PROCEEDINGS A

royalsocietypublishing.org/journal/rspa

Research

Cite this article: Barendregt NW, Webb EG, Kilpatrick ZP. 2023 Adaptive Bayesian inference of Markov transition rates. *Proc. R. Soc. A* **479**: 20220453.
https://doi.org/10.1098/rspa.2022.0453

Received: 28 June 2022
Accepted: 23 January 2023

Subject Areas:
applied mathematics, statistics

Keywords:
adaptive optimal design, transition rates, Markov process, sequential Bayesian inference

Author for correspondence:
Nicholas W. Barendregt
e-mail: nicholas.barendregt@colorado.edu

Adaptive Bayesian inference of Markov transition rates

Nicholas W. Barendregt[1], Emily G. Webb[2] and Zachary P. Kilpatrick[1]

[1]Department of Applied Mathematics, University of Colorado Boulder, 1111 Engineering Center, ECOT 225, 526 UCB, Boulder, CO 80309, USA
[2]Applied Physics Laboratory, Johns Hopkins University, 11100 Johns Hopkins Road, Laurel, MD 20723, USA

NWB, 0000-0002-3268-9426; ZPK, 0000-0002-2835-9416

Optimal designs minimize the number of experimental runs (samples) needed to accurately estimate model parameters, resulting in algorithms that, for instance, efficiently minimize parameter estimate variance. Governed by knowledge of past observations, adaptive approaches adjust sampling constraints online as model parameter estimates are refined, continually maximizing expected information gained or variance reduced. We apply adaptive Bayesian inference to estimate transition rates of Markov chains, a common class of models for stochastic processes in nature. Unlike most previous studies, our sequential Bayesian optimal design is updated with each observation and can be simply extended beyond two-state models to birth–death processes and multistate models. By iteratively finding the best time to obtain each sample, our adaptive algorithm maximally reduces variance, resulting in lower overall error in ground truth parameter estimates across a wide range of Markov chain parameterizations and conformations.

1. Introduction

Using experimental data to infer parameters is essential for accurate quantitative models of natural phenomena. Inherent stochasticity in most physical systems compounds this difficulty, clouding the link between

THE ROYAL SOCIETY PUBLISHING

data and ground truth in ways experimentalists cannot control. Not only does noise cause uncertainty in model parameter estimates, it can also slow the process of model refinement. As a result, researchers historically utilized statistical methods to design experiments that maximize the information obtained from each experimental measurement [1–3]. In particular, Bayesian experimental design, applied to a system with unknown parameters $\mathbf{x}$, starts with a prior $p(\mathbf{x})$, constructs the associated posterior $p(\mathbf{x}|\Theta, \Xi)$ based on data Θ obtained from an experimental design Ξ and finds the design that optimizes a specified objective, such as minimizing a utility function (e.g. variance) that incorporates sampling-associated costs (e.g. time or resources needed to take measurements) [4,5]. These methods have seen wide application in economics [6], queueing theory [7], physics [8,9] and cognitive neuroscience [10,11]. In particular, Bayesian experimental design for inferring transition rates in discrete-state Markov models has seen great success when applied to simple epidemiological [12–14] and ecological [15,16] models. Recent efforts have shown that adaptive designs can speed up the timescale of clinical drug or intervention trials [17], providing an automated, model-based prescription that governs future sampling.

One of the outputs of Bayesian experimental designs, for systems producing time series data, is a sampling schedule, a set of times to measure the state of a system, chosen to optimize an objective function (e.g. minimizing sample number for a fixed estimate tolerance, maximizing sample information). When schedules are planned in advance of experiments, they may require sampling continuously in time or periodically with a fixed sampling frequency [18], which may be infeasible or inefficient given high sampling costs. For example, an ecologist studying the dynamics of several interacting species may be restricted by seasonal patterns of animal activity, the expense or time cost of field work, or a finite project timeline, such that they are unable to implement a fixed schedule of sampling population sizes. In these situations, Bayesian experimental design can be extended to incorporate sequential analysis [19], yielding iterative and adaptive sampling schedules based on prior observations. While approximate versions of these Bayesian adaptive designs have been applied to simple models [20], precise sequential formulations applied to more complicated systems are often limited by intractable likelihood functions. This difficulty has spawned advanced algorithms involving stochastic optimization [21,22], Markov Chain Monte Carlo (MCMC) [5,23,24] and machine learning methods [25,26].

In this work, we develop an adaptive sequential Bayesian inference algorithm that successively optimizes each process sample time to minimize the variance of transition rate parameter estimates for discrete-state Markov processes with arbitrary numbers of states and transition rates. An early version of this work focused on simple two-state processes with a single-transition rate [27]. Starting with two-state Markov chains, we illustrate how sequentially chosen sampling times are selected to minimize expected parametric posterior variance after each observation. We compare the speed and accuracy of this adaptive algorithm to that of a fixed-period sampling algorithm across transition rate parameter space. Considering more complex Markov chains, our algorithm can be extended by minimizing the expected determinant of the covariance matrix associated with the transition rate matrix. We apply this algorithm to three specific Markov chain models: a ring of states, modelling the diffusive degradation of memory for a single circular parameter [28,29], a birth–death process describing population dynamics of epidemics and ecological groups [12] and a general Markov chain inference problem with a binarized prior for each transition rate. Taken collectively, these results demonstrate a simple yet powerful approach to efficiently inferring the dynamics of Markov models.

2. Methods: adaptive Bayesian design for Markov chains

Our Bayesian inferential approach to determining transition rates of a Markov chain from time series observations relies on obtaining accurate representations of the parametric likelihood functions from state observations. This is plausible in the case of simple Markov chains for which likelihoods can be determined analytically. Of course, for continuous-time Markov chains, it

is always possible to write down likelihood formulas, but computation becomes infeasible for sufficiently large chains. In this section, we derive our algorithm for adaptive Bayesian inference through a series of examples with increasing complexity. These examples culminate with our derivation of the adaptive algorithm for general time-homogeneous, continuous-time Markov chains.

(a) Inferring single-transition rates

We start by considering a continuous-time Markov process with two states $X(t) \in \{0, 1\}$ and a single-transition rate $h_0 > 0$, so that

$$\left.\begin{aligned} &\Pr(X(t+\delta t)=1|X(t)=0) = h_0\delta t + o(\delta t), \quad \delta t \downarrow 0, \\ &\Pr(X(t+\delta t)=0|X(t)=1) = 0, \qquad\qquad\quad \forall \delta t \geq 0. \end{aligned}\right\} \tag{2.1}$$

For this Markov process with a single-transition link, we will enforce the initial condition $X(0) = 0$ to guarantee the possibility of observing a transition from $X = 0$ to $X = 1$. Our goal is to design an algorithm for sampling from this process efficiently to optimally decrease the estimate variance of the transition rate parameter h_0 with each state sample (after reinitialization). By using Bayesian inference, we assume a prior distribution $p_0(h_0)$ and obtain measurements of the process $\xi_i = (t_i, X(t_i)) = (t_i, X_i)$, where $i \in \{0, 1, 2, \ldots\}$ represents the sample number index and X evolves according to equation (2.1) and always taking $X(0) = 0$. These samples allow us to construct the posterior $p_n(h_0) = p(h_0|\xi_{1:n})$. Sampling times t_i are chosen to minimize the expected posterior variance after each observation. The algorithm samples observations until a predetermined threshold variance θ is reached. At this point, the transition rate estimator $\hat{h}_0$ is the maximum *a posteriori* estimator.

