1}$. To prove this, we fix some x in $\mathscr{X} \setminus \mathscr{X}_{\mathrm{top}}$ and y in $\mathscr{X}_{\mathrm{top}}$. Observe that $x \rightsquigarrow y$ as y belongs to the top class $\mathscr{X}_{\mathrm{top}}$; hence, there is a sequence $x = x_0, x_1 \dots, x_m = y$ in $\mathscr{X}$ such that, for all i in $\{1, \dots, m\}$, $[Q\mathbb{I}_{x_i}](x_{i-1}) = Q(x_{i-1}, x_i) > 0$.

Observe that, for all k in $\mathbb{N}_0$, z_1 in $\mathscr{X} \setminus A_k$ and z_2 in A_k, $Q(z_1, z_2) \geq 0$ because $z_1 \neq z_2$ and Q is a transition rate matrix. Hence,

$$[Q\mathbb{I}_{A_k}](z_1) = \sum_{z_2 \in A_k} Q(z_1, z_2) > 0 \Leftrightarrow (\exists z^\star \in A_k)\ Q(z_1, z^\star) > 0. \tag{F.1}$$

We now claim that x_{m-1} belongs to A_1. This claim is trivially verified if $x_{m-1} \in A_0$, as $A_0 \subseteq A_1$ by definition. If $x_{m-1} \notin A_0$, then because y belongs to $A_0 = \mathscr{X}_{\mathrm{top}}$ and $Q(x_{m-1}, y) > 0$, it follows from Eqn. (F.1)—with $k = 0$, $z_1 = x$ and $z^\star = y$—that $[Q\mathbb{I}_{A_0}](x_{m-1}) > 0$, which verifies our claim in this case.

Next, we verify that x_{m-2} belongs to A_2. We again distinguish two cases. If $x_{m-2} \in A_1$, then this claim is trivially true because $A_1 \subseteq A_2$ by construction. If $x_{m-2} \notin A_1$, then because x_{m-1} belongs to A_1 and $Q(x_{m-2}, x_{m-1}) > 0$, it follows from Eqn. (F.1)—with $k = 1$, $z_1 = x_{m-2}$ and $z^\star = x_{m-1}$—that $[Q\mathbb{I}_{A_1}](x_{m-2}) > 0$, from which we infer the veracity of our claim.

It is clear that we can repeat the same argument to verify that x_{m-3} belongs to A_3, and so on. If we continue this way, we eventually obtain that $x = x_0$ is an element of A_m. This implies that x is an element of A_n as well because $A_0 \subseteq A_1 \subseteq \cdots \subseteq A_m$ and $A_n = A_{n+k}$ for all k in $\mathbb{N}$. Because x was an arbitrary element of $\mathscr{X} \setminus \mathscr{X}_{\mathrm{top}}$, we infer from this that $\mathscr{X} \setminus \mathscr{X}_{\mathrm{top}} \subseteq A_n$, as required. □

Proposition 10. *Let Q be an ergodic transition rate matrix. Then the matrix Q' on $\mathscr{L}(\mathscr{X}_{\mathrm{top}})$, defined by*

$$Q'(x,y) \coloneqq Q(x,y) \text{ for all } x,y \text{ in } \mathscr{X}_{\mathrm{top}},$$

is an irreducible transition rate matrix. Furthermore, for all f in $\mathscr{L}(\mathscr{X})$, $\langle \pi_\infty, f\rangle = \langle \pi'_\infty, f'\rangle$, where π'_∞ is the limit distribution of Q' and f' is the restriction of f to $\mathscr{X}_{\mathrm{top}}$.

Proof. Our proof hinges on the following claim:

$$Q(x,y) = 0 \text{ for all } x \in \mathscr{X}_{\mathrm{top}} \text{ and } y \in \mathscr{X} \setminus \mathscr{X}_{\mathrm{top}} \tag{F.2}$$

To verify this claim, we fix any such x and y, and assume ex-absurdo that $Q(x,y) \neq 0$. As $x \neq y$, this is equivalent to $Q(x,y) > 0$ because $Q(x,y) \geq 0$ since Q is a transition rate matrix. Fix any arbitrary z in $\mathscr{X}$. As x is a state in the top class, we know that $z \rightsquigarrow x$, or equivalently, that there is a sequence $z = x_0, \dots, x_n = x$ such that $Q(x_{i-1}, x_i) > 0$ for all i in $\{1, \dots, n\}$. If we let $x_{n+1} \coloneqq y$, then clearly $Q(x_{i-1}, x_i) > 0$ for all i in $\{1, \dots, n+1\}$; hence, $z \rightsquigarrow y$. As z was an arbitrary state, this implies that y is a state in the top class $\mathscr{X}_{\mathrm{top}}$, which contradicts our initial assumption.

