Computational Statistics
https://doi.org/10.1007/s00180-024-01454-9

ORIGINAL PAPER

Differentiated uniformization: a new method for inferring Markov chains on combinatorial state spaces including stochastic epidemic models

Kevin Rupp[1] · Rudolf Schill[1] · Jonas Süskind[1] · Peter Georg[2] · Maren Klever[3] · Andreas Lösch[1] · Lars Grasedyck[3] · Tilo Wettig[2] · Rainer Spang[1]

Received: 19 September 2022 / Accepted: 22 December 2023

Abstract

We consider continuous-time Markov chains that describe the stochastic evolution of a dynamical system by a transition-rate matrix Q which depends on a parameter θ. Computing the probability distribution over states at time t requires the matrix exponential $\exp(tQ)$, and inferring θ from data requires its derivative $\partial \exp(tQ)/\partial\theta$. Both are challenging to compute when the state space and hence the size of Q is huge. This can happen when the state space consists of all combinations of the values of several interacting discrete variables. Often it is even impossible to store Q. However, when Q can be written as a sum of tensor products, computing $\exp(tQ)$ becomes feasible by the uniformization method, which does not require explicit storage of Q. Here we provide an analogous algorithm for computing $\partial \exp(tQ)/\partial\theta$, the *differentiated uniformization method*. We demonstrate our algorithm for the stochastic SIR model of epidemic spread, for which we show that Q can be written as a sum of tensor products. We estimate monthly infection and recovery rates during the first wave of the COVID-19 pandemic in Austria and quantify their uncertainty in a full Bayesian analysis. Implementation and data are available at https://github.com/spang-lab/TenSIR.

Keywords Continuous-time Markov chains · Bayesian inference · Uniformization · Matrix exponential · Tensors · Epidemic spread

1 Introduction

Predicting the time evolution of complex dynamical systems has a wide range of applications in medicine and public health. One of them is the SIR model of epidemic spread, which describes how the numbers of susceptible (S), infected (I) and

K. Rupp, R. Schill and J. Süskind have contributed equally to this work.

Extended author information available on the last page of the article

Published online: 26 January 2024

recovered (R) people in a population change over the course of an epidemic by a system of differential equations. In this simple model the total population size stays constant, i.e., there are no births or deaths, there is no migration, and people get infected and recover at most once. The population has no demographic structure and no geographic structure, i.e., all individuals meet each other randomly. More complex models extend the SIR model in order to account for these limitations (Tang et al. 2020).

Until recently even the basic SIR model has been approximated deterministically (Kermack and McKendrick 1927) and was considered computationally intractable in its stochastic formulation (McKendrick 1925). The stochastic SIR model is a continuous-time Markov chain (CTMC) in which infections happen randomly with a rate proportional to S and proportional to I (Allen 2017). The state of the system at a given time is the tuple (S, I, R) which for a constant population size N is already specified by the tuple (S, I) since $R = N - S - I$. The state space is therefore $\{0, \dots, N\} \times \{0, \dots, N\}$ with size $(N + 1) \times (N + 1)$. Since infections happen randomly one must keep track of a huge number of probabilities, one for every possible state.[1]

More generally, we consider a CTMC which describes probability distributions $\mathbf{p}(t) \in \mathbb{R}^{|X|}$ over a discrete state space X, where an entry $\mathbf{p}(t)_x$ denotes the probability that the CTMC is in state $x \in X$ at time $t \in [0, \infty)$. Its change over time is governed by the Kolmogorov forward equation

$$\frac{\mathrm{d}\mathbf{p}(t)}{\mathrm{d}t} = Q\mathbf{p}(t) \tag{1}$$

with transition rate matrix $Q \in \mathbb{R}^{|X| \times |X|}$, where an off-diagonal entry $Q_{y,x}$ is the instantaneous transition rate from state $x \in X$ to state $y \in X$ and diagonal entries are set such that columns sum to zero. The solution to the Kolmogorov equation is given by the action of the matrix exponential,

$$\mathbf{p}(t) = \exp\,(tQ)\,\mathbf{p}(0) = \sum_{n=0}^{\infty} \frac{t^n}{n!} Q^n \mathbf{p}(0) \tag{2}$$

whose complexity is quadratic in $|X|$ when the sum is truncated after an appropriate number of terms. For example, an SIR model of the Austrian population with 9 million people has a state space X of size 9 million × 9 million = 81 trillion. The matrix Q has 81 trillion × 81 trillion entries and naively applying the matrix exponential on a vector is practically impossible, as well as numerically unstable.

Even more dauntingly, when Q depends on an unknown parameter θ, such as the infection or recovery rate in the SIR model, we must first infer θ from data. This can be done, for example, by maximizing its likelihood, which is an optimization problem that can be solved more efficiently if the derivative

[1] Technically half of these states (those with $S + I > N$) always have a probability of 0 which need not be tracked. We track them anyway for the mathematical convenience of writing the state space as a Cartesian product.

$$\frac{\partial \mathbf{p}(t)}{\partial \theta} = \frac{\partial \exp\,(tQ)}{\partial \theta}\,\mathbf{p}(0)$$

is available. Alternatively, θ can by inferred by sampling from its posterior in a full Bayesian analysis, which is also more efficient if the derivative of the likelihood is available.

However, Ho et al. (2018) have recently provided an algorithm that solves the Kolmogorov equation in the Laplace domain and evaluates the inverse Laplace transform numerically, thus avoiding the matrix exponential. Their algorithm is applicable to systems where each discrete variable increases monotonically. This includes the SIR model,[2] for which their algorithm scales quadratically in the population size.

In this paper, we provide an algorithm that directly computes $\exp\,(tQ)$ and, crucially, $\partial \exp\,(tQ)\,/\partial\theta$ at the same time. For the SIR model it scales cubically in the population size but is still practical. Importantly, our approach is applicable to a broader class of CTMCs with large state spaces that arise from interacting discrete variables, without requiring monotonicity. For example, in tumor progression models the states are combinations of possible mutations (Beerenwinkel and Sullivant (2009), Schill et al. (2019)), in stochastic neural networks the states are activation patterns of neurons (Yamanaka et al. 1997), in predator–prey dynamics they are joint population sizes of interacting species (Owen et al. 2014), or in chemical reaction networks they are joint counts of chemical species (Wolf 2007).

For many of these models Q can be written as a sum of tensor products (Buchholz 1999). We provide such a representation for the stochastic SIR model. To the best of our knowledge, this representation is novel. We use it for matrix–vector products that do not require explicit storage of Q (Buis and Dyksen 1996) and make computation of the matrix exponential tractable via the uniformization method (Grassmann 1977). A similar approach by Sherlock (2021) exploits the sparsity of Q. We extend the uniformization method and provide an analogous algorithm that also computes the derivative of the matrix exponential. Finally, we use Hamiltonian Monte Carlo sampling to provide a full Bayesian analysis of the first wave of the COVID-19 pandemic for the Austrian population, shedding new light on the uncertainties associated with the estimation of infection and recovery rates.

2 Differentiated uniformization for parameter estimation

The action of the matrix exponential

$$\mathbf{p}(t) = \exp\,(tQ)\,\mathbf{p}(0) = \sum_{n=0}^{\infty} \frac{t^n}{n!} Q^n \mathbf{p}(0) \tag{3}$$

could be approximated in principle by terminating after a finite number of terms. However, catastrophic cancellations occur (Moler and Van Loan 2003) due to the

[2] By changing variables from susceptibles and infected to infections and recoveries.

fact that Q has negative entries and negative eigenvalues.[3] The uniformization method (Grassmann 1977) addresses this problem by introducing a strictly nonnegative matrix

$$P := \frac{1}{\gamma}Q + \mathrm{I} \quad \text{for some } \gamma \geq \max_x |Q_{x,x}| \tag{4}$$

such that

$$\begin{aligned} \mathbf{p}(t) = \exp\,(tQ)\,\mathbf{p}(0) &= \exp\,(\gamma t(-\mathrm{I} + P))\,\mathbf{p}(0) \\ &= \exp\,(-\gamma t\,\mathrm{I}\,)\,\exp\,(\gamma t P)\,\mathbf{p}(0) \\ &= \sum_{n=0}^{\infty} e^{-\gamma t}\frac{(\gamma t)^n}{n!}P^n\mathbf{p}(0) \end{aligned} \tag{5}$$

does not suffer from cancellations. P can be viewed as the transition probability matrix of a discrete-time Markov chain where the number of transitions is a Poisson-distributed random variable with mean γt.

