Computing Inferences for Large-Scale Continuous-Time Markov Chains by Combining Lumping with Imprecision

Alexander Erreygers and Jasper De Bock

Ghent University, ELIS, SYSTeMS
`{alexander.erreygers, jasper.debock}@ugent.be`

Abstract If the state space of a homogeneous continuous-time Markov chain is too large, making inferences—here limited to determining marginal or limit expectations—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.

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 (CTMCs), this exponentially increasing number of states has as a consequence that using the model to perform inferences—for the sake of brevity here limited to marginal and limit expectations—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, allowing for a reduced state space with considerably fewer states. However, unfortunately, the low-level description and its corresponding larger state space are necessary in order to accurately model the system's dynamics. Therefore, using the reduced state space to make inferences is generally impossible.

In this contribution, we address this problem using imprecise continuous-time Markov chains [5, 11, 15]. In particular, we outline an approach to determine guaranteed lower and upper bounds on marginal and limit expectations using the reduced state space. We introduced a preliminary version of this approach in [8, 14], but the current contribution is—to the best of our knowledge—its first fully general and theoretically justified exposition. Compared to other approaches [3, 9] that also determine lower and upper bounds on expectations,

ours has the advantage that it is not restricted to limit expectations. Furthermore, based on our preliminary experiments, our approach seems to produce tighter bounds.

2 Continuous-Time Markov Chains

We are interested in making inferences about a system, more specifically about the state of this system at some future time t, denoted by X_t. The complication is that we are unable to predict the temporal evolution of the state with certainty. Therefore, at all times $t \in \mathbb{R}_{\geq 0}$,[1] the state X_t of the system is a random variable that takes values—generically denoted by x, y or z—in the state space $\mathcal{X}$.

2.1 Homogeneous Continuous-Time Markov Chains

We assume that the stochastic process that models our beliefs about the system, denoted by $(X_t)_{t \in \mathbb{R}_{\geq 0}}$, is a *continuous-time Markov chain* (CTMC) that is *homogeneous*. For a thorough treatment of the terminology and notation concerning CTMCs, we refer to [1, 11, 13]. Due to length constraints, we here limit ourselves to the bare necessities.

The stochastic process $(X_t)_{t \in \mathbb{R}_{\geq 0}}$ is a CTMC if it satisfies the *Markov property*, which says that for all $t_1, \ldots, t_n, t, \Delta$ in $\mathbb{R}_{\geq 0}$ with $n \in \mathbb{N}$ and $t_1 < \cdots < t_n < t$, and all $x_1, \ldots, x_n, x, y$ in $\mathcal{X}$,

$$P(X_{t+\Delta} = y \mid X_{t_1} = x_1 \ldots, X_{t_n} = x_n, X_t = x) = P(X_{t+\Delta} = y \mid X_t = x). \quad (1)$$

The CTMC $(X_t)_{t \in \mathbb{R}_{\geq 0}}$ is *homogeneous* if for all t, Δ in $\mathbb{R}_{\geq 0}$ and all x, y in $\mathcal{X}$,

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

It is well-known that—both in the classical measure-theoretic framework [1] and the full conditional framework [11]—a homogeneous continuous-time Markov chain is uniquely characterised by a triplet $(\mathcal{X}, \pi_0, Q)$, where $\mathcal{X}$ is a state space, π_0 an initial distribution and Q a transition rate matrix.

The state space $\mathcal{X}$ is taken to be a non-empty, finite and—without loss of generality—ordered set. This way, any real-valued function f on $\mathcal{X}$ can be identified with a column vector, the x-component of which is $f(x)$. The set containing all real-valued functions on $\mathcal{X}$ is denoted by $\mathcal{L}(\mathcal{X})$.

The initial distribution π_0 is defined by

$$\pi_0(x) := P(X_0 = x) \text{ for all } x \text{ in } \mathcal{X}, \quad (3)$$

and hence is a probability mass function on $\mathcal{X}$. We will (almost) exclusively be concerned with positive (initial) distributions, whom we collect in $\mathcal{D}(\mathcal{X})$ and will identify with row vectors.

---

[1] We use $\mathbb{R}_{\geq 0}$ and $\mathbb{R}_{>0}$ to denote the set of 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.