We now use Eqn. (F.2) to verify that Q' is an irreducible transition rate matrix. That Q' is a transition rate matrix—that is, that it has non-negative off-diagonal elements and rows that sum up to zero—follows almost immediately from Eqn. (F.2) and the same properties for Q. Hence, we focus on verifying that Q' is irreducible. To verify this, we need to show that for any arbitrary x and y in $\mathscr{X}_{\mathrm{top}}$, there is a sequence $x = x_0, \dots, x_n = y$ in $\mathscr{X}_{\mathrm{top}}$ such that $Q'(x_{i-1}, x_i) > 0$ for all i in $\{1, \dots, n\}$. To that end, we fix any arbitrary x and y in $\mathscr{X}_{\mathrm{top}}$. Because y belongs to the top class $\mathscr{X}_{\mathrm{top}}$, there is a sequence $x = x_0, \dots, x_n = y$ in $\mathscr{X}$ such that $Q(x_{i-1}, x_i) > 0$ for all i in $\{1, \dots, n\}$. As x_0 is in the top class and $Q(x_0, x_1) > 0$, it follows from Eqn. (F.2) that x_1 belongs to the top class $\mathscr{X}_{\mathrm{top}}$. Repeating this argument, we obtain that the entire sequence $x_1, \dots, x_n$ belongs to the top class $\mathscr{X}_{\mathrm{top}}$. Consequently, $Q'(x_{i-1}, x_i) = Q(x_{i-1}, x_i) > 0$ for all i in $\{1, \dots, n\}$, as required.

To prove the second part of the stated, we recall from Section 7.3 that as Q' is irreducible, it has a unique limit distribution π'_∞ that satisfies the equilibrium condition

$$(\forall y \in \mathscr{X}_{\mathrm{top}}) \sum_{x \in \mathscr{X}_{\mathrm{top}}} \pi'_\infty(x) Q'(x,y) = 0.$$

Observe that, for all y in $\mathscr{X}_{\mathrm{top}}$,

$$0 = \sum_{x \in \mathscr{X}_{\mathrm{top}}} \pi'_\infty(x) Q(x,y) = \sum_{x \in \mathscr{X}} \pi^\star(x) Q(x,y), \tag{F.3}$$

where we let $\pi^\star$ be the distribution on $\mathscr{X}$ given for all x in $\mathscr{X}$ by $\pi^\star(x) \coloneqq \pi'_\infty(x)$ if x is in $\mathscr{X}_{\mathrm{top}}$ and $\pi^\star(x) \coloneqq 0$ otherwise. Furthermore, it now follows from Eqn. (F.2) and our definition of $\pi^\star$ that

$$\sum_{x \in \mathscr{X}} \pi^\star(x) Q(x,y) = 0 \text{ for all } y \in \mathscr{X} \setminus \mathscr{X}_{\mathrm{top}}.$$

The above equality and Eqn. (F.3) imply that $\pi^\star$ satisfies the equilibrium condition for Q. As the only distribution that satisfies this equilibrium condition is the limit distribution π_∞ of Q, we have proven that $\pi^\star = \pi_\infty$. To obtain the second part of the stated, we observe that $\langle \pi_\infty, f\rangle = \langle \pi^\star, f\rangle = \langle \pi', f'\rangle$, where the second equality holds because $\pi^\star(x) = 0$ for all x in $\mathscr{X} \setminus \mathscr{X}_{\mathrm{top}}$. □

Lemma 41. *Let Q be a transition rate matrix and δ in $\mathbb{R}_{>0}$ such that $\delta\|Q\| < 2$. If Q is ergodic, then $(I + \delta Q)$ is ergodic in the sense of [27, Definition 4.7].*