Using the recursions

$$P^n = PP^{n-1}, \tag{6}$$

$$\frac{(\gamma t)^n}{n!} = \frac{\gamma t}{n}\frac{(\gamma t)^{n-1}}{(n-1)!}, \tag{7}$$

$\mathbf{p}(t)$ can be computed according to Eq. (5) by algorithm 1 (Grassmann 1977). Note that $P^n\mathbf{p}(0)$ sums to 1 and hence Eq. (5) sums to less than 1 when terminated after a finite number of terms. The algorithm stops once this probability mass defect

$$1 - \sum_{n=0}^{m} e^{-\gamma t}\frac{(\gamma t)^n}{n!} \tag{8}$$

is smaller than a preset tolerance $\varepsilon > 0$. The required number m of iterations is in $\mathcal{O}(\gamma)$ (Reibman and Trivedi 1988) and can be determined, e.g., using the numerically robust method by Sherlock (2021).

In this paper we are interested in statistical models where Q depends on a parameter θ that we want to estimate from data by maximizing its likelihood or by sampling from its posterior. Both inference approaches benefit from utilizing gradient information. Al-Mohy and Higham (2009) proposed an efficient method to calculate derivatives of the matrix exponential for general matrices. Here, we propose a conceptually similiar algorithm specifically tailored towards transition rate matrices based on the uniformization method:

[3] Negative entries directly lead to cancellations in Eq. (3). If Eq. (3) is transformed to the eigenbasis of Q, negative eigenvalues of Q lead to cancellations as well.

$$\begin{aligned}\frac{\partial \mathbf{p}(t)}{\partial\theta} &= \frac{\partial \exp\,(tQ)}{\partial\theta}\,\mathbf{p}(0)\\ &= \frac{\partial}{\partial\theta}\left(\sum_{n=0}^{\infty} e^{-\gamma t}\frac{(\gamma t)^n}{n!}P^n\mathbf{p}(0)\right)\\ &= \sum_{n=0}^{\infty} e^{-\gamma t}\frac{(t\gamma)^n}{n!}\frac{\partial P^n}{\partial\theta}\mathbf{p}(0) + e^{-\gamma t}\frac{\partial\gamma}{\partial\theta}\left(-\frac{t^{n+1}\gamma^n}{n!}+\frac{t^n\gamma^{n-1}}{(n-1)!}\right)P^n\mathbf{p}(0)\\ &= \sum_{n=0}^{\infty} e^{-\gamma t}\frac{(t\gamma)^n}{n!}\left(\frac{\partial P^n}{\partial\theta}\mathbf{p}(0) + \frac{\partial\gamma}{\partial\theta}\left(\frac{n}{\gamma}-t\right)P^n\mathbf{p}(0)\right)\end{aligned}\tag{9}$$

We use the recursions (6), (7) and additionally

$$\begin{aligned}\frac{\partial P^n}{\partial\theta} &= \frac{\partial P}{\partial\theta}P^{n-1} + P\frac{\partial P}{\partial\theta}P^{n-2} + \ldots + P^{n-2}\frac{\partial P}{\partial\theta}P + P^{n-1}\frac{\partial P}{\partial\theta}\\ &= \frac{\partial P}{\partial\theta}P^{n-1} + P\left(\frac{\partial P}{\partial\theta}P^{n-2} + \ldots + P^{n-3}\frac{\partial P}{\partial\theta}P + P^{n-2}\frac{\partial P}{\partial\theta}\right)\\ &= \frac{\partial P}{\partial\theta}P^{n-1} + P\left(\frac{\partial P^{n-1}}{\partial\theta}\right)\end{aligned}\tag{10}$$

to compute $\mathbf{p}(t)' := \partial\mathbf{p}(t)/\partial\theta$ according to Eq. (9) by algorithm 2.

Algorithm 1 Uniformization

```
input  : $\mathbf{p}(0), t, P, \gamma, \varepsilon$
output: $\mathbf{p}(t)$
$n \leftarrow 0$
$w \leftarrow 1$
$\mathbf{p}(t) \leftarrow \mathbf{0}$
$\mathbf{q} \leftarrow \mathbf{p}(0)$
repeat
    $\mathbf{p}(t) \leftarrow \mathbf{p}(t) + e^{-\gamma t} w\mathbf{q}$
    $n \leftarrow n + 1$
    $\mathbf{q} \leftarrow P\mathbf{q}$
    $w \leftarrow \frac{\gamma t}{n} w$
until $1 - |\mathbf{p}(t)|_1 < \varepsilon$;
return $\mathbf{p}(t)$
```

Algorithm 2 Differentiated Uniformization

```
   input  : $\mathbf{p}(0), t, P, P', \gamma, \gamma', \varepsilon$
   output: $\mathbf{p}(t), \mathbf{p}(t)'$
 1 $n \leftarrow 0$
 2 $w \leftarrow 1$
 3 $\mathbf{p}(t) \leftarrow \mathbf{0}$
 4 $\mathbf{p}(t)' \leftarrow \mathbf{0}$
 5 $\mathbf{q} \leftarrow \mathbf{p}(0)$
 6 $\mathbf{q}' \leftarrow \mathbf{0}$
 7 repeat
 8     $\mathbf{p}(t) \leftarrow \mathbf{p}(t) + e^{-\gamma t} w \mathbf{q}$
 9     $\mathbf{p}(t)' \leftarrow \mathbf{p}(t)' + e^{-\gamma t} w \left( \mathbf{q}' + \gamma' \left( \frac{n}{\gamma} - t \right) \mathbf{q} \right)$
10     $n \leftarrow n + 1$
11     $\mathbf{q}' \leftarrow P' \mathbf{q} + P \mathbf{q}'$
12     $\mathbf{q} \leftarrow P \mathbf{q}$
13     $w \leftarrow \frac{\gamma t}{n} w$
14 until $1 - |\mathbf{p}(t)|_1 < \varepsilon$;
15 return $\mathbf{p}(t), \mathbf{p}(t)'$
```

Applying the differentiated uniformization for a particular statistical model requires the scalar

$$\gamma \geq \max_x |Q_{x,x}| \quad \text{and its derivative} \quad \gamma' := \frac{\partial \gamma}{\partial \theta}. \tag{11}$$

A generic choice for γ can be the 2-norm of the diagonal of Q or any p-norm with even p. It also requires the operators

$$P = \frac{1}{\gamma} Q + \mathrm{I} \quad \text{and} \quad P' := \frac{\partial P}{\partial \theta} = -\frac{1}{\gamma^2} \frac{\partial \gamma}{\partial \theta} Q + \frac{1}{\gamma} \frac{\partial Q}{\partial \theta}. \tag{12}$$

Crucially, these operators are only needed for matrix–vector products in lines 11 and 12 of algorithm 2 and do not need to be stored explicitly. This makes our method especially useful for models where Q is large but has a compact representation as a sum of tensor products, which allows one to cheaply compute matrix–vector products (Buis and Dyksen 1996).

Differentiated uniformization thus opens the door to **parameter inference** for CTMCs on huge discrete state spaces. Let $\{x_1, \ldots, x_K\}$ be observations of the Markov chain at corresponding time points $\{t_1, \ldots, t_K\}$. We represent each data point by an empirical probability distribution $\boldsymbol{\delta}(t_k) \in \mathbb{R}^{|X|}$, where $\boldsymbol{\delta}(t_k)_{x_k} = 1$ and all other entries are zero. The likelihood of θ for a single observation of state x_k at time t_k with $k > 1$ is

$$\mathbf{p}(t_k)_{x_k}, \text{ where } \mathbf{p}(t_k) = \exp\left((t_k - t_{k-1}) Q \right) \boldsymbol{\delta}(t_{k-1}). \tag{13}$$

The log-likelihood for the whole data set,

$$\ell(\theta) = \sum_{k=2}^{K} \log(\mathbf{p}(t_k)_{x_k}), \tag{14}$$

can be maximized using its derivative

$$\frac{\partial \ell(\theta)}{\partial \theta} = \sum_{k=2}^{K} \frac{\mathbf{p}(t_k)'_{x_k}}{\mathbf{p}(t_k)_{x_k}}, \tag{15}$$

for example by gradient ascent. This derivative can also be used for sampling a posterior distribution of θ in a full Bayesian model using a Hamiltonian Monte Carlo method (Gelman et al. 2013).

3 Modeling epidemic spread

The most basic models of epidemic spread are SIR models, which describe the numbers of susceptible (S), infected/infectious (I) and recovered (R) people during an epidemic in a closed population of constant size N.