To illustrate the process of selecting each sample time t_n, suppose we have previously observed $n-1$ measurements, $\xi_{1:n-1} = \{(t_1, X_1), \ldots (t_{n-1}, X_{n-1})\}$. Given a particular planned subsequent observation time t_n, the expected posterior variance on the next (nth) timestep $\overline{\mathrm{Var}}_n(h_0|\xi_{1:n-1}, t_n)$ is given by marginalizing over the possible future measurements ξ_n (i.e. possible state observations $X(t_n)$) assuming the history of observations $\xi_{1:n-1}$ and a sample time t_n:

$$\begin{aligned} \overline{\mathrm{Var}}_n(h_0|\xi_{1:n-1}, t_n) = &\,\mathrm{Var}(h_0|X_n = 0, t_n, \xi_{1:n-1}) \Pr(X_n = 0|t_n, \xi_{1:n-1}) \\ &+ \mathrm{Var}(h_0|X_n = 1, t_n, \xi_{1:n-1}) \Pr(X_n = 1|t_n, \xi_{1:n-1}). \end{aligned}$$

Note, that since we always reset $X(0) = 0$ preceding each sample, the relevant conditional observation probabilities are as follows:

$$p(\xi_n|h_0) = \begin{cases} 1 - e^{-h_0 t_n}, & X(t_n) = 1 \\ e^{-h_0 t_n}, & X(t_n) = 0 \end{cases} \tag{2.2}$$

and

$$p(\xi_n|\xi_{1:n-1}) = \int_0^\infty p(\xi_n|h_0) p_{n-1}(h_0) dh_0.$$

Thus, we obtain the expected variance formula:

$$\begin{aligned} \overline{\mathrm{Var}}_n(h_0|\xi_{1:n-1}, t_n) = \int_0^\infty p_{n-1}(h_0) \Bigg[&\mathrm{e}^{-h_0 t_n} \left\{ h_0 - \frac{\int_0^\infty h_0 \,\mathrm{e}^{-h_0 t_n} p_{n-1}(h_0)\, \mathrm{d}h_0}{\int_0^\infty \mathrm{e}^{-h_0 t_n} p_{n-1}(h_0)\, \mathrm{d}h_0} \right\}^2 \\ &+ (1 - \mathrm{e}^{-h_0 t_n}) \left\{ h_0 - \frac{\int_0^\infty h_0 (1 - \mathrm{e}^{-h_0 t_n}) p_{n-1}(h_0)\, \mathrm{d}h_0}{\int_0^\infty (1 - \mathrm{e}^{-h_0 t_n}) p_{n-1}(h_0)\, \mathrm{d}h_0} \right\}^2 \Bigg] \mathrm{d}h_0. \end{aligned} \tag{2.3}$$

Algorithm 1. Single-transition adaptive Bayesian inference.

Require: $n = 0, \theta > 0, p_0(h_0)$ ▷ p_0: prior with support $[0, \infty)$.
  while $\mathrm{Var}_n(h_0) > \theta$ **do**
    $n \leftarrow n + 1$
    $t_n \leftarrow \arg\min_{t \geq 0} \overline{\mathrm{Var}}_n(h_0; t)$ ▷ Calculate $\overline{\mathrm{Var}}_n$ using equation (2.3).
    Draw $\xi_n = (t_n, X_n)$
    $p_n(h_0) \leftarrow \frac{p(\xi_n|h_0)p_{n-1}(h_0)}{\int_0^\infty p(\xi_n|h_0)p_{n-1}(h_0)\,\mathrm{d}h_0}$
  end while

The sample time t_n that minimizes equation (2.3) depends on the posterior p_{n-1}, computed from the sequence of previous observations of the stochastic process. Because each sample time depends on the previous posterior distribution, this algorithm performs sequential and *adaptive Bayesian inference*. The expected variance formula in equation (2.3) can be defined iteratively. Moreover, once t_n is chosen by minimizing equation (2.3) and an observation is made for X_n, we can calculate the true variance after the nth observation as follows:

$$\mathrm{Var}_n(h_0) = \int_0^\infty p_n(h_0) \left\{ h_0 - \int_0^\infty h_0 p_n(h_0)\,\mathrm{d}h_0 \right\}^2 \mathrm{d}h_0, \tag{2.4}$$

where

$$p_n(h_0) = p_{n-1}(h_0) \cdot \begin{cases} 1 - \mathrm{e}^{-h_0 t_n}, & X(t_n) = 1 \\ \mathrm{e}^{-h_0 t_n}, & X(t_n) = 0. \end{cases}$$

This leads us to propose algorithm 1. Unless otherwise noted, we will take the variance threshold, which terminates the accumulation of observations, to be $\theta = 0.1$ throughout this work.

(b) Multi-dimensional inference on simple chains

Our adaptive inference algorithm easily extends to chains with multiple transition rates, as we simply need to compute the state probability distribution for the Markov chain and include that in our Bayesian update. To illustrate, consider the same two-state Markov process, but with transitions occurring bidirectionally with transition rates h_0 $(0 \mapsto 1)$ and h_1 $(1 \mapsto 0)$. Expanding the inference problem beyond a single dimension requires defining a new objective function to minimize, which will now involve multiple transition rate parameters: variability in the estimate is now defined by the posterior covariance matrix Σ rather than the variance. There are several ways to 'minimize' a covariance matrix [2,30]. Here, we take the approach of minimizing the determinant of the expected covariance, known as a 'D-optimal' method in optimal experimental design. Such an approach is also equivalent to maximizing the product of the eigenvalues of the Fisher information matrix [31]. Maximizing information gain here is preferable to reducing averaging variance (as in A-optimal designs), since there could be strong asymmetry in the transition rate parameters.

The process is always guaranteed to eventually switch from one state to another as long as both rates are non-zero, and the transition rate parameters can both be inferred to arbitrarily small variances given enough observations. Thus, we avoid the need to reset the chain's state after each observation. For simplicity, we assume $X_0 = X(t_0 = 0) = 0$ to begin, but it is not difficult to extend the algorithm to the case where X_0 is chosen randomly, and we subsequently allow the variable to evolve according to a Markov chain whose transition rates are chosen from the prior, $(h_0, h_1) \sim p_0(h_0, h_1)$. Thereafter, $X_n = X(t_n)$ is drawn and compared with $X_{n-1} = X(t_{n-1})$ to update the posterior over the transition rates (h_0, h_1).

As with the adaptive inference procedure for a single-transition rate, we determine the next sample time t_n after the current time t_{n-1} by minimizing the determinant of the expected covariance matrix. For an arbitrary subsequent sampling time t_n, the expected covariance $[\Sigma_n]_{ij} \equiv \overline{\mathrm{Cov}}_n(h_i, h_j)$ is computed by marginalizing over the possible observations X_n and conditioning on

royalsocietypublishing.org/journal/rspa Proc. R. Soc. A 479: 20220453

Algorithm 2. Bidirectional two-state chain adaptive Bayesian inference.

Require: $n = 0, \theta > 0, p_0(h_0, h_1)$ ▷ p_0: prior with support $[0, \infty)^2$.
  while $\det(\mathrm{Cov}_n(h_0, h_1)) > \theta$ **do**
    $n \leftarrow n + 1$
    $t_n \leftarrow \arg\min_{t \geq t_{n-1}} \det\left(\overline{\mathrm{Cov}}_n(h_0, h_1; t)\right)$ ▷ Calculate $\overline{\mathrm{Cov}}_n$ using Eq. (2.6).
    Draw $\xi_n = (t_n, X_n)$
    $p_n(h_0, h_1) \leftarrow \frac{p(\xi_n|h_0,h_1)p_{n-1}(h_0,h_1)}{\iint_{\mathbb{R}^2_{\geq 0}} p(\xi_n|h_0,h_1)p_{n-1}(h_0,h_1)\,dh_0\,dh_1}$
  end while

the past observations $\xi_{1:n-1}$:

$$\begin{aligned}\overline{\mathrm{Cov}}_n(h_i, h_j) = \mathrm{Cov}_n(h_i, h_j|X_n = 0, t_n, \xi_{1:n-1})p(X_n = 0|t_n, \xi_{1:n-1}) \\ + \mathrm{Cov}_n(h_i, h_j|X_n = 1, t_n, \xi_{1:n-1})p(X_n = 1|t_n, \xi_{1:n-1}).\end{aligned}$$