Proof. Fix any δ in $\mathbb{R}_{>0}$ such that $\delta\|Q\| < 2$. By Lemma 28, $T \coloneqq (I + \delta Q)$ is a transition matrix. That T is ergodic if Q is ergodic follows from several other results. It follows from [25, Theorem 8], [28, Proposition 7] and [28, Definition 2] that for all f in $\mathscr{L}(\mathscr{X})$, $\lim_{n \to +\infty} \langle \pi_0, T^n f\rangle$ exists and does not depend on the initial distribution π_0 in $\mathscr{D}(\mathscr{X})$. Because this is clearly equivalent to the condition in [27, Definition 4.7], we have proven the stated. □

Appendix F.2. Ergodicity and irreducibility in the imprecise case

Just like precise Markov chains, their imprecise counterparts also have some nice ergodic properties. For a detailed exposition of these properties, we refer the interested reader to our previous work [13, 25]. We here only mention the definitions and results that we will need in the remainder.

Definition 7 (Definition 6 in [13])**.** The lower transition rate operator $\underline{Q}$ is *ergodic* if, for any $f \in \mathscr{L}(\mathscr{X})$, $\lim_{t\to+\infty} \underline{T}_0^t f$ is a constant function.

Note the similarity between Definition 2 and Definition 7. In fact, the former is just the precise version of the latter. Therefore, it should not come as a surprise that the accessibility relation $\cdot \rightsquigarrow \cdot$ has an imprecise counterpart. Following [13, Definition 7], we say that a state x is *upper reachable* from the state y, denoted by $y \rightarrowtail x$, if there is a sequence $y = x_0, \dots, x_n = x$ in $\mathscr{X}$ such that $[\overline{Q}\mathbb{I}_{x_i}](x_{i-1}) = -[\underline{Q}(-\mathbb{I}_{x_i})](x_{i-1}) > 0$ for all i in $\{1, \dots, n\}$.

Definition 8. The lower transition rate operator $\underline{Q}$ is called *irreducible* if

$$\mathscr{X}_{\text{top}} := \{x \in \mathscr{X} : (\forall y \in \mathscr{X})\ y \rightarrowtail x\} = \mathscr{X}.$$

Corollary 42. *If the lower transition rate operator $\underline{Q}$ is irreducible, then it is ergodic.*

Proof. This is an immediate consequence of [13, Theorem 19]. □

Note that Corollary 42 resembles Proposition 9, although the latter establishes a necessary and sufficient condition for the ergodicity of a transition rate matrix Q while the former only establishes a sufficient condition for the ergodicity of a lower transition rate operator $\underline{Q}$. Similarly, the following result resembles Proposition 11, in the sense that it establishes the convergence of $(I + \delta\underline{Q})^n f$.

Corollary 43. *If the lower transition rate operator $\underline{Q}$ is ergodic, then for any f in $\mathscr{L}(\mathscr{X})$ and δ in $\mathbb{R}_{>0}$ with $\delta\|\underline{Q}\| < 2$, $(I + \delta\underline{Q})^n f$ converges to a constant function in the limit for $n \to +\infty$.*

Proof. Fix any δ in $\mathbb{R}_{>0}$ such that $\delta\|\underline{Q}\| < 2$ and let $\underline{T} := (I + \delta\underline{Q})$. Then by [25, Proposition 3], $\underline{T}$ is a lower transition operator. Furthermore, since $\underline{Q}$ is ergodic, it follows from [25, Theorem 8] and either [28, Proposition 7] or [29, Theorem 21] that the lower transition operator $\underline{T}$ is also *ergodic*, meaning that, for all f in $\mathscr{L}(\mathscr{X})$, $\lim_{n\to+\infty} \underline{T}^n f = \lim_{n\to+\infty}(I + \delta\underline{Q})^n f$ exists and is a constant function. □

Appendix F.3. Bounding limit expectations

Lemma 44. *Let Q be a transition rate matrix and $\Lambda\colon \mathscr{X} \to \hat{\mathscr{X}}$ a lumping map. If Q is irreducible, then $\hat{\underline{Q}}$ is irreducible as well.*

Proof. To prove that $\hat{\underline{Q}}$ is irreducible, we need to verify that $\hat{x} \rightarrowtail \hat{y}$ for all $\hat{x}, \hat{y}$ in $\hat{\mathscr{X}}$. Observe that, by definition, $\hat{x} \rightarrowtail \hat{x}$ for all $\hat{x}$ in $\hat{\mathscr{X}}$. Hence, we now consider any $\hat{x}, \hat{y}$ in $\hat{\mathscr{X}}$ such that $\hat{x} \neq \hat{y}$, and verify that indeed $\hat{x} \rightarrowtail \hat{y}$.