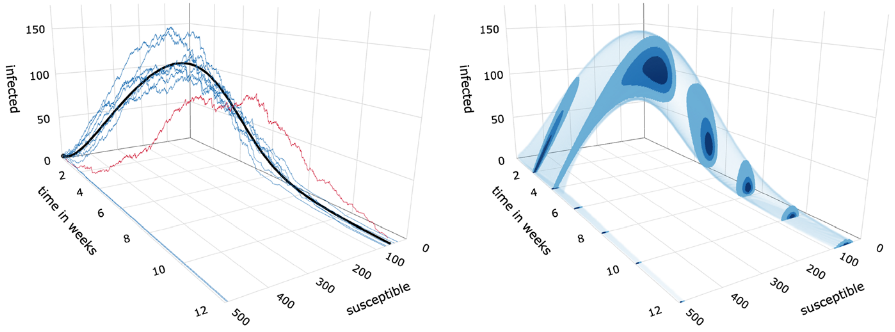
The **deterministic SIR model** (Kermack and McKendrick 1927) assumes that $S(t), I(t), R(t) \in [0, N]$ are continuous and describes their evolution over time $t \in [0, \infty)$ by the following system of nonlinear ordinary differential equations:

$$\begin{aligned}
\frac{\mathrm{d}S(t)}{\mathrm{d}t} &= \overbrace{-\beta\frac{I(t)S(t)}{N}}^{\text{infections}} \quad \overbrace{}^{\text{recoveries}},\\
\frac{\mathrm{d}I(t)}{\mathrm{d}t} &= +\beta\frac{I(t)S(t)}{N} \quad - \alpha I(t),\\
\frac{\mathrm{d}R(t)}{\mathrm{d}t} &= \phantom{+\beta\frac{I(t)S(t)}{N}} \quad + \alpha I(t),
\end{aligned} \tag{16}$$

where $\alpha, \beta \in \mathbb{R}^+$ are parameters. Note that once $S(t)$ and $I(t)$ are given, $R(t) = N - S(t) - I(t)$ is already determined and can be omitted in further analysis.

In words, an infection occurs when a susceptible person comes in sufficiently close contact with an infected person, which happens proportionally to the number of susceptible and to the density of infected people in the population and proportionally to an infection rate β. This rate β encompasses, for example, disease characteristics, people's behavior, public policy and weather. An infected person recovers with rate α and can then no longer become susceptible or infected again. The basic reproduction number $\mathcal{R}_0 := \beta/\alpha$ is the number of people (in a fully susceptible population) that one infected person infects before recovering.

There is no analytical solution to system (16), but it can be solved numerically, for example by Euler's method:



(a) Solution of a deterministic SIR model (black curve) and 10 randomly sampled trajectories of the corresponding stochastic SIR model (blue and red). The trajectory highlighted in red deviates drastically from the deterministic solution.

(b) Analytic solution of the Kolmogorov equation for a stochastic SIR model. The time slices show distributions $\mathbf{p}(t)$ where the shades of blue show the smallest (not necessarily contiguous) areas that contain 30%, 60% and 90% of the probability mass at time t.

Fig. 1 Illustration of SIR models with $N = 500$, $\alpha = 1w^{-1}$, $\beta = 2.5w^{-1}$, $I(0) = 3$, $S(0) = 497$

$$
\begin{aligned}
S(t + \Delta t) &= S(t) - \beta \frac{S(t)I(t)}{N} \Delta t, \\
I(t + \Delta t) &= I(t) + \beta \frac{S(t)I(t)}{N} \Delta t - \alpha I(t) \Delta t.
\end{aligned}
\tag{17}
$$

The black curve in Fig. 1a illustrates this solution for given parameters $\alpha = 1w^{-1}$, $\beta = 2.5w^{-1}$ and initial conditions $N = 500$, $I(0) = 3$, $S(0) = 497$.

This deterministic model has several limitations. First, an epidemic is in fact not a deterministic dynamical system but a stochastic process that depends on the random behavior of people and random duration of each infection. A person does not recover after exactly one week, but only after one week on average. Especially if the very first infected people recover by chance before they come in contact with other people, the epidemic may not even take off (flat blue curves in Fig. 1a). Also, a person does not infect exactly 2.5 people per week, but only 2.5 people per week on average. Whether a person infected early on happens to come in close contact with someone else after one week or two weeks may shift the whole course of the epidemic (red curve in Fig. 1a). Hence stochastic fluctuations especially at the beginning of the epidemic can drastically alter the shape of the curve compared to its deterministic counterpart. Only by considering the uncertainty in the course of the epidemic can policy makers make informed decisions, e.g., for allocating limited hospital capacities over time.

Another limitation of the deterministic model is that without modeling the stochastics explicitly it is not possible to quantify the uncertainties of inferred parameters, which contributes to the uncertainties in the course of the epidemic.

These limitations are alleviated by the **stochastic SIR model** (McKendrick 1925; Allen 2017) which is a continuous-time Markov chain over all possible states of the population. A state is a pair of integers $(S, I) \in \{0, \dots, N\} \times \{0, \dots, N\}$ denoting the number of susceptible and infected people. Because of the very large number of possible states the stochastic SIR model is more challenging and less widely adopted than the deterministic model.

Let $\mathbf{p}(t) \in \mathbb{R}^{(N+1)^2}$ denote the probability distribution at time t over all states (S, I). That is, $\mathbf{p}(t)_{(S,I)}$ is the probability that at time t there are S susceptible and I infected people. Its time evolution is governed by the Kolmogorov forward equation

$$\frac{\mathrm{d}\mathbf{p}(t)}{\mathrm{d}t} = Q\mathbf{p}(t), \tag{18}$$

where the matrix $Q \in \mathbb{R}^{(N+1)^2 \times (N+1)^2}$ contains the transition rates from a state (S, I) to a state $(S + \Delta S, I + \Delta I)$:

$$Q_{(S+\Delta S, I+\Delta I),(S,I)} = \begin{cases} \beta\frac{SI}{N} & \text{if } \Delta S = -1, \Delta I = +1, \\ \alpha I & \text{if } \Delta S = 0, \Delta I = -1, \\ -\beta\frac{SI}{N} - \alpha I & \text{if } \Delta S = 0, \Delta I = 0, S \neq 0, I \neq N, \\ -\alpha I & \text{if } \Delta S = 0, \Delta I = 0, S = 0 \text{ or } I = N, \\ 0 & \text{otherwise}. \end{cases} \tag{19}$$

The blue and red curves in Fig. 1a depict 10 randomly sampled trajectories where transitions happen according to the rates in Eq. (19), generated by the Gillespie (1976) algorithm. Figure 1b shows the analytic solution to Eq. (18) and further illustrates that the stochasticity is not merely additive noise around the deterministic solution.

The parameters $\alpha, \beta \in \mathbb{R}^+$ can be inferred from data using differentiated uniformization. This requires multiple matrix–vector products with Q which is, however, too large to be stored explicitly, even for populations of only thousands of people. Hence, we propose a novel representation of Q that does not require explicit storage. To this end, we introduce band matrices of size $(N + 1) \times (N + 1)$:

$$\begin{aligned} \mathcal{S}^+_{\text{inf}} &= \text{superdiag}(1, \dots, N), & \mathcal{I}^+_{\text{inf}} &= \text{subdiag}(0, \dots, N-1), \\ \mathcal{S}^-_{\text{inf}} &= \text{diag}(0, \dots, N), & \mathcal{I}^-_{\text{inf}} &= \text{diag}(0, \dots, N-1, 0), \\ \mathcal{S}^+_{\text{rec}} &= \text{diag}(1, 1, \dots, 1) = \text{ I}, & \mathcal{I}^+_{\text{rec}} &= \text{superdiag}(1, \dots, N), \\ \mathcal{S}^-_{\text{rec}} &= \text{diag}(1, 1, \dots, 1) = \text{ I}, & \mathcal{I}^-_{\text{rec}} &= \text{diag}(0, \dots, N). \end{aligned} \tag{20}$$

This yields a representation of the transition-rate matrix

$$Q = \frac{\beta}{N}(\mathcal{S}^+_{\text{inf}} \otimes \mathcal{I}^+_{\text{inf}}) + \alpha(\mathcal{S}^+_{\text{rec}} \otimes \mathcal{I}^+_{\text{rec}}) - \frac{\beta}{N}(\mathcal{S}^-_{\text{inf}} \otimes \mathcal{I}^-_{\text{inf}}) - \alpha(\mathcal{S}^-_{\text{rec}} \otimes \mathcal{I}^-_{\text{rec}}) \tag{21}$$

as a sum of tensor products[4] (see Fig. 2 for an illustrated explanation). Note that Eq. (21) is not an approximation but an exact reformulation of Eq. (19). The benefit of this representation is that its storage complexity is $\mathcal{O}(N)$ rather than $\mathcal{O}(N^4)$ and that performing matrix–vector products has a complexity in only $\mathcal{O}(N^2)$ (Buis and Dyksen 1996) rather than $\mathcal{O}(N^4)$.