The transition rate matrix Q is a real-valued $|\mathcal{X}| \times |\mathcal{X}|$ matrix—or equivalently, a linear map from $\mathcal{L}(\mathcal{X})$ to $\mathcal{L}(\mathcal{X})$—with non-negative off-diagonal entries and rows that sum up to zero. If for any t in $\mathbb{R}_{\geq 0}$ we define the *transition matrix over t* as

$$T_t := e^{tQ} = \lim_{n \to +\infty} \left(I + \frac{t}{n} Q \right)^n, \tag{4}$$

then for all t in $\mathbb{R}_{\geq 0}$ and all x, y in $\mathcal{X}$,

$$P(X_t = y \mid X_0 = x) = T_t(x, y). \tag{5}$$

Finally, we denote by E the expectation operator with respect to the homogeneous CTMC $(X_t)_{t \in \mathbb{R}_{\geq 0}}$ in the usual sense. It follows immediately from (3) and (5) that $E(f(X_t)) = \pi_0 T_t f$ for any f in $\mathcal{L}(\mathcal{X})$ and any t in $\mathbb{R}_{\geq 0}$.

2.2 Irreducibility

In order not to be tangled up in edge cases, in the remainder we are only concerned with irreducible transition rate matrices. Many equivalent necessary and sufficient conditions exist; see for instance [13, Theorem 3.2.1]. For the sake of brevity, we here say that a transition rate matrix Q is *irreducible* if, for all t in $\mathbb{R}_{>0}$ and x, y in $\mathcal{X}$, $T_t(x, y) > 0$.

Consider now a homogeneous CTMC that is characterised by $(\mathcal{X}, \pi_0, Q)$. It is then well-known that for any f in $\mathcal{L}(\mathcal{X})$, the limit $\lim_{t \to +\infty} E(f(X_t))$ exists. Even more, since we assume that Q is irreducible, this limit value is the same for all initial distributions π_0 [13, Theorem 3.6.2]! This common limit value, denoted by $E_\infty(f)$, is called the *limit expectation of f*. Furthermore, the irreducibility of Q also implies that there is a unique stationary distribution π_∞ in $\mathcal{D}(\mathcal{X})$ that satisfies the *equilibrium condition* $\pi_\infty Q = 0$. This unique distribution is called the *limit distribution*, as $E_\infty(f) = \pi_\infty f$.

In the remainder of this contribution, a *positive and irreducible CTMC* is any homogeneous CTMC characterised by a positive initial distribution π_0 and an irreducible transition rate matrix Q.

3 Lumping and the Induced (Imprecise) Process

In many practical applications—see for instance [3, 8, 9, 14]—we have a positive and irreducible CTMC that models our system and we want to use this chain to make inferences of the form $E(f(X_t)) = \pi_0 T_t f$ or $E_\infty(f)$. As analytically evaluating the limit in (4) is often infeasible, we usually have to resort to one of the many available numerical methods—see for example [12]—that approximate T_t. However, unfortunately these numerical methods turn out to be computationally intractable when the state space becomes large. Similarly, determining the unique distribution π_∞ that satisfies the equilibrium condition also becomes intractable for large state spaces.

Fortunately, as previously mentioned in Sect. 1, the state space $\mathcal{X}$ 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 $\hat{\mathcal{X}}$. However, this provides no immediate solution as the motive behind using the detailed state space $\mathcal{X}$ in the first place is that this allows us to accurately model the (uncertain) dynamics of the system using a homogeneous CTMC; see [3, 8–10, 14] for practical examples. In contrast, the dynamics of the induced stochastic process on the the reduced state space $\hat{\mathcal{X}}$ are often unknown and/or intractable to obtain, which inhibits us from making exact inferences using the induced stochastic process. We now set out to address this by allowing for imprecision.

3.1 Notation and Terminology Concerning Lumping

We assume that the lumped state space $\hat{\mathcal{X}}$ is obtained by *lumping*—sometimes called grouping or aggregating, see [2, 4]—states in $\mathcal{X}$, such that $1 < |\hat{\mathcal{X}}| \le |\mathcal{X}|$. This lumping is formalised by the surjective *lumping map* $\Lambda\colon \mathcal{X} \to \hat{\mathcal{X}}$, which maps every state x in $\mathcal{X}$ to a state $\Lambda(x) = \hat{x}$ in $\hat{\mathcal{X}}$. In the remainder, we also use the inverse lumping map Γ, which maps every $\hat{x}$ in $\hat{\mathcal{X}}$ to a subset $\Gamma(\hat{x}) \coloneqq \{x \in \mathcal{X}\colon \Lambda(x) = \hat{x}\}$ of $\mathcal{X}$. Given such a lumping map Λ, a function f in $\mathcal{L}(\mathcal{X})$ is *lumpable with respect to* Λ if there is an $\hat{f}$ in $\mathcal{L}(\hat{\mathcal{X}})$ such that $f(x) = \hat{f}(\Lambda(x))$ for all x in $\mathcal{X}$. We use $\mathcal{L}_\Lambda(\mathcal{X}) \subseteq \mathcal{L}(\mathcal{X})$ to denote the set of all real-valued functions on $\mathcal{X}$ that are lumpable with respect to Λ.