Now, marginalizing over transition probabilities from the previous state X_{n-1}, which we can define for all possible cases

$$p(X_n = j|X_{n-1} = i) = \begin{cases} \dfrac{h_1}{h_0+h_1} + \dfrac{h_0}{h_0+h_1}\,\mathrm{e}^{-(h_0+h_1)(t_n-t_{n-1})}, & i = j = 0 \\ \dfrac{h_0}{h_0+h_1} - \dfrac{h_0}{h_0+h_1}\,\mathrm{e}^{-(h_0+h_1)(t_n-t_{n-1})}, & i = 0,\ j = 1 \\ \dfrac{h_1}{h_0+h_1} - \dfrac{h_1}{h_0+h_1}\,\mathrm{e}^{-(h_0+h_1)(t_n-t_{n-1})}, & i = 1,\ j = 0 \\ \dfrac{h_0}{h_0+h_1} + \dfrac{h_1}{h_0+h_1}\,\mathrm{e}^{-(h_0+h_1)(t_n-t_{n-1})}, & i = j = 1 \end{cases}, \tag{2.5}$$

yields the expected future covariance

$$\begin{aligned}\overline{\mathrm{Cov}}_n(h_i, h_j; t_n) = \sum_{k=0}^{1} \iint_{\mathbb{R}^2_{\geq 0}} \Bigg[\left\{ h_i - \frac{\int_0^\infty h_i \left(\int_0^\infty p(X_n = k|X_{n-1})p_{n-1}(h_0,h_1)^2\,\mathrm{d}h_j \right) \mathrm{d}h_i}{\iint_{\mathbb{R}^2_{\geq 0}} p(X_n = k|X_{n-1})p_{n-1}(h_0,h_1)^2\,\mathrm{d}h_0\,\mathrm{d}h_1} \right\} \\ \times \left\{ h_j - \frac{\int_0^\infty h_j \left(\int_0^\infty p(X_n = k|X_{n-1})p_{n-1}(h_0,h_1)^2\,\mathrm{d}h_i \right) \mathrm{d}h_j}{\iint_{\mathbb{R}^2_{\geq 0}} p(X_n = k|X_{n-1})p_{n-1}(h_0,h_1)^2\,\mathrm{d}h_0\,\mathrm{d}h_1} \right\} \\ \times p(X_n = k|X_{n-1})p_{n-1}(h_0,h_1)^2 \Bigg]\,\mathrm{d}h_0\,\mathrm{d}h_1.\end{aligned} \tag{2.6}$$

Note the extra factor of the posterior from the previous sequence of observations p_{n-1} appears to properly weight the probability of transitioning from X_{n-1} to X_n. Using the determinant of the expected covariance as computed by equation (2.6), we modify algorithm 1 to obtain the multi-dimensional adaptive inference algorithm shown in algorithm 2. After a sample $\xi_n = (t_n, X_n)$, the resulting covariance is

$$\begin{aligned}\mathrm{Cov}_n(h_i, h_j) = \iint_{\mathbb{R}^2_{\geq 0}} \Bigg[\left\{ h_i - \iint_{\mathbb{R}^2_{\geq 0}} h_i p_n(h_0,h_1)\,\mathrm{d}h_j\,\mathrm{d}h_i \right\} \\ \times \left\{ h_j - \iint_{\mathbb{R}^2_{\geq 0}} h_j p_n(h_0,h_1)\,\mathrm{d}h_i\,\mathrm{d}h_j \right\} p_n(h_0,h_1) \Bigg]\,\mathrm{d}h_0\,\mathrm{d}h_1.\end{aligned}$$

(c) Adaptive Bayesian inference for arbitrary Markov chains

While the algorithms proposed earlier considered chains with two states, we can generalize our approach to arbitrary chains by considering systems with higher-dimensional covariance matrices whose determinants we treat as our objective function. Let $X(t)$ be a discrete-state Markov process with m states and d transition rates, denoting X_n as the nth state sample, the state $k \in \{0, \ldots, m-1\}$, and the transition rate h_i indexed by $i \in \{1, \ldots, d\}$. Note that we could index transition rates as h_{ij} using the ordered pair for the rate of transition from state $X^j \to X^i$, but the single index formulation leads to a more concise form in the terms hereafter. Moreover, we do not always consider Markov chains with complete digraph transition rate conformations that would benefit from ordered pair notation.

To infer the d-dimensional vector $\mathbf{h}$ of transition rates, we construct a posterior distribution, for instance, after the $(n-1)$th sample from the sequence $\xi_{1:n-1} = \{(t_1, X_1), \ldots, (t_{n-1}, X_{n-1})\}$ with a covariance matrix $\Sigma_{n-1} \equiv \mathrm{Cov}_{n-1} \in \mathbb{R}^{d\times d}$ having entries $[\Sigma_{n-1}]_{ij} \equiv \mathrm{Cov}_{n-1}(h_i, h_j)$ defining the covariance between the estimates of the transition rates h_i and h_j. Our algorithm then proceeds in choosing the next sample time t_n that minimizes the determinant of the expected covariance matrix $\det(\overline{\mathrm{Cov}}_n(t))$. By marginalizing over possible observable states $X_n \in \{0, \ldots, m-1\}$, the entries of $\overline{\mathrm{Cov}}_n$, averaged for a particular choice of the next sample time t_n, are given by

$$\overline{\mathrm{Cov}}_n(h_i, h_j) = \sum_{k=0}^{m-1} \mathrm{Cov}_n(h_i, h_j | X_n = k, t_n, \xi_{1:n-1}) \Pr(X_n = k | t_n, \xi_{1:n-1}). \tag{2.7}$$

As mentioned earlier, we consider the marginalization in equation (2.7) in the context of transitions from the previous (known) state X_{n-1}. This requires introducing the transition probabilities $p(X_n = k | X_{n-1}, \mathbf{h})$ and the posterior p_{n-1} into equation (2.7). Formulas for the transition probabilities can be obtained explicitly in a number of cases, but they are not as concise as in the case of two state chains. In general, the explicit update rule for the entries of the expected covariance matrix will be

$$\begin{aligned}\overline{\mathrm{Cov}}_n(h_i, h_j) = \sum_{k=0}^{m-1} \iint_{\mathbb{R}^2_{\geq 0}} \Bigg[\Bigg\{ h_i - \frac{\int_0^\infty h_i \left(\int_0^\infty p(X_n = k | X_{n-1}, \mathbf{h}) p_{n-1}(\mathbf{h})^2 \, d_i^{d-1}\mathbf{h} \right) \mathrm{d}h_i}{\int_0^\infty p(X_n = k | X_{n-1}, \mathbf{h}) p_{n-1}(\mathbf{h})^2 \, d^d\mathbf{h}} \Bigg\} \\ \times \Bigg\{ h_j - \frac{\int_0^\infty h_j \left(\int_0^\infty p(X_n = k | X_{n-1}, \mathbf{h}) p_{n-1}(\mathbf{h})^2 \, d_j^{d-1}\mathbf{h} \right) \mathrm{d}h_j}{\int_0^\infty p(X_n = k | X_{n-1}, \mathbf{h}) p_{n-1}(\mathbf{h})^2 \, d^d\mathbf{h}} \Bigg\} \\ \times \int_0^\infty p(X_n = k | X_{n-1}, \mathbf{h}) p_{n-1}(\mathbf{h})^2 \, d_{ij}^{d-2}\mathbf{h} \Bigg] \mathrm{d}h_i \, \mathrm{d}h_j. \end{aligned} \tag{2.8}$$

In equation (2.8), we use the notation $\int \cdot \, d^x_{y_1 y_2 \ldots}\mathbf{z}$ to denote that the integral is x-dimensional, and the directions $\{y_1, y_2, \ldots\}$ are the directions *not* integrated over. For example, $\int \cdot \, d_i^{d-1}\mathbf{h}$ indicates we integrate over all directions in $\mathbf{h}$ except h_i. Note that for conservation, the number of subindices y_i and the dimension of the integral x must add to the dimension of the space $\mathbf{z}$. To use equation (2.8) to infer a network's transition rates, we substitute this covariance update in place of equation (2.6) in algorithm 2 and modify the normalization step of the posterior appropriately to define algorithm 3. In the following, we apply this generalized algorithm to infer transition rates to perform rate inference in some canonical Markov chain models.