Fix any x in $\Lambda^{-1}(\hat{x})$ and y in $\Lambda^{-1}(\hat{y})$. Then as Q is irreducible, it follows from Proposition 9 that there is a sequence $x = x_0, \dots, x_n = y$ in $\mathscr{X}$ such that $Q(x_{i-1}, x_i) > 0$ for all i in $\{1, \dots, n\}$. If for all i in $\{0, \dots, n\}$ we let $\hat{x}_i := \Lambda(x_i)$, then $\hat{x}_0, \dots, \hat{x}_n$ is obviously a sequence in $\hat{\mathscr{X}}$ such that $\hat{x}_0 = \hat{x}$ and $\hat{x}_n = \hat{y}$. It may occur for several indices j in $\{0, \dots, n-1\}$ that there are consecutive entries $\hat{x}_j, \hat{x}_{j+1}, \dots$ that are all equal to $\hat{x}_j$. For each of those indices j we delete these consecutive entries $\hat{x}_{j+1}, \dots$ from the sequence; this way, we end up with the shorter sequence $\hat{x}_{i_0}, \dots, \hat{x}_{i_m}$ in $\hat{\mathscr{X}}$, where $\{i_0, \dots, i_m\}$ is an increasing subsequence of $\{1, \dots, n\}$. Note that by construction $\hat{x}_{i_0} = \hat{x}$, $\hat{x}_{i_m} = \hat{y}$ and $\hat{x}_{i_{(k-1)}} \neq \hat{x}_{i_k}$ for all k in $\{1, \dots, m\}$.

Fix now any k in $\{1, \dots, m\}$. Unfortunately, it does not necessarily hold that $Q(x_{i_{(k-1)}}, x_{i_k}) > 0$. However, we have removed the consecutive entries in such a way that $Q(x_{i_k-1}, x_{i_k}) > 0$. Because clearly $x_{i_k-1} \in \hat{x}_{i_{(k-1)}}$

and $x_{i_k} \in \hat{x}_{i_k}$, it now follows that

$$\begin{aligned}
[\hat{\overline{Q}}\mathbb{I}_{\hat{x}_{i_k}}](\hat{x}_{i_{(k-1)}}) &= -[\underline{\hat{Q}}(-\mathbb{I}_{\hat{x}_{i_k}})](\hat{x}_{i_{(k-1)}}) \\
&= -\min\left\{-\sum_{\hat{y}\in\hat{\mathscr{X}}} \mathbb{I}_{\hat{x}_{i_k}}(\hat{y}) \sum_{y\in\hat{y}} Q(x,y) \colon x \in \hat{x}_{i_{(k-1)}}\right\} \\
&= \max\left\{\sum_{\hat{y}\in\hat{\mathscr{X}}} \mathbb{I}_{\hat{x}_{i_k}}(\hat{y}) \sum_{y\in\hat{y}} Q(x,y) \colon x \in \hat{x}_{i_{(k-1)}}\right\} = \max\left\{\sum_{y\in\hat{x}_{i_k}} Q(x,y) \colon x \in \hat{x}_{i_{(k-1)}}\right\} \\
&\geq \sum_{y\in\hat{x}_{i_k}} Q(x_{i_k-1},y) \geq Q(x_{i_k-1},x_{i_k}) > 0,
\end{aligned}$$

where the second inequality holds because $\hat{x}_{i_{(k-1)}} \neq \hat{x}_{i_k}$ implies that $x_{i_k-1} \neq y$ for all y in $\Lambda^{-1}(\hat{x}_{i_k})$. Consequently, $\hat{y} \rightarrowtail \hat{x}$, as desired. □

Proposition 11. *If Q is an ergodic transition rate matrix, then for all f in $\mathscr{L}(\mathscr{X})$, δ in $\mathbb{R}_{>0}$ with $\delta\|Q\| < 2$, and n in $\mathbb{N}_0$,*

$$\min(I+\delta Q)^n f \leq \langle \pi_\infty, f\rangle \leq \max(I+\delta Q)^n f.$$

Furthermore, the lower and upper bounds in this expression become monotonously tighter with increasing n, and converge to $\langle \pi_\infty, f\rangle$ as n approaches $+\infty$.