[4] For clarity, we did not factor out $\mathcal{S}^+_{\text{rec}} = \mathcal{S}^-_{\text{rec}} = \text{ I}$, which would allow for a representation with only three terms.

$$Q = \begin{array}{c|cccc|cccc|cccc|cccc|}
 & 00 & 01 & 02 & 03 & 10 & 11 & 12 & 13 & 20 & 21 & 22 & 23 & 30 & 31 & 32 & 33 \\
\hline
00 & & 1\alpha & & & & & & & & & & & & & & \\
01 & & -1\alpha & 2\alpha & & & & & & & & & & & & & \\
02 & & & -2\alpha & 3\alpha & & 1\cdot1\beta/N & & & & & & & & & & \\
03 & & & & -3\alpha & & & 1\cdot2\beta/N & & & & & & & & & \\
\hline
10 & & & & & & 1\alpha & & & & & & & & & & \\
11 & & & & & & \substack{-1\cdot1\beta/N\\-1\alpha} & 2\alpha & & & & & & & & & \\
12 & & & & & & & \substack{-1\cdot2\beta/N\\-2\alpha} & 3\alpha & & 2\cdot1\beta/N & & & & & & \\
13 & & & & & & & & -3\alpha & & & 2\cdot2\beta/N & & & & & \\
\hline
20 & & & & & & & & & & 1\alpha & & & & & & \\
21 & & & & & & & & & & \substack{-2\cdot1\beta/N\\-1\alpha} & 2\alpha & & & & & \\
22 & & & & & & & & & & & \substack{-2\cdot2\beta/N\\-2\alpha} & 3\alpha & & 3\cdot1\beta/N & & \\
23 & & & & & & & & & & & & -3\alpha & & & 3\cdot2\beta/N & \\
\hline
30 & & & & & & & & & & & & & & 1\alpha & & \\
31 & & & & & & & & & & & & & & \substack{-3\cdot1\beta/N\\-1\alpha} & 2\alpha & \\
32 & & & & & & & & & & & & & & & \substack{-3\cdot2\beta/N\\-2\alpha} & 3\alpha \\
33 & & & & & & & & & & & & & & & & -3\alpha \\
\hline
\end{array}$$

$$= \frac{\beta}{N}\underset{\mathcal{S}^{+}_{\text{inf}}}{\begin{array}{c|cccc|} & 0 & 1 & 2 & 3 \\ \hline 0 & & 1 & & \\ 1 & & & 2 & \\ 2 & & & & 3 \\ 3 & & & & \\ \hline \end{array}} \otimes \underset{\mathcal{I}^{+}_{\text{inf}}}{\begin{array}{c|cccc|} & 0 & 1 & 2 & 3 \\ \hline 0 & & & & \\ 1 & 0 & & & \\ 2 & & 1 & & \\ 3 & & & 2 & \\ \hline \end{array}} + \alpha \underset{\mathcal{S}^{+}_{\text{rec}}}{\begin{array}{c|cccc|} & 0 & 1 & 2 & 3 \\ \hline 0 & 1 & & & \\ 1 & & 1 & & \\ 2 & & & 1 & \\ 3 & & & & 1 \\ \hline \end{array}} \otimes \underset{\mathcal{I}^{+}_{\text{rec}}}{\begin{array}{c|cccc|} & 0 & 1 & 2 & 3 \\ \hline 0 & & 1 & & \\ 1 & & & 2 & \\ 2 & & & & 3 \\ 3 & & & & \\ \hline \end{array}}$$

$$- \frac{\beta}{N}\underset{\mathcal{S}^{-}_{\text{inf}}}{\begin{array}{c|cccc|} & 0 & 1 & 2 & 3 \\ \hline 0 & 0 & & & \\ 1 & & 1 & & \\ 2 & & & 2 & \\ 3 & & & & 3 \\ \hline \end{array}} \otimes \underset{\mathcal{I}^{-}_{\text{inf}}}{\begin{array}{c|cccc|} & 0 & 1 & 2 & 3 \\ \hline 0 & 0 & & & \\ 1 & & 1 & & \\ 2 & & & 2 & \\ 3 & & & & 0 \\ \hline \end{array}} - \alpha \underset{\mathcal{S}^{-}_{\text{rec}}}{\begin{array}{c|cccc|} & 0 & 1 & 2 & 3 \\ \hline 0 & 1 & & & \\ 1 & & 1 & & \\ 2 & & & 1 & \\ 3 & & & & 1 \\ \hline \end{array}} \otimes \underset{\mathcal{I}^{-}_{\text{rec}}}{\begin{array}{c|cccc|} & 0 & 1 & 2 & 3 \\ \hline 0 & 0 & & & \\ 1 & & 1 & & \\ 2 & & & 2 & \\ 3 & & & & 3 \\ \hline \end{array}}$$

Fig. 2 Illustration of Q for a population of size $N = 3$ given by its entry-wise representation in Eq. (19) (top) and its tensor representation in Eq. (21) (bottom). Blue numbers indicate susceptibles, red numbers indicate infected and blank entries in the matrices are zero. Transitions should be read from columns to rows. $\mathcal{S}^{+}_{\text{inf}}$: An infection decreases the number of susceptibles by one and happens proportionally to the current number of susceptibles. $\mathcal{I}^{+}_{\text{inf}}$: At the same time, an infection increases the number of infected by one and happens proportionally to the current number of infected. The tensor product $\otimes$ combines both these transitions for a single infection. Moreover, an infection happens inversely proportional to the total population size N and proportionally to the parameter β. $\mathcal{S}^{+}_{\text{rec}}$: A recovery does not change the number of susceptibles. $\mathcal{I}^{+}_{\text{rec}}$: At the same time, a recovery decreases the number of infected by one and happens proportionally to the current number of infected. The tensor product $\otimes$ combines both these transitions for a single recovery. Moreover, a recovery happens proportionally to the parameter α. The matrices $\mathcal{S}^{-}_{\text{inf}}$, $\mathcal{I}^{-}_{\text{inf}}$, $\mathcal{S}^{-}_{\text{rec}}$, $\mathcal{I}^{-}_{\text{rec}}$ generate corresponding negative entries for the diagonal of Q

Additionally, differentiated uniformization requires the derivative $\partial Q/\partial\theta$. Here we perform inference with respect to logarithmic parameters $\theta = (\log\alpha, \log\beta)$ in order to ensure the positivity constraint on α and β:

$$\frac{\partial Q}{\partial \log\alpha} = \alpha(\mathcal{S}^{+}_{\text{rec}} \otimes \mathcal{I}^{+}_{\text{rec}}) - \alpha(\mathcal{S}^{-}_{\text{rec}} \otimes \mathcal{I}^{-}_{\text{rec}}), \tag{22}$$

$$\frac{\partial Q}{\partial \log \beta} = \frac{\beta}{N}(\mathcal{S}^{+}_{\text{inf}} \otimes \mathcal{I}^{+}_{\text{inf}}) - \frac{\beta}{N}(\mathcal{S}^{-}_{\text{inf}} \otimes \mathcal{I}^{-}_{\text{inf}}). \tag{23}$$

Finally, differentiated uniformization requires a differentiable upper bound γ on the absolute diagonal entries of Q. For the SIR model we choose the exact maximum

$$\begin{aligned} \gamma = \max_x |Q_{x,x}| &= \max\left\{|Q_{(N-1,N-1),(N-1,N-1)}|, |Q_{(N,N),(N,N)}|\right\} \\ &= \max\left\{N(N-1)\frac{\beta}{N} + (N-1)\alpha, N\alpha\right\} \\ &= \max\{(N-1)\beta + (N-1)\alpha, \alpha + (N-1)\alpha\} \\ &= (N-1)\alpha + \max\{(N-1)\beta, \alpha\}. \end{aligned} \tag{24}$$

It is differentiable[5] for $\alpha \neq (N-1)\beta$ with

$$\frac{\partial \gamma}{\partial \log \alpha} = \begin{cases} N\alpha & \text{if } \alpha > (N-1)\beta, \\ (N-1)\alpha & \text{if } \alpha < (N-1)\beta, \end{cases} \tag{25}$$

$$\frac{\partial \gamma}{\partial \log \beta} = \begin{cases} 0 & \text{if } \alpha > (N-1)\beta, \\ (N-1)\beta & \text{if } \alpha < (N-1)\beta. \end{cases} \tag{26}$$

Overall, differentiated uniformization performs $\mathcal{O}(\gamma)$ matrix–vector products and thus has a total runtime complexity in $\mathcal{O}(\gamma N^2) = \mathcal{O}(N^3)$ for the SIR model. It requires storage of the result $\mathbf{p}(t)$, which has complexity $\mathcal{O}(N^2)$.