(d) Numerical implementation of adaptive Bayesian inference algorithms

To numerically implement algorithms 1–3, we must specify a method for finding the optimal sampling times t_n and for updating the posterior density after an observation. To solve the optimization problem of finding the sampling time t_n that minimizes the expected covariance

Algorithm 3. Adaptive Bayesian inference for general Markov chains.

Require: $n = 0, \theta > 0, p_0(\mathbf{h})$ ▷ p_0: prior with support $[0, \infty)^d$.
  while $\det(\mathrm{Cov}_n(\mathbf{h})) > \theta$ **do**
    $n \leftarrow n + 1$
    $t_n \leftarrow \arg\min_{t \geq t_{n-1}} \det\left(\overline{\mathrm{Cov}}_n(\mathbf{h}; t)\right)$ ▷ Calculate $\overline{\mathrm{Cov}}_n$ using Eq. (2.8).
    Draw $\xi_n = (t_n, X_n)$
    $p_n(\mathbf{h}) \leftarrow \frac{p(\xi_n|\mathbf{h})p_{n-1}(\mathbf{h})}{\iint_{\substack{\mathbb{R}^d \\ \geq 0}} p(\xi_n|\mathbf{h})p_{n-1}(\mathbf{h})\, d\mathbf{h}}$
  end while

$\overline{\mathrm{Cov}}_n$, subject to the constraint $t_n \geq 0$, we used MATLAB's `fminbnd`, which combines golden section search and parabolic interpolation to find optima on a bounded interval. Each successive guess is generated as a convex linear combination of two endpoints of the proposed interval in which the optimum lies [32]. We found that standard constrained optimization routines, such as `fmincon` in MATLAB or `scipy.optimize.minimize` in Python, provided good results. However, we also found that using a bounded optimization routine provided more accurate results for a sufficiently large upper bound on t_n. To update the posterior density, we first constructed a prior density p_0 on a mesh of possible transition rates $\mathbf{h}$ and then calculated the posterior update on the same mesh using the exact update rule for the chain configuration (see equation (3.4)). Because we have access to the exact posterior update rule for any finite-state Markov chain, we were able to avoid the usual computational concerns associated with evaluating posterior densities. For further information about our numerical implementations, see https://github.com/nwbarendregt/AdaptMarkovRateInf.

3. Results: adaptive Bayesian inference applied to classic Markov chain models

Having developed our adaptive Bayesian inference algorithm for arbitrary discrete-state Markov chains, we now proceed to study the algorithm's performance on a variety of Markov chain models. In all our results, we run the inference algorithm until it has converged, which we define as the first time the determinant of the posterior covariance (or, in the case of inferring a single-transition rate, the posterior variance) drops below a threshold θ. To measure algorithm performance, we use two metrics: (1) the number of samples ξ needed for the algorithm to converge, which we define as N_s, and (2) the mean-squared error (MSE) of the final posterior output of the algorithm after convergence. The MSE associated with a transition rate h_i is defined with respect to the whole posterior, not just the maximum likelihood or the posterior mode:

$$\mathrm{MSE}_i = \int_0^\infty (h_i - h_i^{\mathrm{true}})^2 \left[\int_0^\infty p_{N_s}(\mathbf{h})\, d_i^{d-1}\mathbf{h} \right] \mathrm{d}h_i. \tag{3.1}$$

In equation (3.1), N_s is the total number of samples to convergence and h_i^{true} is the true transition rate, and the integral notation $\int \cdot\, d_i^{d-1}\mathbf{h}$ has the same meaning as in equation (2.8). Note that we use MSE as a measure of inference error rather than the posterior covariance. Posterior covariance is used to guide sampling time choice and as the convergence criterion since it implies reduction of uncertainty. It is therefore equal across converged algorithms. However, posterior covariance does not necessarily imply how well the algorithm can recover a ground truth model generating observations. In fact, some parameter fitting procedures can lead to biased estimates, in which case uncertainty implied by the parametric posterior is reduced by maximum likelihood estimates straying from the ground truth [33]. This is why we use MSE to evaluate algorithm performance, but this quantity requires knowledge of the true transition rate, which we have access to in our simulation-based studies. Alternatively, an experimentalist who wants to study algorithmic performance applied to a given physical system, in which the true transition rates may be unknown, could instead use a performance metric that measures how well the algorithm's output replicates experimental data, such as Bayes factor analysis [34].

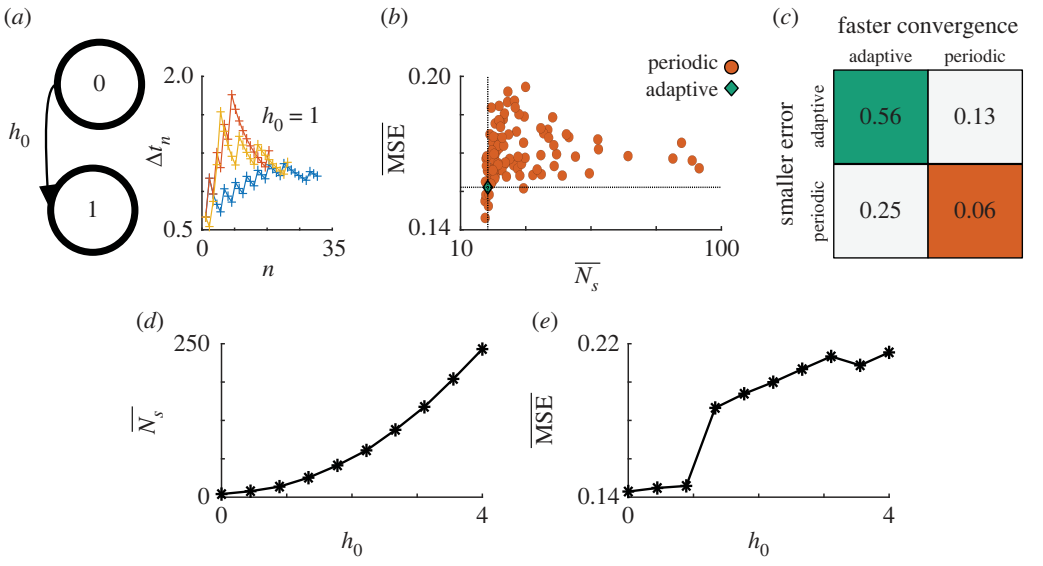

Figure 1. Inferring single-transition rates. (*a*) Schematic of two-state Markov process with single-transition rate h_0. Inset plot shows realizations of the optimal sampling interval computed by the adaptive inference algorithm. (*b*) Scatter plot of algorithms' performance metrics, measured in the number of samples to converge ($\overline{N_s}$) and average error ($\overline{\text{MSE}}$) for inferring the transition rate $h_0 = 1$. Green diamond shows performance of the adaptive algorithm, and each orange dot shows the performance of the periodic algorithm for a different fixed sampling period T. Periodic algorithm dots that lie in the upper-right quadrant relative to the adaptive algorithm have strictly worse performance than the adaptive algorithm. (*c*) Average performance comparison between adaptive and fixed-period algorithms, taken over 10^3 values of h_0. The periodic algorithm only outperforms the adaptive algorithm 6% of the time. (*d*) Average number of samples to converge $\overline{N_s}$ of adaptive algorithm as a function of fixed transition rate h_0. Averages taken over 10^3 realizations with the same h_0. (*e*) $\overline{\text{MSE}}$ of adaptive algorithm as a function of h_0, using the same simulations as in (*d*).