Proof. Fix any δ in $\mathbb{R}_{>0}$ such that $\delta\|Q\| < 2$, and let $T := I + \delta Q$. We recall from Lemma 41 that T is an ergodic transition matrix. For all f in $\mathscr{L}(\mathscr{X})$,

$$\langle \pi_\infty, Tf\rangle = E_{\pi_\infty}(Tf) = E_{\pi_\infty}((I+\delta Q)f) = E_{\pi_\infty}(f) + \delta E_{\pi_\infty}(Qf) = \langle \pi_\infty, f\rangle + \delta\langle \pi_\infty, Qf\rangle = \langle \pi_\infty, f\rangle,$$

where the third equality follows from the linearity of the expectation E_{π_∞} and the final equality holds because $\langle \pi_\infty, Qf\rangle = 0$ due to the equilibrium condition for Q. This equality is exactly the equilibrium condition for the ergodic transition matrix T. As π_∞ satisfies this condition, it follows from [27, Definition 4.7 and Theorem 4.5] that $\lim_{n\to+\infty}[T^n f](x) = \langle \pi_\infty, f\rangle$ for all f in $\mathscr{L}(\mathscr{X})$ and x in $\mathscr{X}$.

Fix now any f in $\mathscr{L}(\mathscr{X})$ and consider the sequence

$$\{\min(I+\delta Q)^n f\}_{n\in\mathbb{N}_0} = \{\min T^n f\}_{n\in\mathbb{N}_0}.$$

From the previous, we know that this sequence converges to $\langle \pi_\infty, f\rangle$ in the limit for $n \to +\infty$. Since T is a transition matrix—that is, has non-negative elements and rows that sum up to one—$\min g \leq \min Tg$ for all g in $\mathscr{L}(\mathscr{X})$. It now follows from repeated application of this inequality that the sequence $\{\min T^n f\}_{n\in\mathbb{N}_0}$ is non-decreasing. A similar argument shows that the sequence $\{\max T^n f\}_{n\in\mathbb{N}_0}$ is non-increasing and converges to $\langle \pi_\infty, f\rangle$ as well. □

Theorem 12. *Consider an ergodic transition rate matrix Q and a lumping map $\Lambda \colon \mathscr{X} \to \hat{\mathscr{X}}$. Then for all f in $\mathscr{L}(\mathscr{X})$, δ in $\mathbb{R}_{>0}$ with $\delta\|Q\| < 2$, and n in $\mathbb{N}_0$,*

$$\min(I+\delta\underline{\hat{Q}})^n \hat{f}_{\mathrm{L}} \leq \langle \pi_\infty, f\rangle \leq \max(I+\delta\hat{\overline{Q}})^n \hat{f}_{\mathrm{U}}.$$

Moreover, for fixed δ, the lower and upper bounds in this expression become monotonously tighter with increasing n, and each converges to a—possibly different—limit as n approaches $+\infty$. If Q is furthermore irreducible, $(I+\delta\underline{\hat{Q}})^n \hat{f}_{\mathrm{L}}$ and $(I+\delta\hat{\overline{Q}})^n \hat{f}_{\mathrm{U}}$ both converge to a—possibly different—constant function as n approaches $+\infty$.

Proof. As we have seen multiple times before, it suffices to prove the stated for the lower bound, as the stated for the upper bound follows from applying the stated for the lower bound to $g := -f$, because $\langle \pi_\infty, g \rangle = -\langle \pi_\infty, f \rangle$ and $\hat{g}_{\mathrm{L}} = -\hat{f}_{\mathrm{U}}$.

We now set out to prove the statement for the lower bound. As the expectation operator E_{π_∞} defined by the limit distribution π_∞ is monotonous, Lemma 40 implies that

$$\langle \pi_\infty, f \rangle = E_{\pi_\infty}(f) \geq E_{\pi_\infty}(\hat{f}_{\mathrm{L}} \circ \Lambda) = \langle \pi_\infty, \hat{f}_{\mathrm{L}} \circ \Lambda \rangle. \tag{F.4}$$

Observe now that

$$\langle \pi_\infty, f \rangle \geq \langle \pi_\infty, \hat{f}_{\mathrm{L}} \circ \Lambda \rangle \geq \min (I + \delta Q)^n (\hat{f}_{\mathrm{L}} \circ \Lambda) \geq \min [(I + \delta \underline{\hat{Q}})^n \hat{f}_{\mathrm{L}} \circ \Lambda] = \min (I + \delta \underline{\hat{Q}})^n \hat{f}_{\mathrm{L}},$$

where the second inequality follows from Proposition 11 and the third inequality follows from Lemma 37.