As far as our results are concerned, it does not matter in which way the states are lumped. For a given f in $\mathcal{L}(\mathcal{X})$—recall that we are interested in the (limit) expectation of $f(X_t)$—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 $\hat{\mathcal{X}}$ correspond to a natural higher-level description of the state of the system; see for example [3, 8, 9] for some positive results. An extra benefit of this approach is that the resulting model can be used to determine the (limit) expectation of multiple functions.

3.2 The Lumped Stochastic Process

Let $(X_t)_{t \in \mathbb{R}_{\geq 0}}$ be a positive and homogeneous continuous-time Markov chain. Then any lumping map $\Lambda\colon \mathcal{X} \to \hat{\mathcal{X}}$ unequivocally induces a *lumped stochastic process* $(\hat{X}_t)_{t \in \mathbb{R}_{\geq 0}}$. It has $\hat{\mathcal{X}}$ as state space and is defined by the relation

$$(\hat{X}_t = \hat{x}) \Leftrightarrow (X_t \in \Gamma(\hat{x})) \text{ for all } t \text{ in } \mathbb{R}_{\geq 0} \text{ and all } \hat{x} \text{ in } \hat{\mathcal{X}}. \tag{6}$$

In some cases, this lumped stochastic process is a homogeneous CTMC, and the inference of interest can then be computed using this reduced CTMC. See for example [2, Theorem 2.3(i)] for a necessary condition and [2, Theorem 2.4] or [4, Theorem 3] for a necessary and sufficient one. However, these conditions are very stringent. Indeed, in general, the lumped stochastic process is not

homogeneous nor Markov. For this general case, we are not aware of any previous work that characterises the dynamics of the lumped stochastic process efficiently— i.e., directly from Λ, Q and π_0 and without ever determining T_t.

3.3 The Induced Imprecise Continuous-Time Markov Chain

Nevertheless, that is exactly what we now set out to do. Due to length constraints, we will here restrict ourselves to providing an intuitive explanation of our methodology, becoming formal only when stating our main results; see Theorems 1 and 2 further on. For a detailed exposition, we refer to the appendix of the extended preprint of this contribution [7].

The essential point is that, while we cannot exactly determine the dynamics of the lumped stochastic process $(\hat{X}_t)_{t \in \mathbb{R}_{\geq 0}}$, we can consider a *set of possible stochastic processes*, not necessarily homogeneous and/or Markovian but all with $\hat{\mathcal{X}}$ as state space, that definitely contains the lumped stochastic process $(\hat{X}_t)_{t \in \mathbb{R}_{\geq 0}}$. In the remainder, we will denote this set by $\mathbb{P}_{\pi_0, Q, \Lambda}$. As is indicated by our notation, $\mathbb{P}_{\pi_0, Q, \Lambda}$ is fully characterised by π_0, Q and Λ.

Crucially, it turns out that $\mathbb{P}_{\pi_0, Q, \Lambda}$ takes the form of a so-called *imprecise continuous-time Markov chain*. For a formal definition of general imprecise CTMCs, and an extensive study of their properties, we refer the reader to the work of Krak et. al. [11] and De Bock [5]. For our present purposes, it suffices to know that tight lower and upper bounds on the expectations that correspond to the set of stochastic processes of an imprecise CTMC 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.

There are many parallels between homogeneous CTMCs and imprecise CTMCs. For instance, the counterpart of a transition rate matrix is a *lower transition rate operator*. For our imprecise CTMC $\mathbb{P}_{\pi_0, Q, \Lambda}$, this lower transition rate operator is $\underline{\hat{Q}} \colon \mathcal{L}(\hat{\mathcal{X}}) \to \mathcal{L}(\hat{\mathcal{X}}) \colon g \mapsto \underline{\hat{Q}}g$ where, for every g in $\mathcal{L}(\hat{\mathcal{X}})$, $\underline{\hat{Q}}g$ is defined by

$$[\underline{\hat{Q}}g](\hat{x}) := \min\left\{ \sum_{\hat{y} \in \hat{\mathcal{X}}} g(\hat{y}) \sum_{y \in \Gamma(\hat{y})} Q(x, y) \colon x \in \Gamma(\hat{x}) \right\} \text{ for all } \hat{x} \text{ in } \hat{\mathcal{X}}. \quad (7)$$

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 (7) is fairly straightforward, as is for instance observed in [8, 14].