For parameter inference we are typically only interested in the likelihood that an earlier data point (S, I) is followed by a later data point $(S + \Delta S, I + \Delta I)$ after time t. Since the number of susceptibles cannot increase $(\Delta S \leq 0)$ and the number of recovered cannot decrease $(\Delta R = -\Delta S - \Delta I \geq 0)$ along a trajectory, it is sufficient to compute $\mathbf{p}(t)$ and $\mathbf{p}(t)'$ on the restricted state space

$$\{S + \Delta S, \dots, S\} \times \{I - \Delta R, \dots, I - \Delta S\},$$

as explained in Appendix A. Following Ho et al. (2018) we use this state-space restriction to reduce the time complexity of our algorithm to $\mathcal{O}\big((I + |\Delta S|)(\Delta S^2 + |\Delta S|\Delta R)\big)$ and its storage complexity to $\mathcal{O}(\Delta S^2 + |\Delta S|\Delta R)$.

4 COVID-19 pandemic

Here we model the first wave of the COVID-19 pandemic in Austria as a stochastic SIR model. We employ differentiated uniformization to estimate the parameters α and β and quantify their uncertainty. We use daily numbers on S, I and R between 2020-03-01 and 2020-09-01 from public health data provided by the Austrian Bundesministerium für Soziales (2021) (Fig. 4). I and R are given directly, and we set $S = N - I - R$ assuming that the initial population size $N = 8{,}932{,}664$ stays

[5] For $\alpha = (N-1)\beta$ a differentiable upper bound for $\max\{(N-1)\beta, \alpha\}$ is $\log(e^{(N-1)\beta} + e^{\alpha})$.

Table 1 Diagnostics of HMC sampling

Month	Runtime	ESS α	ESS β
March	3.30×10^6s	3096	5068
April	8.50×10^5s	5019	4830
May	3.00×10^4s	4988	5005
June	2.70×10^4s	4755	4956
July	1.00×10^5s	4862	4574
August	6.30×10^5s	4716	4972

constant. People who have died from COVID-19 are counted under "recovered" in a technical sense as they are no longer infectious. We do not correct for undiscovered cases and biases in testing and reporting. We also assume that parameters are piecewise constant for each month.

We do a full Bayesian analysis for parameter pairs $(\log \alpha, \log \beta)$ with a uniform prior for each parameter in the interval between $\log(0.01/\text{day})$ and $\log(1/\text{day})$. This highlights the shape of the likelihood of the model but can be substituted by any other prior informed by expert knowledge. Following Ho et al. (2018) we sample from the joint posterior using a Hamiltonian Monte Carlo (HMC) scheme (Duane et al. 1987; Neal 2011) as implemented in the software package PyMC (Oriol et al. 2023). Unlike a standard Metropolis-Hastings scheme, HMC makes use of the gradient of the likelihood, which we compute using differentiated uniformization. This makes sampling more efficient with less samples needed to cover the posterior distribution (Gelman et al. 2013).

We estimated the joint posterior of $(\log \alpha, \log \beta)$ for every month between March 2020 and August 2020 separately. For each month we performed 10 parallel Monte Carlo chains with length 1000, where we discarded the first 100 points each, resulting in 9000 points per month. These calculations were done on the QPACE 3 cluster (Georg et al. 2018). For each posterior we recorded the runtime (averaged over the 10 chains) and measured the marginal folded effective sample sizes (ESS) (Vehtari et al. 2021) for α and β, see Table 1.

Figure 5 shows the results of this analysis. The estimated posterior is plotted (α, β) on logarithmic scales. The gray shaded areas were generated using Gaussian-kernel density estimation applied to the posterior samples. The crosses mark the least-squares estimators of the corresponding deterministic SIR models. The dashed lines represent parameter constellations where $\alpha = \beta$ and thus $\mathcal{R}_0 = 1$. Here the epidemic switches between growing and decreasing numbers of infected. From April-August 2020 the posterior of the recovery rate α varies around a value of 0.07 per day, corresponding to the realistic mean time to recovery of about 2 weeks (Faes et al. 2020). In contrast, the posterior of α in March 2020 appears to be off, with a mean of about 0.03 per day corresponding to a mean time to recovery of one month. Inspecting the original numbers, we observed that the numbers of recovered are unexpectedly low (less than 100 people until 2020-03-23) possibly due to lagging declaration of recoveries because of cautious hospital policies in the beginning of the pandemic.

Table 2 Diagnostics of MH sampling

Month	Runtime	ESS α	ESS β
March	0.79×10^6s	159	31
April	2.30×10^5s	31	38
May	0.98×10^4s	55	73
June	0.88×10^4s	140	84
July	0.27×10^5s	114	75
August	1.40×10^5s	60	34

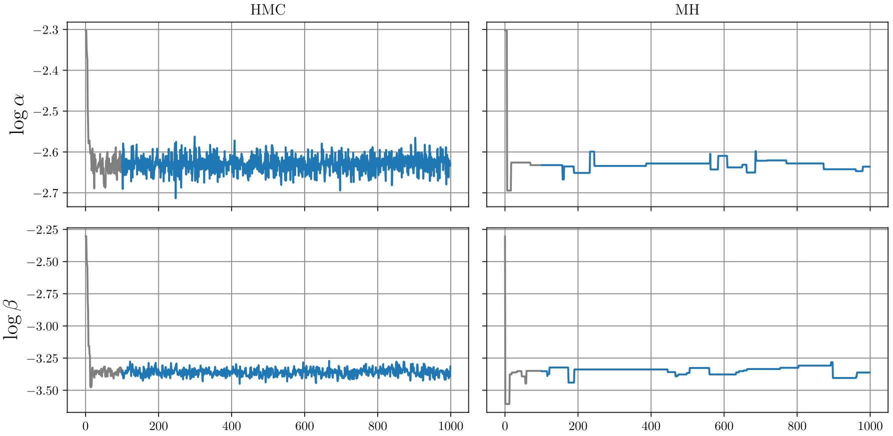


Fig. 3 Marginal trace plots of α and β for May in a single chain for the HMC sampling and MH sampling. The first 100 sample points are discarded as burn-in and shown in grey

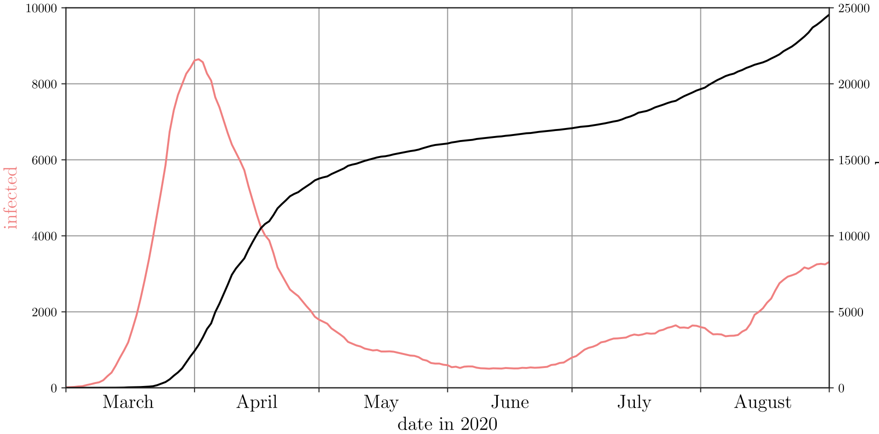


Fig. 4 Daily reported numbers of people infected by and recovered from SARS-CoV-2 in Austria

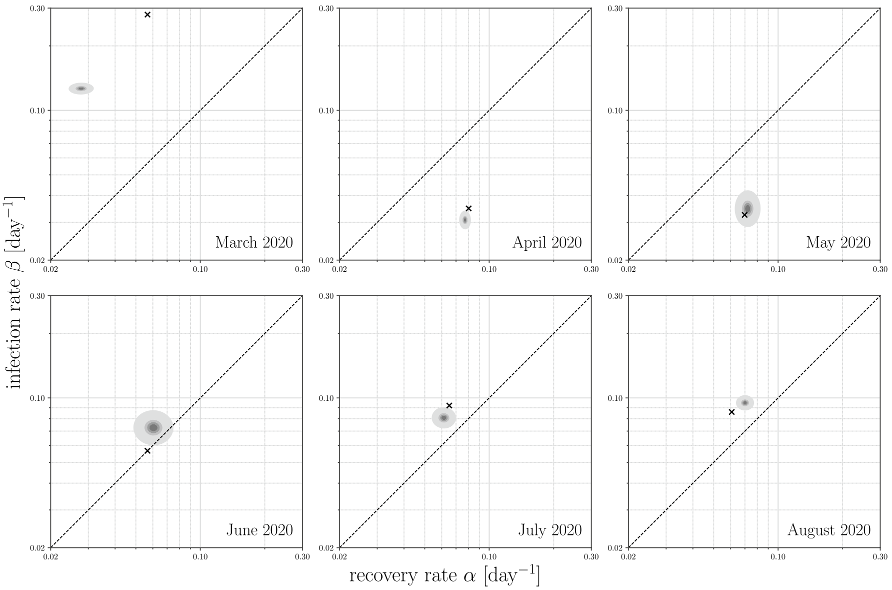


Fig. 5 Posterior probability densities over parameter pairs (α, β) for separate stochastic SIR models of the first six months of the COVID-19 pandemic in Austria. The dashed lines indicate parameters where the basic reproduction number $\mathcal{R}_0 = \beta/\alpha = 1$. The crosses mark the least-squares estimators of the corresponding deterministic SIR models

Finally, we compared the HMC sampling to a random walk Metropolis Hastings (MH) sampling, see Table 2. We evaluated both methods on the same hardware and used their respective implementations in PyMC with 100 tuning iterations for their hyperparameters. As expected, MH required a much lower runtime for a fixed total sample size. However, the effective sample sizes were substantially larger for HMC, such that HMC outperformed MH in terms of ESS per runtime. Figure 3 shows the marginal trace plots of α and β for May in a single chain for each of both samplings. Trace plots for the other months and autocorrelation plots are available in the supplement.