We end this proof by verifying the statement concerning the monotonous convergence of the lower bound. To that end, we first prove that

$$\|\underline{\hat{Q}}\| \leq \|Q\|. \tag{F.5}$$

By [25, Proposition 4],

$$\|\underline{\hat{Q}}\| = 2 \max \{ -[\underline{\hat{Q}} \mathbb{I}_{\hat{x}}](\hat{x}) \colon \hat{x} \in \hat{\mathscr{X}} \}.$$

We now use some properties of the transition rate matrix Q and execute some straightforward manipulations, to yield Eqn. (F.5):

$$\begin{aligned}
\|\underline{\hat{Q}}\| &= 2 \max \left\{ -\min \left\{ \sum_{\hat{y} \in \hat{\mathscr{X}}} \mathbb{I}_{\hat{x}}(\hat{y}) \sum_{y \in \hat{y}} Q(x, y) \colon x \in \hat{x} \right\} \colon \hat{x} \in \hat{\mathscr{X}} \right\} \\
&= 2 \max \left\{ -\min \left\{ \sum_{y \in \hat{x}} Q(x, y) \colon x \in \hat{x} \right\} \colon \hat{x} \in \hat{\mathscr{X}} \right\} \\
&= 2 \max \left\{ \max \left\{ -\sum_{y \in \hat{x}} Q(x, y) \colon x \in \hat{x} \right\} \colon \hat{x} \in \hat{\mathscr{X}} \right\} \\
&\leq 2 \max \Big\{ \max \{ -Q(x, x) \colon x \in \hat{x} \} \colon \hat{x} \in \hat{\mathscr{X}} \Big\} = 2 \max \{ -Q(x, x) \colon x \in \mathscr{X} \} = \|Q\|.
\end{aligned}$$

Since $\delta \|Q\| < 2$, it follows from Eqn. (F.5) that $\delta \|\underline{\hat{Q}}\| < 2$. Therefore, $(I + \delta \underline{\hat{Q}})$ is a lower transition operator by Lemma 26. Recall from (LT5) that $\min \hat{g} \leq \min (I + \delta \underline{\hat{Q}}) \leq \max (I + \delta \underline{\hat{Q}}) \hat{g} \leq \max \hat{g}$ for all $\hat{g}$ in $\mathscr{L}(\hat{\mathscr{X}})$. By repeatedly applying these inequalities, we obtain for all n in $\mathbb{N}_0$ that

$$\min \hat{f}_{\mathrm{L}} \leq \min (I + \delta \underline{\hat{Q}})^n \hat{f}_{\mathrm{L}} \leq \min (I + \delta \underline{\hat{Q}})^{n+1} \hat{f}_{\mathrm{L}} \leq \max (I + \delta \underline{\hat{Q}})^{n+1} \hat{f}_{\mathrm{L}} \leq \max (I + \delta \underline{\hat{Q}})^n \hat{f}_{\mathrm{L}} \leq \max \hat{f}_{\mathrm{L}}.$$

From this, we infer that $\{\min (I + \delta \underline{\hat{Q}})^n \hat{f}_{\mathrm{L}}\}_{n \in \mathbb{N}_0}$ is a bounded, non-decreasing sequence of real numbers, and therefore this sequence converges to some real number by the Monotonous Convergence Theorem. If Q is furthermore irreducible, then it follows immediately from Lemma 44 and Corollary 43 that $(I + \delta \underline{\hat{Q}})^n \hat{f}_{\mathrm{L}}$ converges to a constant function as n approaches $+\infty$. □

Theorem 13. *Consider an ergodic transition rate matrix Q and a lumping map $\Lambda \colon \mathscr{X} \to \hat{\mathscr{X}}$. Then for all $\hat{\mathcal{A}} \subseteq \mathcal{P}(\hat{\mathscr{X}})$[6] and f in $\mathscr{L}(\mathscr{X})$,*

$$\min \{ \langle \hat{\pi}, \hat{f}_{\mathrm{L}} \rangle \colon \hat{\pi} \in \mathscr{D}_{\hat{\mathcal{A}}} \} \leq \langle \pi_\infty, f \rangle \leq \max \{ \langle \hat{\pi}, \hat{f}_U \rangle \colon \hat{\pi} \in \mathscr{D}_{\hat{\mathcal{A}}} \}$$

with

$$\mathscr{D}_{\hat{\mathcal{A}}} := \{ \hat{\pi} \in \mathscr{D}(\hat{\mathscr{X}}) \colon (\forall \hat{A} \in \hat{\mathcal{A}}) \ \langle \hat{\pi}, \underline{\hat{Q}} \mathbb{I}_{\hat{A}} \rangle \leq 0 \}.$$

[6] Here and in the remainder, we denote the power set of the set S by $\mathcal{P}(S)$.