The counterpart of the transition matrix over t is now the *lower transition operator over t*, denoted by $\underline{\hat{T}}_t \colon \mathcal{L}(\hat{\mathcal{X}}) \to \mathcal{L}(\hat{\mathcal{X}})$ and defined for all g in $\mathcal{L}(\hat{\mathcal{X}})$ by

$$\underline{\hat{T}}_t g := \lim_{n \to +\infty} \left(I + \frac{t}{n} \underline{\hat{Q}} \right)^n g, \quad (8)$$

where the n-th power should be interpreted as consecutively applying the operator n times. Note how strikingly (8) resembles (4). Analogous to the precise case, one needs numerical methods—see for instance [6] or [11, Sect. 8.2]—to approximate $\underline{\hat{T}}_t g$ because analytically evaluating the limit in (8) is, at least in general, impossible.

4 Performing Inferences Using The Lumped Process

Everything is now set up to present our main results. Due to length constraints, we have relegated our proofs to the appendix of the extended arXiv version of this contribution [7].

4.1 Guaranteed Bounds On Marginal Expectations

We first turn to marginal expectations. Once we have $\mathbb{P}_{\pi_0,Q,\Lambda}$, the following result is a—not quite immediate—consequence of [11, Corollary 8.3].

Theorem 1. *Consider a positive and irreducible CTMC characterised by $(\mathcal{X}, \pi_0, Q)$ and a lumping map $\Lambda\colon \mathcal{X} \to \hat{\mathcal{X}}$. Let f in $\mathcal{L}(\mathcal{X})$ be lumpable with respect to Λ and let $\hat{f}$ be the corresponding element of $\mathcal{L}(\hat{\mathcal{X}})$. Then for any t in $\mathbb{R}_{\geq 0}$,*

$$\hat{\pi}_0 \hat{\underline{T}}_t \hat{f} \leq E(f(X_t)) = \pi_0 T_t f \leq -\hat{\pi}_0 \hat{\underline{T}}_t(-\hat{f}),$$

where $\hat{\pi}_0$ in $\mathcal{D}(\hat{\mathcal{X}})$ is defined by $\hat{\pi}_0(\hat{x}) := \sum_{x \in \Gamma(\hat{x})} \pi_0(x)$ for all $\hat{x}$ in $\hat{\mathcal{X}}$.

This result is highly useful in the setting that was outlined in Sect. 3. Indeed, for large systems we can use Theorem 1 to compute guaranteed lower and upper bounds on marginal expectations that cannot be computed exactly.

4.2 Guaranteed Bounds on Limit Expectations

Our second result provides guaranteed lower and upper bounds on limit expectations. This is extremely useful because the limit expectation is (almost surely) equal to the long-term temporal average due to the ergodic theorem [13, Theorem 3.8.1], and in practice—see for instance [8]—the inference one is interested in is often a long-term temporal average.

Theorem 2. *Consider an irreducible CTMC and a lumping map $\Lambda\colon \mathcal{X} \to \hat{\mathcal{X}}$. Let f in $\mathcal{L}(\mathcal{X})$ be lumpable with respect to Λ and let $\hat{f}$ be the corresponding element of $\mathcal{L}(\hat{\mathcal{X}})$. Then for all n in $\mathbb{N}_0$ and δ in $\mathbb{R}_{>0}$ such that $\delta \max\{|Q(x,x)|\colon x \in \mathcal{X}\} < 1$,*

$$\min(I + \delta\hat{\underline{Q}})^n \hat{f} \leq E_\infty(f) \leq -\min(I + \delta\hat{\underline{Q}})^n(-\hat{f}).$$

Furthermore, for fixed δ, the lower and upper bounds in this expression become monotonously tighter with increasing n, and each converges to a (possibly different) constant as n approaches $+\infty$.

This result can be used to devise an approximation method similar to [8, Algorithm 1]: we fix some value for δ, set $g_0 = \hat{f}$ (or $g_0 = -\hat{f}$) and then repeatedly compute $g_i := (I + \delta\hat{\underline{Q}})g_{i-1} = g_{i-1} + \delta\underline{Q}g_{i-1}$ until we empirically observe convergence of $\min g_i$ (or $-\min g_i$). In general, the lower and upper bounds obtained in this way are dependent on the choice of δ and this choice can therefore influence the tightness of the obtained bounds. Empirically, we have seen that smaller δ tend to yield tighter bounds, at the expense of requiring more iterations—that is, larger n—before empirical convergence.