(a) Single-transition rate inference

We compared the performance of adaptive inference for the two-state, single-transition rate chain schematized in figure 1*a*, defined by algorithm 1, to that of an algorithm using a predetermined, fixed sampling period T, which for a general Markov chain is given by algorithm 4. Randomly selecting parameters from a gamma-distributed prior with hyperparameters $(\alpha, \beta) = (2, 1)$, we determined the average number of samples for both algorithms to converge and the average mean-squared error (MSE) of both algorithms once they have converged. Note that the periodic algorithm's performance will depend on the choice of the fixed sampling period T; for the adaptive inference algorithm, which does not take T as an input, performance will be independent of T. Comparing the performance of both algorithms across a range of values of T, we found that the adaptive algorithm tended to strongly outperform the periodic algorithm in convergence time and inference error. This performance advantage can be clearly observed both when inferring a single ground truth transition rate parameter (figure 1*b*) and across many realizations of transition rates (figure 1*c*). In addition, the adaptive algorithm inherently identifies the best choice of the sampling time for each sample based on the posterior over transition rates inferred so far, integrating the process of parameter inference with sampling method as an online experimental design. We also measured the number of samples required and the MSE of the adaptive algorithm for fixed h_0 to investigate any systematic biases in adaptive inference. Inferring larger transition rates require more samples (figure 1*d*) due to (a) the sensitivity of the posterior to state observations at short observation times and (b) low prior likelihood due to their place in the tail of the gamma distribution. Inference error, defined as the MSE between the posterior and true

Algorithm 4. Fixed-period inference algorithm.

Require: $n = 0, \theta > 0, p_0(\mathbf{h}), T > 0$ ▷ p_0: prior with support $[0, \infty)^d$.
  while $\det(\mathrm{Cov}_n(\mathbf{h})) > \theta$ **do**
    $n \leftarrow n + 1$
    $t_n \leftarrow t_{n-1} + T$ ▷ Update sample time using fixed-period input T.
    Draw $\xi_n = (t_n, X_n)$
    $p_n(\mathbf{h}) \leftarrow \frac{p(\xi_n|\mathbf{h})p_{n-1}(\mathbf{h})}{\iint_{\mathbb{R}^d_{\geq 0}} p(\xi_n|\mathbf{h})p_{n-1}(\mathbf{h})\,d\mathbf{h}}$
  end while

transition rates, was uniformly low across all values of h_0 (figure 1*e*). Note that termination at low posterior variance does not ensure low MSE, since observations could guide the mean estimate away from the true value, and Bayesian inference of parameters with insufficient sample sizes or priors can be strongly biased [33].

(b) Higher-dimensional inference on two-state chains

To assess the adaptive algorithm's performance on a slightly higher-dimensional problem, we considered a two-state Markov chain with two transition rates (schematized in figure 2*a*) and compared algorithm 2 to a fixed-period sampling algorithm, identifying the best period for a given prior. To mirror the gamma prior used previously, we chose a bivariate gamma prior with joint distribution function [35]

$$p_0(h_0, h_1) = C\Gamma(b)(h_0 h_1)^{c-1}\left(\frac{h_0}{\mu_1} + \frac{h_1}{\mu_2}\right)^{((a-1)/2)-c} \exp\left\{-\frac{1}{2}\left(\frac{h_0}{\mu_1} + \frac{h_1}{\mu_2}\right)\right\} \times W_{c-b+((1-a)/2),c-(a/2)}\left(\frac{h_0}{\mu_1} + \frac{h_1}{\mu_2}\right), \tag{3.2}$$

where C is given by

$$\frac{1}{C} = (\mu_1\mu_2)^c \Gamma(c)\Gamma(a)\Gamma(b),$$

$c = a + b$, and W is the Whittaker function given by

$$W_{\lambda,\mu}(a) = \frac{a^{\mu+(1/2)}\mathrm{e}^{-(a/2)}}{\Gamma\left(\mu - \lambda + \frac{1}{2}\right)} \int_0^\infty t^{\mu-\lambda-\frac{1}{2}}(1+t)^{\mu+\lambda-\frac{1}{2}}\,\mathrm{e}^{-at}\,\mathrm{d}t.$$

We will take $\mu_1 = \mu_2 = 2$ and $a = b = 1$ throughout when using the bivariate gamma prior given by equation (3.2). Sampling different pairs of transition rates (h_0, h_1) from this prior, we compared the convergence time and inference error of both algorithms using the same approach we implemented for the single-transition problem (figure 2*b*,*c*). For this higher-dimensional inference problem, the adaptive algorithm still showed better performance across a range of different transition rate pairs. Because an experimentalist cannot know the best sampling period *a priori*, these results suggest the adaptive algorithm is generally faster and more accurate than the naive approach.

Does increasing the dimension of the inference problem introduce any new biases in adaptive inference? We measured average convergence time (figure 2*d*) and average inference errors (figure 2*e*,*f*) of the adaptive inference algorithm for fixed transition rates using the same bivariate gamma prior. Similar to the single-transition network in §3a, inferring larger transition rates takes more samples, as these rates are in the tails of the bivariate gamma prior and observables (state sequences) are less sensitive to subtle changes in parameters (transition rates). In addition, the average MSE associated with the each rate is consistently low across most of parameter space.

The notable exception to this low-error behaviour is when one of the transition rates (h_i) is identically zero in which case the estimate of the other transition rate (h_j) is poor. To understand this behaviour, consider a network where $h_0 = 0$ and $h_1 \geq 0$. If the initial state is $X_0 = 0$, then no

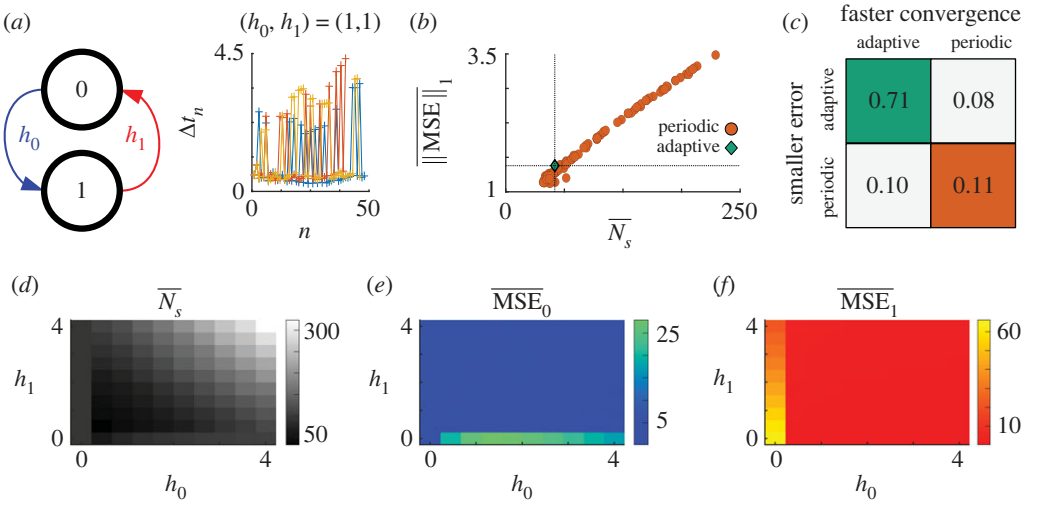

Figure 2. Inferring multiple transition rates. (*a*) Schematic of two-state Markov process with two transition rates h_0, h_1 (originating from state 0 and 1, respectively). Inset plot shows realizations of the optimal sampling interval computed by the adaptive inference algorithm. Note the distinct dichotomous nature of the set of time choices. (*b*) Scatter plot of algorithms' performance metrics, measured in the number of samples to converge ($\overline{N}_s$) and average norm of the error ($\overline{||\mathrm{MSE}||}_1$) for inferring the transition rate pair $(h_0, h_1) = (1, 1)$. Green diamond shows performance of the adaptive algorithm, and each orange dot shows the performance of the periodic algorithm for a different fixed sampling period T. Periodic algorithms that lie in the upper-right quadrant relative to the adaptive algorithm have strictly worse performance than the adaptive algorithm. (*c*) Average performance comparison between adaptive and fixed-period algorithms, taken over 10^3 pairs of (h_0, h_1). The periodic algorithm only outperforms the adaptive algorithm 11% of the time. (*d*) Mean number of samples $\overline{N}_s$ required for adaptive algorithm to reach the covariance determinant threshold as a function of fixed true transition rates h_0, h_1. Averages taken over 10^3 realizations with the same pair of transition rates. (*e*) $\overline{\mathrm{MSE}}$ for h_0 inference of adaptive algorithm as a function of h_0, h_1, taken using the same simulations as in (*d*). (*f*) Same as (*e*), but for $\overline{\mathrm{MSE}}$ for h_1 inference.

transitions will occur. Obtaining the same $X_n = 0$ measurements implies either (1) the rate h_0 is small or (2) the rate h_1 is large. The posterior of the adaptive algorithm converges to account for both of these possibilities, which results in small errors in h_0 inference and potentially large errors in h_1 inference. Heuristically, we can explain this behaviour by noting that repeated $X_n = 0$ measurements means that, in the large-sample limit and for fixed inter-sample interval δt, the posterior scales as powers of the first term in equation (2.5):

$$p_n(h_0, h_1) \propto \left[\frac{h_1}{h_0 + h_1} + \frac{h_0}{h_0 + h_1} e^{-(h_0+h_1)\delta t} \right]^n.$$

Fixing $h_1 \geq 0$, this posterior is maximized at $h_0 = 0$, so as n increases, p_n will asymptotically converge to a delta distribution along the h_0 dimension at $h_0 = 0$. Simultaneously, for fixed $h_0 \to 0$, p_n appears as a flat distribution along the h_1 dimension, implying the posterior contains no information about h_1. These results demonstrate that achieving accurate inference requires utilizing a prior with dimension equal to the dimension of the problem.