5 Discussion

We provide a novel method for computing the transient distribution and its derivative for continuous-time Markov Chains on huge discrete state spaces. This makes parameter inference tractable for a large family of statistical models, including the stochastic SIR model of epidemic spread.

Our key observation is that the transition-rate matrix of an SIR model can be written as a sum of tensor products, which allows us to cheaply compute matrix–vector products without storing the matrix itself. This operation alone is sufficient to compute the transient distribution by the uniformization method (Grassmann 1977), a numerically stable power-series expansion of the matrix exponential. We propose the *differentiated uniformization method*, an analogous power series for computing the derivative of the transient distribution with respect to parameters of a CTMC.

For the SIR model our algorithm scales cubically in the size of the population, which is one order slower than the state-of-the-art method for multivariate birth processes (Ho et al. 2018). On the other hand, our general-purpose algorithm also applies to birth-death processes such as predator–prey dynamics (Owen et al. 2014), which have been considered intractable so far (Ho et al. 2017). We illustrate this in Appendix B. In general our algorithm is applicable to any CTMC of interacting discrete variables. It scales exponentially in the number of variables but polynomially in the size of each variable's state space. These variables could be additional compartments in an epidemic model, such as the number of exposed but not yet infected people, asymptomatic carriers or deceased people.

Beyond epidemiology we are interested in tumor progression modeling using mutual hazard networks (Schill et al. 2019). Similar to an epidemiological model that scales exponentially in the number of compartments, a tumor progression model is a CTMC that scales exponentially in the number of possible mutations. While both differentiated uniformization and the algorithm of Ho et al. (2018) have the potential to advance this field, large scale inference remains an open problem for tumor progression models with up to hundreds of mutations. The tensor representation of the transition-rate matrix could serve as a starting point for representing the transient distribution itself in a low-rank tensor format. These formats reduce the exponential cost (e.g., in the number of mutations or compartments) to linear cost provided certain low-rank structures exist (Hackbusch 2012). For large-scale CTMCs, low-rank tensor formats were already successfully used, e.g., for the computation of transient (Johnson et al. 2010) and stationary distributions (Benson et al. 2017; Buchholz et al. 2016; Kressner and Macedo 2014) and also for a variant of the uniformization method (Georg et al. 2020). Therefore, the combination of low-rank tensor formats and differentiated uniformization could be a promising new avenue for large-scale inference problems in computational oncology and epidemiology.

From this perspective our work can also be seen as an attempt to connect these two communities.

Appendix A: State-space restriction

In order to compute the likelihood that an earlier data point (S, I) is followed by a later data point $(S + \Delta S, I + \Delta I)$ after time t, it is sufficient to compute $\mathbf{p}(t)$ on a small subset of the entire state space. Since the number of susceptibles cannot increase $(\Delta S \leq 0)$ and the number of recovered cannot decrease $(\Delta R = -\Delta S - \Delta I \geq 0)$, all possible trajectories from (S, I) to $(S + \Delta S, I + \Delta I)$ must necessarily stay within the restricted state space

$$\{S_{\min}, \dots, S_{\max}\} \times \{I_{\min}, \dots, I_{\max}\},$$

where

$$\begin{aligned} S_{\min} &= S + \Delta S, & S_{\max} &= S, \\ I_{\min} &= I - \Delta R, & I_{\max} &= I - \Delta S. \end{aligned} \tag{27}$$

All probability mass that leaves this space must be accounted for, but we do not need to keep track of its destination. To this end, we introduce the modified band matrices

$$\underbrace{\begin{aligned} \tilde{\mathcal{S}}^{+}_{\text{inf}} &= \operatorname{superdiag}(S_{\min} + 1, \dots, S_{\max}) \\ \tilde{\mathcal{S}}^{-}_{\text{inf}} &= \operatorname{diag}(S_{\min}, \dots, S_{\max}) \\ \tilde{\mathcal{S}}^{+}_{\text{rec}} &= \operatorname{diag}(1, 1, \dots, 1) = \mathrm{I} \\ \tilde{\mathcal{S}}^{-}_{\text{rec}} &= \operatorname{diag}(1, 1, \dots, 1) = \mathrm{I} \end{aligned}}_{(|\Delta S| + 1) \times (|\Delta S| + 1)} \qquad \underbrace{\begin{aligned} \tilde{\mathcal{I}}^{+}_{\text{inf}} &= \operatorname{subdiag}(I_{\min}, \dots, I_{\max} - 1) \\ \tilde{\mathcal{I}}^{-}_{\text{inf}} &= \operatorname{diag}(I_{\min}, \dots, I_{\max}) \\ \tilde{\mathcal{I}}^{+}_{\text{rec}} &= \operatorname{superdiag}(I_{\min} + 1, \dots, I_{\max}) \\ \tilde{\mathcal{I}}^{-}_{\text{rec}} &= \operatorname{diag}(I_{\min}, \dots, I_{\max}). \end{aligned}}_{(|\Delta S| + \Delta R + 1) \times (|\Delta S| + \Delta R + 1)} \tag{28}$$

and define a smaller transition-rate matrix on the restricted state space as

$$\tilde{Q} = \frac{\beta}{N}(\tilde{\mathcal{S}}^{+}_{\text{inf}} \otimes \tilde{\mathcal{I}}^{+}_{\text{inf}}) + \alpha(\tilde{\mathcal{S}}^{+}_{\text{rec}} \otimes \tilde{\mathcal{I}}^{+}_{\text{rec}}) - \frac{\beta}{N}(\tilde{\mathcal{S}}^{-}_{\text{inf}} \otimes \tilde{\mathcal{I}}^{-}_{\text{inf}}) - \alpha(\tilde{\mathcal{S}}^{-}_{\text{rec}} \otimes \tilde{\mathcal{I}}^{-}_{\text{rec}}) \tag{29}$$

with derivatives

$$\frac{\partial \tilde{Q}}{\partial \log \alpha} = \alpha(\tilde{\mathcal{S}}^{+}_{\text{rec}} \otimes \tilde{\mathcal{I}}^{+}_{\text{rec}}) - \alpha(\tilde{\mathcal{S}}^{-}_{\text{rec}} \otimes \tilde{\mathcal{I}}^{-}_{\text{rec}}), \tag{30}$$

$$\frac{\partial \tilde{Q}}{\partial \log \beta} = \frac{\beta}{N}(\tilde{\mathcal{S}}^{+}_{\text{inf}} \otimes \tilde{\mathcal{I}}^{+}_{\text{inf}}) - \frac{\beta}{N}(\tilde{\mathcal{S}}^{-}_{\text{inf}} \otimes \tilde{\mathcal{I}}^{-}_{\text{inf}}). \tag{31}$$

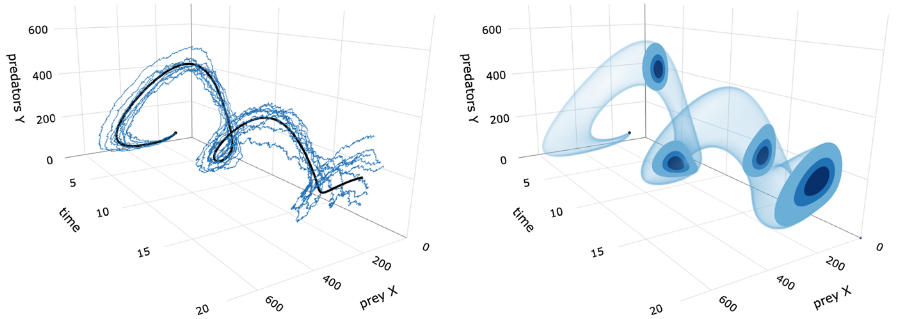
Note that the columns of $\tilde{Q}$ sum to less than zero and that $\mathbf{p}(t)$ therefore sums to less than 1 on the restricted state space. Computing matrix–vector products using these operators has a time complexity in $\mathcal{O}(|\Delta S|^2 + |\Delta S|\Delta R)$.