4.3 Some Preliminary Numerical Results

Due to length constraints, we leave the numerical assessment of Theorem 1 for future work. For an extensive numerical assessment of—the method implied by—Theorem 2, we refer the reader to [8]. We believe that in this contribution, it is more fitting to compare our method to the only existing method—at least the only one that we are aware of—that also uses lumping to provide guaranteed lower and upper bounds on limit expectations. This method was first outlined by Franceschinis and Muntz [9], and then later improved by Buchholz [3]. In order to display the benefit of their methods, they use them to determine bounds on several performance measures for a closed queueing network that consists of a single server in series with multiple parallel servers. We use the method outlined in Sect. 4.2 to also compute bounds on these performance measures, as reported in Table 1. Note that our bounds are tighter than those of [9]. We would very much like to compare our method with the improved method of [3] as well. Unfortunately, the system parameters Buchholz uses do not—as far as we can tell—correspond to the number of states and the values for the performance measures he reports in [3, Fig. 3], thus preventing us from comparing our results.

Table 1. Comparison of the bounds obtained by using Theorem 2 with those obtained by the method presented in [9, Sect. 3.2] for the closed queueing network of [9].

	Exact	[9, Tab. 1]		Theorem 2	
		Lower	Upper	Lower	Upper
Mean queue length	1.2734	1.2507	1.3859	1.2664	1.2802
Throughput	0.9828	0.9676	0.9835	0.9826	0.9831

5 Conclusion

Broadly speaking, the conclusion of this contribution is that imprecise CTMCs 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 CTMCs. More concretely, the first important observation of this contribution is that lumping states in a homogeneous CTMC inevitably introduces imprecision, in the sense that we cannot exactly determine the parameters that describe the dynamics of the lumped stochastic process without also explicitly determining the original process. The second is that we can easily characterise a set of processes that definitely contains the lumped process, in the form of an imprecise CTMC. Using this imprecise CTMC, we can then determine guaranteed lower and upper bounds on marginal and limit expectations with respect to the original chain. From a practical point of view, these results are helpful 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 Sect. 4 is necessary. Furthermore, it would be of theoretical as well as practical interest to determine bounds on the *conditional* expectation of a lumpable function, or to consider functions that depend on the state at *multiple* time points. Finally, we are developing a method to determine lower and upper bounds on limit expectations that only requires the solution of a simple linear program.

Acknowledgements

References

1. Anderson, W.J.: Continuous-Time Markov Chains. Springer-Verlag (1991)
2. Ball, F., Yeo, G.F.: Lumpability and marginalisability for continuous-time Markov chains. Journal of Applied Probability **30**(3), 518–528 (1993)
3. Buchholz, P.: An improved method for bounding stationary measures of finite Markov processes. Performance Evaluation **62**(1), 349–365 (2005)
4. Burke, C.J., Rosenblatt, M.: A Markovian function of a Markov chain. The Annals of Mathematical Statistics **29**(4), 1112–1122 (1958)
5. De Bock, J.: The limit behaviour of imprecise continuous-time Markov chains. Journal of Nonlinear Science **27**(1), 159–196 (2017)
6. Erreygers, A., De Bock, J.: Imprecise continuous-time Markov chains: Efficient computational methods with guaranteed error bounds. In: Proceedings of ISIPTA'17. pp. 145–156. PMLR (2017), extended pre-print: 1702.07150math.PR
7. Erreygers, A., De Bock, J.: Computing inferences for large-scale continuous-time Markov chains by combining lumping with imprecision (2018), 1804.01020math.PR
8. Erreygers, A., Rottondi, C., Verticale, G., De Bock, J.: Imprecise Markov models for scalable and robust performance evaluation of flexi-grid spectrum allocation policies (2018), submitted, 1801.05700cs:NI
9. Franceschinis, G., Muntz, R.R.: Bounds for quasi-lumpable Markov chains. Performance Evaluation **20**(1), 223–243 (1994)
10. Ganguly, A., Petrov, T., Koeppl, H.: Markov chain aggregation and its applications to combinatorial reaction networks. Journal of Mathematical Biology **69**(3), 767–797 (2014)
11. Krak, T., De Bock, J., Siebes, A.: Imprecise continuous-time Markov chains. International Journal of Approximate Reasoning **88**, 452–528 (2017)
12. Moler, C., Van Loan, C.: Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. SIAM Review **45**(1), 3–49 (2003)
13. Norris, J.R.: Markov chains. Cambridge University Press (1997)
14. Rottondi, C., Erreygers, A., Verticale, G., De Bock, J.: Modelling spectrum assignment in a two-service flexi-grid optical link with imprecise continuous-time Markov chains. In: Proceedings of DRCN 2017. pp. 39–46. VDE Verlag (2017)
15. Škulj, D.: Efficient computation of the bounds of continuous time imprecise Markov chains. Applied Mathematics and Computation **250**, 165–180 (2015)