(i) Effect of convergence tolerance on adaptive inference

So far we have investigated the adaptive algorithm's convergence speed and inference error for a fixed convergence tolerance θ. However, the choice of θ may impact the algorithm's performance. By using the simple two-state Markov processes discussed earlier, we measured the convergence time (figure 3*a*) and inference error (figure 3*b*) of the adaptive inference algorithm as θ is varied. Changing θ leads to a trade-off between the error in the estimate and the number of samples

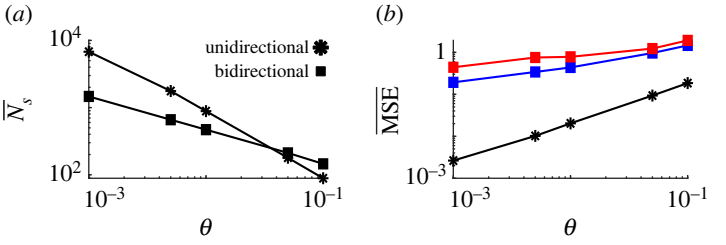

Figure 3. Performance of adaptive algorithm for varied convergence tolerances. (*a*) Average number of samples $\overline{N}_s$ required for the adaptive algorithm to converge for different tolerances θ applied to both two-state networks in §§3(a) and 3(b). Averages taken over 10^3 transition rates drawn from a $\Gamma(2, 1)$ prior for the unidirectional network and 10^3 pairs of rates from the bivariate gamma prior, equation (3.2), for the bidirectional network. (*b*) Mean-squared error $\overline{\text{MSE}}$ of adaptive algorithm with varied θ applied to the same networks and sampled transition rates as in (*a*).

required for convergence: lower (tighter) tolerances require more samples to converge, but yield low-error estimates. By comparing the algorithm's performance between unidirectional versus bidirectional Markov chains, we found that inferring multiple transition rates requires fewer samples than inferring a single transition for nearly all values of θ. This trend is likely a result of the increase in the inference problem's dimension: convergence in the multi-dimensional inference algorithm is measured using covariance as opposed to variance, and minimizing the determinant of the covariance can be achieved both by minimizing the individual variances and by maximizing the correlations between the two rate estimations.

(c) Multi-dimensional inference on complex chains

(i) Inferring birth and death rates in an M/M/1 queuing process

We start by considering an M/M/1 queue (birth–death process) Markov chain $X(t) \in \{0, 1, 2, \ldots\}$ ($m \to \infty$), schematized in figure 4*a*, with birth rates $\lambda \geq 0$ and death rates $\mu > 0$, with the restriction $\mu > \lambda$ for boundedness, so that

$$\begin{aligned}
\Pr(X(t+\delta t) = i+1 | X(t) = i) &= \lambda \delta t + o(\delta t), && \forall i \geq 0,\\
\Pr(X(t+\delta t) = i-1 | X(t) = i) &= \mu \delta t + o(\delta t), && \forall i \geq 1,\\
\Pr(X(t+\delta t) = i | X(t) = i) &= 1 - (\lambda+\mu)\delta t + o(\delta t), && \forall i \geq 0.
\end{aligned}$$

One can show (see [36] for details) that the transition probabilities $p(X(t_n) = j | X(t_{n-1}) = i, \lambda, \mu)$ for this M/M/1 queue are given by the formula involving the time interval $\Delta t = t_n - t_{n-1}$ and states i and j:

$$\begin{aligned}
p(X(t_n) = j | X(t_{n-1}) = i, \lambda, \mu) &= \mathrm{e}^{-(\lambda+\mu)\Delta t}\\
&\times \left\{ \rho^{(j-i)/2} I_{j-i}(a\Delta t) + \rho^{(j-i-1)/2} I_{j+i+1}(a\Delta t) + (1-\rho)\rho^j \sum_{k=j+i+2}^{\infty} \rho^{-(k/2)} I_k(a\Delta t) \right\}, \qquad (3.3)
\end{aligned}$$

where $\rho = \lambda/\mu < 1$ by the condition placed on the birth/death rates, I_k is the modified Bessel function of the first kind and $a = 2\sqrt{\lambda\mu}$. By substituting equation (3.3) into equation (2.8), we can proceed with inferring the transition rates λ and μ using algorithm 3. As mentioned, restricting possible transition rates so that $\mu > \lambda$ guarantees a well-defined and finite stationary distribution for the Markov chain. We accomplish this restriction by drawing the birth λ and death μ rates from a truncated version of the bivariate gamma prior from equation (3.2) which ensures death rates are always larger.

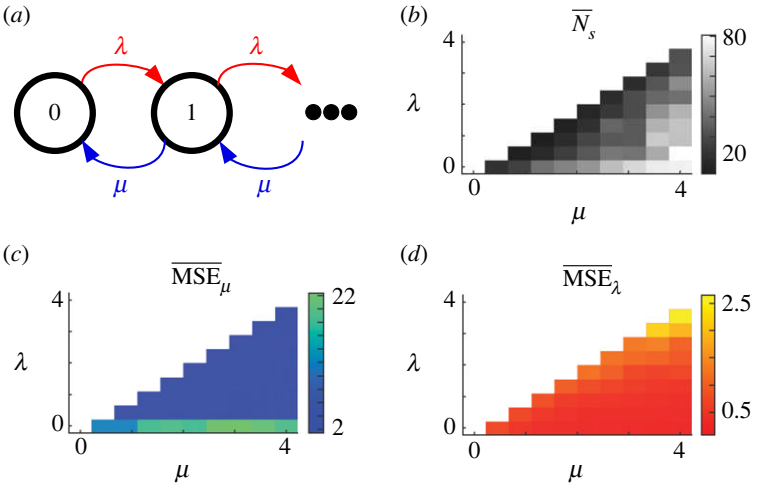

Figure 4. Inferring transition rates in a basic queueing process. (*a*) Schematic of an M/M/1 queue process with 'birth rates' λ and 'death rates' μ. (*b*) Average number of samples $\overline{N}_s$ required for the adaptive algorithm to infer fixed transition rates μ, λ, with restriction $\lambda < \mu$. Average taken over 10^2 realizations with the same pair of transition rates. (*c*) Mean-squared error $\overline{\text{MSE}}$ estimating μ, inferred using the adaptive algorithm for fixed transition rates, taken over the same samples as in (*b*). (*d*) Same as (*c*), but for $\overline{\text{MSE}}$ estimating λ.

To determine how adaptive inference fares in specifying the birth and death rates of this countably infinite Markov chain, we again quantify estimation error and convergence time. Across all transition rates considered, the adaptive algorithm quickly converges to an accurate estimate of the transition rates (figure 4*b*). As in the simple networks discussed earlier, the algorithm converges slower when the transition rates are in the tails of the prior. In addition, as in the two-state network (figure 2*e*), the algorithm has poor accuracy for inferring μ when $\lambda = 0$ (figure 4*c*). For this pure death process, the chain will always eventually converge to the absorbing state $X = 0$ for all $\mu > 0$, yielding little information about the magnitude of μ itself. However, unlike the two-state network (figure 2*f*), the algorithm had very low error in inferring λ for all values considered (figure 4*d*). The adaptive algorithm is effective in inferring the parameters of this birth–death process, particularly when the two parameters have similar value but the death rate is still larger than the birth rate.