The largest absolute diagonal entry of $\tilde{Q}$ is

$$\gamma = \max_x |\tilde{Q}_{x,x}| = \frac{\beta}{N} S_{\max} I_{\max} + \alpha I_{\max} \tag{32}$$

with derivatives

$$\frac{\partial \gamma}{\partial \log \alpha} = \alpha I_{\max}, \tag{33}$$



(a) Solution of a deterministic predator-prey model (black curve) and 10 randomly sampled trajectories (blue) of the corresponding stochastic model.

(b) Analytic solution of the Kolmogorov equation for a stochastic predator-prey model. The time slices show distributions $\mathbf{p}(t)$ where the shades of blue show the smallest areas that contain 30%, 60% and 90% of the probability mass at time t.

Fig. 6 Illustration of predator–prey models with $\alpha = 1$, $\beta = 0.004$, $\delta = 0.8$, $X(0) = 100$, $Y(0) = 40$, $X_{\max} = Y_{\max} = 1200$

$$\frac{\partial \gamma}{\partial \log \beta} = \frac{\beta}{N} S_{\max} I_{\max}. \tag{34}$$

We perform m iterations of algorithm 2 such that the entire probability mass (including that which left the restricted state space) according to Eq. (8) reaches the required tolerance. Hence, the overall time complexity of the algorithm is

$$\mathcal{O}\big(\gamma(|\Delta S|^2 + |\Delta S|\Delta R)\big) = \mathcal{O}\big(I_{\max}(|\Delta S|^2 + |\Delta S|\Delta R)\big) = \mathcal{O}\big((I + |\Delta S|)(|\Delta S|^2 + |\Delta S|\Delta R)\big). \tag{35}$$

Storing the result $\mathbf{p}(t)$ has complexity $\mathcal{O}(|\Delta S|^2 + |\Delta S|\Delta R)$.

Appendix B: Predator–prey dynamics

We consider the following deterministic predator–prey equations, based on (Lotka 1925; Volterra 1926),

$$\frac{\mathrm{d}X(t)}{\mathrm{d}t} = -\beta X(t)Y(t) \overbrace{+\alpha X(t) - \alpha \frac{X(t)^2}{X_{\max}}}^{\text{prey birth}} \tag{36}$$

$$\frac{\mathrm{d}Y(t)}{\mathrm{d}t} = \underbrace{+\beta X(t)Y(t)}_{\substack{\text{prey consumption}\\ \text{\&predator birth}}} \quad \underbrace{-\delta Y(t)}_{\text{predator death}} \tag{37}$$

which describe how the population size $X(t) \in \mathbb{R}^+$ of a prey species and the population size $Y(t) \in \mathbb{R}^+$ of a predator species change continuously over time as the species interact (black curve in Fig. 6a). The parameter α is the birth rate of prey, δ is the death rate of predators and β is the contact rate between predators and prey. Upon contact a prey is consumed and, we assume for simplicity, exactly one predator is born. X_{max} is a finite carry capacity representing the available plant resources for the prey species, which would result in logistic growth in the absence of predators. This is neither necessary nor commonly assumed in the literature on the deterministic model, since the prey population is always limited by a nonzero number of predators in $\mathbb{R}^+$.

Here we are interested in a corresponding stochastic model (Owen et al. 2014) (blue curves in Fig. 6a) in which the number of predators is an integer and may drop to zero, which would lead to exponential growth of the prey population without finite carry capacity. We define a stochastic predator–prey model as a CTMC over the state space

$$\{0, \dots, X_{\mathrm{max}}\} \times \{0, \dots, Y_{\mathrm{max}}\}$$

with transition-rate matrix

$$Q_{(X+\Delta X, Y+\Delta Y),(X,Y)} = \begin{cases} \beta XY & \text{if } \Delta X = -1, \Delta Y = +1, \\ \delta Y & \text{if } \Delta X = 0, \Delta Y = -1, \\ \alpha X - \alpha X^2 / X_{\mathrm{max}} & \text{if } \Delta X = +1, \Delta Y = 0, \\ -\beta XY - \delta Y - \alpha X + \alpha X^2 / X_{\mathrm{max}} & \text{if } \Delta X = 0, \Delta Y = 0, \\ 0 & \text{otherwise} \end{cases} \tag{38}$$

whose columns sum to less than zero. This is because the upper limit Y_{max} of the predator population is a computational cutoff and not enforced by the model. Transitions that leave the state space result in missing probability mass, which must be mitigated by choosing a sufficiently high Y_{max}.

We introduce the band matrices

$$\underbrace{\begin{aligned} \mathcal{X}^+_{\mathrm{cons}} &= \mathrm{superdiag}(1, \dots, X_{\mathrm{max}}), \\ \mathcal{X}^-_{\mathrm{cons}} &= \mathrm{diag}(0, \dots, X_{\mathrm{max}}), \\ \mathcal{X}^+_{\mathrm{birth}} &= \mathrm{subdiag}(0, \dots, X_{\mathrm{max}} - 1), \\ \mathcal{X}^-_{\mathrm{birth}} &= \mathrm{diag}(0, \dots, X_{\mathrm{max}} - 1, 0), \\ \mathcal{X}^+_{\mathrm{cap}} &= \mathrm{subdiag}(0^2, 1^2, 2^2, \dots, (X_{\mathrm{max}} - 1)^2), \\ \mathcal{X}^-_{\mathrm{cap}} &= \mathrm{diag}(0^2, 1^2, 2^2, \dots, (X_{\mathrm{max}} - 1)^2, 0), \\ \mathcal{X}^+_{\mathrm{death}} &= \mathrm{diag}(1, 1, \dots, 1) = \mathrm{I}, \\ \mathcal{X}^-_{\mathrm{death}} &= \mathrm{diag}(1, 1, \dots, 1) = \mathrm{I}, \end{aligned}}_{(X_{\mathrm{max}} + 1) \times (X_{\mathrm{max}} + 1)} \qquad \underbrace{\begin{aligned} \mathcal{Y}^+_{\mathrm{cons}} &= \mathrm{subdiag}(0, \dots, Y_{\mathrm{max}} - 1), \\ \mathcal{Y}^-_{\mathrm{cons}} &= \mathrm{diag}(0, \dots, Y_{\mathrm{max}}), \\ \mathcal{Y}^+_{\mathrm{birth}} &= \mathrm{diag}(1, 1, \dots, 1) = \mathrm{I}, \\ \mathcal{Y}^-_{\mathrm{birth}} &= \mathrm{diag}(1, 1, \dots, 1) = \mathrm{I}, \\ \mathcal{Y}^+_{\mathrm{cap}} &= \mathrm{diag}(1, 1, \dots, 1) = \mathrm{I}, \\ \mathcal{Y}^-_{\mathrm{cap}} &= \mathrm{diag}(1, 1, \dots, 1) = \mathrm{I}, \\ \mathcal{Y}^+_{\mathrm{death}} &= \mathrm{superdiag}(1, \dots, Y_{\mathrm{max}}), \\ \mathcal{Y}^-_{\mathrm{death}} &= \mathrm{diag}(0, \dots, Y_{\mathrm{max}}), \end{aligned}}_{(Y_{\mathrm{max}} + 1) \times (Y_{\mathrm{max}} + 1)} \tag{39}$$

in order to represent the transition-rate matrix

$$\begin{aligned} Q = &+ \beta(\mathcal{X}^+_{\text{cons}} \otimes \mathcal{Y}^+_{\text{cons}}) + \alpha(\mathcal{X}^+_{\text{birth}} \otimes \mathcal{Y}^+_{\text{birth}}) - \frac{\alpha}{X_{\text{max}}}(\mathcal{X}^+_{\text{cap}} \otimes \mathcal{Y}^+_{\text{cap}}) + \delta(\mathcal{X}^+_{\text{death}} \otimes \mathcal{Y}^+_{\text{death}}) \\ &- \beta(\mathcal{X}^-_{\text{cons}} \otimes \mathcal{Y}^-_{\text{cons}}) - \alpha(\mathcal{X}^-_{\text{birth}} \otimes \mathcal{Y}^-_{\text{birth}}) + \frac{\alpha}{X_{\text{max}}}(\mathcal{X}^-_{\text{cap}} \otimes \mathcal{Y}^-_{\text{cap}}) - \delta(\mathcal{X}^-_{\text{death}} \otimes \mathcal{Y}^-_{\text{death}}) \end{aligned} \tag{40}$$

as a sum of tensor products. This allows us to efficiently compute solutions $\mathbf{p}(t)$ of the Kolmogorov equation for the stochastic predator–prey model (see Fig. 6b).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00180-024-01454-9.