Proof. Recall from the beginning of the proof of Theorem 12 that we only need to prove the left inequality of the stated. Furthermore, recall from Eqn. (F.4) that

$$\langle \pi_\infty, f \rangle \geq \langle \pi_\infty, \hat{f}_{\mathrm{L}} \circ \Lambda \rangle.$$

Note that

$$\langle \pi_\infty, \hat{f}_{\mathrm{L}} \circ \Lambda \rangle = \sum_{x \in \mathscr{X}} \hat{f}_{\mathrm{L}}(\Lambda(x)) \pi_\infty(x) = \sum_{\hat{x} \in \hat{\mathscr{X}}} \hat{f}_{\mathrm{L}}(\hat{x}) \sum_{x \in \hat{x}} \pi_\infty(x) = \sum_{\hat{x} \in \hat{\mathscr{X}}} \hat{f}_{\mathrm{L}}(\hat{x}) \hat{\pi}_\infty(\hat{x}) = \langle \hat{\pi}_\infty, \hat{f}_{\mathrm{L}} \rangle,$$

where $\hat{\pi}_\infty \colon \hat{\mathscr{X}} \to \mathbb{R} \colon \hat{x} \mapsto \sum_{x \in \hat{x}} \pi_\infty(x)$. Combining the previously obtained inequality with the above equality yields

$$\langle \hat{\pi}_\infty, \hat{f}_{\mathrm{L}} \rangle \leq \langle \pi_\infty, f \rangle.$$

Hence, the stated follows if $\hat{\pi}_\infty$ is contained in $\hat{\mathscr{D}}_{\hat{\mathcal{A}}}$.

We now set out to prove this using the equilibrium condition of Eqn. (24). To that end, we observe that it suffices to verify that for all $\hat{A}$ in $\hat{\mathcal{A}}$,

$$\langle \hat{\pi}_\infty, \underline{\hat{Q}} \mathbb{I}_{\hat{A}} \rangle \leq 0.$$

Therefore, we fix an arbitrary $\hat{A}$ in $\hat{\mathcal{A}}$. By repeatedly applying Eqn. (24), we obtain

$$0 = \sum_{\hat{x} \in \hat{A}} \sum_{x \in \hat{x}} \sum_{y \in \mathscr{X}} \pi_\infty(y) Q(y, x).$$

We now reorder the summations, to yield

$$\begin{aligned} 0 &= \sum_{y \in \mathscr{X}} \pi_\infty(y) \sum_{\hat{x} \in \hat{A}} \sum_{x \in \hat{x}} Q(y, x) = \sum_{y \in \mathscr{X}} \pi_\infty(y) \sum_{x \in \mathscr{X}} Q(y, x) [\mathbb{I}_{\hat{A}} \circ \Lambda](x) = \sum_{y \in \mathscr{X}} \pi_\infty(y) [Q(\mathbb{I}_{\hat{A}} \circ \Lambda)](y) \\ &\geq \sum_{y \in \mathscr{X}} \pi_\infty(y) [\underline{\hat{Q}} \mathbb{I}_{\hat{A}}](\Lambda(y)) = \sum_{\hat{y} \in \hat{\mathscr{X}}} [\underline{\hat{Q}} \mathbb{I}_{\hat{A}}](\hat{y}) \sum_{y \in \hat{y}} \pi_\infty(y) = \sum_{\hat{y} \in \hat{\mathscr{X}}} [\underline{\hat{Q}} \mathbb{I}_{\hat{A}}](\hat{y}) \hat{\pi}_\infty(\hat{y}) \\ &= \langle \hat{\pi}_\infty, \underline{\hat{Q}} \mathbb{I}_{\hat{A}} \rangle, \end{aligned}$$

where the inequality follows from Lemma 30. □