(ii) Inferring transition rates in a ring network

Every chain we have considered so far has shared a key feature: we can introduce an absorbing state in the chain by setting one of the transition rates to zero. Cutting a single link of the chain causes one state to have no link out of it. This can lead to high errors in inference for both the two-state (figure 2*e*,*f*) and M/M/1 queue (figure 4*c*) networks. To move away from these cases, we consider a ring chain with identical clockwise transition rates h_+ and counterclockwise transition rates h_- (figure 5*a*). This chain, which is equivalent to a periodic random walk, possesses absorbing states only if h_+ and h_- are identically zero, so we can further test if adaptive inference error increases due to the reduction in a problem's dimension. For such a ring network with m states, the transition probabilities from state j to i given a time interval and the clockwise and counterclockwise transition rates, $p(X(t_n) = j | X(t_{n-1}) = i, h_+, h_-)$, are given by the matrix exponential [37]

$$p(X(t_n) = j | X(t_{n-1}) = i, h_+, h_-) = [e^{\mathbf{A}(h_+, h_-)\Delta t}]_{ij}, \tag{3.4}$$

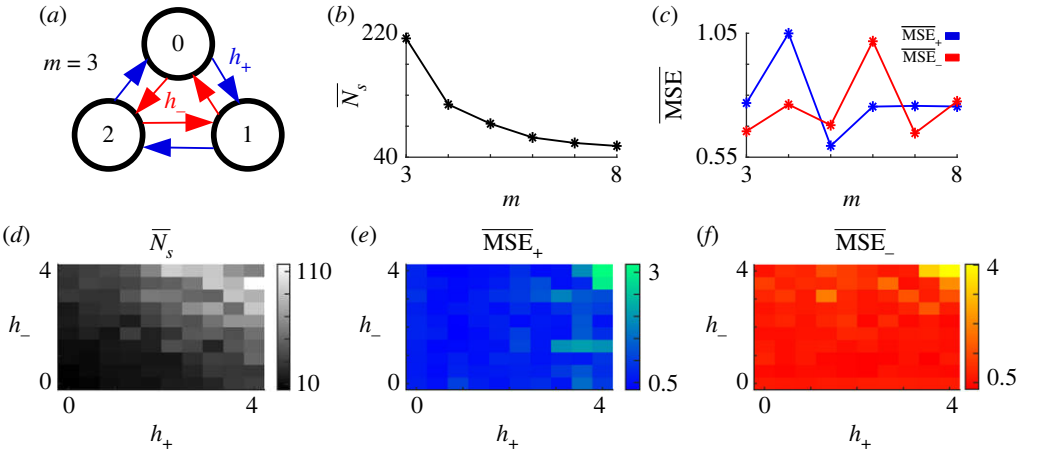

Figure 5. Inferring transition rates in a symmetric ring network. (*a*) Schematic of a ring network of size $m = 3$ with clockwise transition rates h_+ and counterclockwise transition rates h_-. (*b*) Average number of samples $\overline{N}_s$ required for the adaptive algorithm to infer transition rates given a ring of size m. Averages taken over 10^2 sampled pairs of transition rates generated from the bivariate gamma prior, equation (3.2). (*c*) Mean-squared error $\overline{\text{MSE}}$ in estimating the transition rates of the ring chain using the adaptive algorithm for varied network size m and the same network realizations from (*b*). (*d*) Mean number of samples $\overline{N}_s$ required for the adaptive algorithm's estimate of the transition rates to converge in a network of size $m = 8$ for a fixed pair (h_-, h_+) of true counterclockwise and clockwise transition rates. Averages at each parameter set value taken over 10^2 trials. (*e*) Mean-squared error $\overline{\text{MSE}}$ in the estimate of h_+ as the two transition rates are varied. Averages taken using the same trials as in (*d*). (*f*) Same as (*e*), but for $\overline{\text{MSE}}$ estimating h_-.

where $\Delta t = t_n - t_{n-1}$ and $\mathbf{A} \in \mathbb{R}^{m\times m}$ is the infinitesimal generator matrix for the network. In the case of an m-state ring network, $\mathbf{A}(h_+, h_-)$ is given by

$$\mathbf{A} = \begin{pmatrix} -(h_+ + h_-) & h_+ & 0 & \dots & 0 & h_- \\ h_- & -(h_+ + h_-) & h_+ & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & & \\ \vdots & & & & & \vdots \\ & & & & & 0 \\ \vdots & \dots & 0 & h_- & -(h_+ + h_-) & h_+ \\ h_+ & 0 & \dots & 0 & h_- & -(h_+ + h_-) \end{pmatrix}.$$

We were first interested in how the size of the Markov chain affected the error in inference and the time for the algorithm to converge. Recall that inference of the transition rate parameters converged faster in the M/M/1 queue chain than in the simple two-state chain, which we attributed to an increase in the number of chain states (i.e. increase in the possible measurements), allowing for a more refined sampling of the stochastic dynamics of the chain with each observation. We looked to see if this trend extended to the context of a chain with ring topology, computing the convergence time (figure 5*b*) and error (figure 5*c*) in transition rate estimation as a function of chain size. The results suggest that increasing the chain size does in fact speed up convergence of the adaptive algorithm. Moreover, this speed-up does not generate any additional error in rate inference, as the adaptive algorithm displays uniformly low error for all chain sizes considered.

We also looked further at how strong asymmetries in the true transition rate parameters impacted the performance of the adaptive algorithm on the rate inference problem on the ring. To do so, we fixed the ring size $m = 8$ and measured convergence time (figure 5*d*) and inference errors of each transition rate (figure 5*e*,*f*) for different fixed pairs of h_+ and h_-. As we observed in all

previous chains, the adaptive algorithm takes longer to converge when the true transition rates are large and fall in the tails of the bivariate gamma prior (equation (3.2)). Unlike the previous examples we studied, the average inference error for both transition rates is uniformly low across the range of values considered. These findings provide further credence to our speculation that the appearance of absorbing states in Markov chains can drastically increase the error in rate parameter inference. The only way for the ring chain to become absorbing would be for both transition rates to be identically zero, causing the system to be frozen in the initial state from the start. Otherwise, the observations of the Markov chain are guaranteed to span the entire state space and continually provide new information about both transition rates to the adaptive algorithm. In addition, because the only absorbing network occurs when $h_+ = h_- = 0$, the algorithm quickly infers this configuration by obtaining repeated measurements of the same state, so even these parameters can be rapidly inferred to high precision.

(iii) Inferring network structure

As a final test of the adaptive algorithm, we considered the problem of inferring the structure of a Markov chain with a strongly restrictive prior on the rates, requiring that they are either 0 or 1. Doing so isolates the problem of identifying the presence or absence of a link in the Markov chain without the further problem of inferring the amplitude of the rate. Thus, each transition rate is a binary variable drawn independently from the set $\{0, 1\}$ with Bernoulli parameter p (figure 6*a*). In this way, we reduce the set of possible chain link conformations by only allowing for one possible non-zero value for all transition rates. The transition probabilities $p(X(t_n) = j | X(t_{n-1}) = i, \mathbf{h})$ for these networks are given by the same matrix exponential as in the case of ring chains, described in equation (3.4), where the infinitesimal generator matrix $\mathbf{A}$ is changed to reflect the specific chain's structure. To measure our algorithm's performance on this class of chains, we compute the average number of samples required to converge and, due to the binary prior over the transition rates, measure inference using MAE, given by a normalized L_1 error:

$$\mathrm{MAE} = \frac{\sum_{k=1}^{d_{\max}} |h_k - \hat{h}_k|}{d_{\max}}, \tag{3.5}$$

where $d_{\max} = m(m-1)$ is the maximum number of possible non-zero transition rates for a chain of size m, h_k is the true value of the k-th transition rate and $\hat{h}_k$ is the maximum *a posteriori* estimate of the kth transition rate. To account for the fact that the initial determinant of the covariance matrix changes as $d_{\max}$ increases, we modified the convergence threshold to depend on this initial covariance determinant. For example, if the initial determinant for a simulation was D, we ran our adaptive algorithm until the covariance determinant was less than θD. We took $\theta = 10^{-2}$ for all simulations on these binarized networks.