Acknowledgements This work was funded by the German Research Foundation (DFG) Grant Nos. TRR-305, SFB/TRR-55 and GR-3179/6-1. We thank Stefan Solbrig for his assistance with the simulations on QPACE 3.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Allen LJS (2017) A primer on stochastic epidemic models: formulation, numerical simulation, and analysis. Infect Disease Modell 2(2):128–142. https://doi.org/10.1016/j.idm.2017.03.001

Al-Mohy AH, Higham NJ (2009) Computing the Fréchet derivative of the matrix exponential, with an application to condition number estimation. SIAM J Matrix Anal Appl 30(4):1639–1657. https://doi.org/10.1137/080716426

Beerenwinkel N, Sullivant S (2009) Markov models for accumulating mutations. Biometrika 96(3):645–661. https://doi.org/10.1093/biomet/asp023

Benson AR, Gleich DF, Lim L-H (2017) The spacey random walk: a stochastic process for higher-order data. SIAM Rev 59(2):321–345. https://doi.org/10.1137/16m1074023

Buchholz Peter, Dayar Tuğrul, Kriege Jan, Orhan M Can (2016) Compact representation of solution vectors in Kronecker-Based Markovian Analysis. Quantitative Evaluation of Systems. Springer, New York pp 260–276. https://doi.org/10.1007/978-3-319-43425-4_18

Buchholz P (1999) Structured analysis approaches for large Markov chains. Appl Numer Math 31(4):375–404. https://doi.org/10.1016/S0168-9274(99)00005-7. (**issn: 0168-9274**)

Buis PE, Dyksen Wayne R (1996) Efficient vector and parallel manipulation of tensor products. ACM Trans Math Softw 22(1):18–23. https://doi.org/10.1145/225545.225548

Bundesministerium für Soziales Gesundheit, Pflege und Konsumentenschutz (BMSGPK) (2021) Open Data Österreich. url:https://www.data.gv.at/katalog/dataset/ef8e980b-9644-45d8-b0e9-c6aaf0eff0c0 (visited on 10/15/2021)

Christel F, Steven A, Dominique VB, Geert M, Erika V, Niel H (2020) Time between symptom onset, hospitalisation and recovery or death: statistical analysis of Belgian COVID-19 Patients. Int J Environ Res Public Health 17(20):7560. https://doi.org/10.3390/ijerph17207560

Duane S, Kennedy AD, Pendleton BPJ, Roweth D (1987) Hybrid Monte Carlo. Phys Lett B 195(2):216–222. https://doi.org/10.1016/0370-2693(87)91197-X

Gelman A, Carlin JB, Stern HS, Rubin DB (2013) Bayesian data analysis, 3rd edn. Chapman and Hall/CRC, Florida. https://doi.org/10.1201/b16018

Georg P, Grasedyck L, Klever M, Schill R, Spang R, Wettig T (2020) Low-rank tensor methods for Markov chains with applications to tumor progression models. arXiv: 2006.08135 [math.NA]

Georg P, Richtmann D, Wettig T (2018) DD-α AMG on QPACE 3. In: EPJ Web of Conferences 175. Ed. by M. Della Morte, P. Fritzsch, E. Gámiz Sánchez, and C. Pena Ruano, p. 02007. issn: 2100-014X. https://doi.org/10.1051/epjconf/201817502007

Gillespie DT (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. J Comput Phys 22(4):403–434. https://doi.org/10.1016/0021-9991(76)90041-3

Grassmann WK (1977) Transient solutions in markovian queueing systems. Comput Operat Res 4(1):47–53. https://doi.org/10.1016/0305-0548(77)90007-7

Hackbusch W (2012) Tensor Spaces and Numerical Tensor Calculus. Springer, Berlin and Heidelberg. 524 pp. https://doi.org/10.1007/978-3-642-28027-6

Ho LS, Tung JX, Crawford FW, Minin VN, Suchard MA (2017) Birth/birthdeath processes and their computable transition probabilities with biological applications. J Math Biol 76(4):911–944. https://doi.org/10.1007/s00285-017-1160-3

Ho LS, Tung FWC, Suchard MA (2018) Direct likelihood-based inference for discretely observed stochastic compartmental models of infectious disease. Annals Appl Stat 12(3):1993–2021. https://doi.org/10.1214/18-AOAS1141

Johnson TH, Clark SR, Jaksch D (2010) Dynamical simulations of classical stochastic systems using matrix product states. Phys Rev E 82(3):036702. https://doi.org/10.1103/physreve.82.036702

Kermack WO, McKendrick AG (1927) A contribution to the mathematical theory of epidemics. Proc R Soc Lond Ser A 115(772):700–721. https://doi.org/10.1098/rspa.1927.0118

Kressner Daniel, Macedo Francisco (2014) Low-rank tensor methods for communicating markov processes. Quantitative Evaluation of Systems. Springer, New York, pp. 25–40. https://doi.org/10.1007/978-3-319-10696-0_4

Lotka AJ (1925) Elements of physical biology. Williams & Wilkins, Philadelphia

McKendrick AG (1925) Applications of mathematics to medical problems. Proc Edinb Math Soc 44:98–130. https://doi.org/10.1017/S0013091500034428

Moler C, Van Loan C (2003) Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. SIAM Rev 45(1):3–49. https://doi.org/10.1137/S00361445024180

Neal RM (2011) MCMC using Hamiltonian dynamics. Handbook of markov chain Monte Carlo 2:11. https://doi.org/10.1201/b10905

Oriol A-P, Virgile A, Colin C, Larry D, Fonnesbeck CJ, Maxim K, Ravin K, Jupeng L, Luhmann CC, Martin OA, Michael O, Ricardo V, Thomas W, Robert Z (2023) PyMC: a modern and comprehensive probabilistic programming framework in python. PeerJ Comput Sci 9:e1516. https://doi.org/10.7717/peerj-cs.1516

Owen J, Wilkinson DJ, Gillespie CS (2014) Scalable inference for Markov processes with intractable likelihoods. Stat Comput 25(1):145–156. https://doi.org/10.1007/s11222-014-9524-7

Reibman A, Trivedi K (1988) Numerical transient analysis of markov models. Comput Operat Res 15(1):19–36. https://doi.org/10.1016/0305-0548(88)90026-3

Schill R, Solbrig S, Wettig T, Spang R (2019) Modelling cancer progression using Mutual Hazard Networks. Bioinformatics 36(1):241–249. https://doi.org/10.1093/bioinformatics/btz513

Sherlock C (2021) Direct statistical inference for finite Markov jump processes via the matrix exponential. Comput Stat 36(4):2863–2887. https://doi.org/10.1007/s00180-021-01102-6

Tang L, Zhou Y, Wang L, Purkayastha S, Zhang L, He J, Wang F, Song PX-K (2020) A review of multi-compartment infectious disease models. Int Stat Rev 88(2):462–513. https://doi.org/10.1111/insr.12402

Vehtari A, Gelman A, Simpson D, Carpenter B, Bürkner P-C (2021) Rank-normalization, folding, and localization: an improved $\hat{R}$ for assessing convergence of MCMC (with discussion). Bayesian Anal 16(2):667–718. https://doi.org/10.1214/20-BA1221

Volterra V (1926) Fluctuations in the abundance of a species considered mathematically. Nature 118(2972):558–560. https://doi.org/10.1038/118558a0

Wolf V (2007) Modelling of biochemical reactions by stochastic automata networks. Electr Notes Theoret Comp Sci 171(2):197–208. https://doi.org/10.1016/j.entcs.2007.05.017

Yamanaka K, Agu M, Miyajima T (1997) A continuous-time asynchronous boltzmann machine. Neural Netw 10(6):1103–1107. https://doi.org/10.1016/S0893-6080(97)00006-3

Authors and Affiliations

Kevin Rupp[1] · Rudolf Schill[1] · Jonas Süskind[1] · Peter Georg[2] · Maren Klever[3] · Andreas Lösch[1] · Lars Grasedyck[3] · Tilo Wettig[2] · Rainer Spang[1]

✉ Rudolf Schill
Rudolf.Schill@klinik.uni-regensburg.de

✉ Rainer Spang
Rainer.Spang@klinik.uni-regensburg.de

[1] Department of Statistical Bioinformatics, University of Regensburg, 93040 Regensburg, Germany

[2] Department of Physics, University of Regensburg, 93040 Regensburg, Germany

[3] Institut für Geometrie und Praktische Mathematik, RWTH Aachen University, 52062 Aachen, Germany