How does adaptive inference handle this problem? For a fixed chain size, inferring the structure is faster when the chain is more disconnected (figure 6*b*). More connected networks provide a higher diversity of possible measurements, increasing the possible number of network configurations that may generate those measurements. Note that because these networks have binary transition rates, our algorithm does not have to infer the magnitude of a non-zero transition rate and therefore avoids the errors shown in figure 2*e*,*f*. In addition, inference error heavily depends on both the true connectivity of the network and the prior likelihood over each transition rate (figure 6*c*). However, error is lowest when the true connectivity is aligned with the prior (i.e. when both $d/d_{\max} < 0.5$ and $p < 0.5$ or $d/d_{\max} > 0.5$ and $p > 0.5$), and increases when the two are mismatched (e.g. when p is closer to 1 but d is closer to 0). For a fixed prior, rate estimate performance displays similar behaviours across different chain sizes if there are more than two states: the algorithm converges faster when chains are more disconnected (figure 6*d*), and error is lower when the chain structure and the prior agree (figure 6*e*). These performance trends are consistent with our findings from applying adaptive inference to other Markov chain models: transition rates in the tails of the prior or that are highly asymmetric take more measurements to infer, but generally have low inference error. When some transition rates are set to zero,

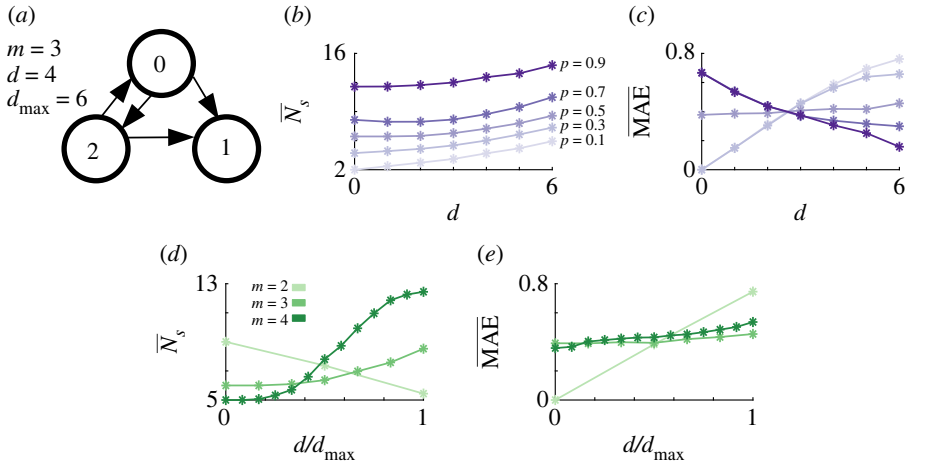

Figure 6. Inferring structure of Markov chains with rates drawn from binary sets. (*a*) Schematic of a sample binary chain of size $m = 3$ with $d = 4$ non-zero transition rates. All transition rates are each independently chosen from the set {0, 1} with Bernoulli parameter p (see text for details). (*b*) Average number of samples $\overline{N}_s$ needed for the adaptive algorithm to infer the transition rates to the required degree of accuracy, defined by equation (3.5), as a function of the number of non-zero transition rates d with fixed network size $m = 3$. Averages taken over 10^3 sampled network structures with independent Bernoulli parameter p; several values of p are superimposed (labelled). (*c*) Average inference error, measured using mean absolute error ($\overline{\text{MAE}}$, L_1-error), for varied d. Averages computed using the same trials as in (*b*). (*d*) Average number of samples $\overline{N}_s$ required for the adaptive algorithm to converge to a set accuracy for different chain sizes (legend). Averages taken over 10^3 sampled network structures with fixed Bernoulli prior $p = 0.5$. (*e*) Mean absolute error $\overline{\text{MAE}}$ to which the adaptive algorithm converges as the density of links in the chain is increased for several different network sizes. Averages computed using the same data from (*d*).

creating absorbing states in the chain, inferring the existence of non-zero transition rates is less difficult than inferring the magnitude of those rates. However, across a range of chain structures and transition rate magnitudes, adaptive inference is able to rapidly and accurately infer state transition dynamics.

4. Conclusion

In this work, we developed a simple algorithm to infer transition rates of arbitrary discrete-state Markov processes that determines optimal sampling times to minimize a posterior covariance. Starting with small chains made up of two states, we found that using a previously developed adaptive algorithm by [27] had lower error than a naive algorithm that samples with a fixed period. Because of the simplicity of our approach, we showed how to extend the adaptive algorithm to infer generic structures of transition rates, where sample times are chosen to minimize the determinant of the posterior covariance matrix. Applying this extension to more complex Markov chains, we found that the adaptive algorithm rapidly converged to accurately estimate rate parameters when the true chain structure was more likely according to the prior. When the chain link conformation was less likely according to the prior, the adaptive algorithm still converged fairly quickly, but inference error was higher for one or more of the transition rates.

Our work builds on previous development of sequential Bayesian methods for experimental design, harnessing stepwise posterior updates to guide design parameter choices. In addition to the work in traditional experimental design, our framework is also closely related to standard approaches used in active learning [38], and leverages advances in classic Bayesian optimization problems [39]. However, when inferring complex models, these fields are often faced with likelihood functions that quickly become intractable and make posterior updates challenging, much like the problems facing the experimental design studies we have previously mentioned.

Here, we demonstrate that for the large class of models that can be described using Markov chains, adaptive Bayesian design can be implemented to great effect, has analytically tractable observational likelihoods and does not require specialized, computational techniques to perform posterior updates. This moves considerably beyond a recent study of adaptive sampling for Markov chains [20], which would update the sampling time after several samples rather than each time, and only ensured asymptotic equivalence to optimal fixed time designs.

While we only considered Markovian networks with constant transition rates, our posterior-covariance-minimization approach can theoretically be extended to transition rates of arbitrary functional forms. These extensions could prove useful for inferring transition rates in stochastic systems with variable rates, as found in chemical kinetic systems modelled by M/G/1 queueing processes [40] and stochastic implementations of Hodgkin–Huxley neuronal dynamics with voltage-dependent transition rates [41,42]. Our adaptive inference algorithm can also be adapted to more complex chain structures, such as those used in age-structured epidemiological models [43] and models for chaperone-assisted protein folding [44]. The only hard constraints to our approach are that the possible transition functions be fully specified and the chain itself be Markovian. However, our algorithm cannot escape the curse of dimensionality for more complex chains. For a Markov chain with d distinct transition rates (or equivalently, d parameters that specify the transition rate functions) and a numerical discretization that allows each transition rate to take on n possible values, the size of the posterior grows as n^d. Future work utilizing our algorithm for more complex transition rate inference problems would necessitate efficient matrix methods or posterior approximation techniques.

Data accessibility. For the MATLAB code used to generate all results and figures, see https://github.com/nwbarendregt/AdaptMarkovRateInf.

Authors' contributions. N.W.B.: conceptualization, formal analysis, software, visualization, writing—original draft and writing—review and editing; E.G.W.: conceptualization, software and writing—review and editing; Z.P.K.: conceptualization, funding acquisition, supervision and writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. This work was supported by CRCNS/NIH R01-MH-115557, NIH R01-EB029847-01, and NSF- DMS-1853630.

References

1. Lindley DV. 1956 On a measure of the information provided by an experiment. *Ann. Math. Stat.* **27**, 986–1005. (doi:10.1214/aoms/1177728069)
2. Nishii R. 1993 Optimality of experimental designs. *Discrete Math.* **116**, 209–225. (doi:10.1016/0012-365X(93)90402-F)
3. Johnson RT, Montgomery DC, Jones BA. 2011 An expository paper on optimal design. *Qual. Eng.* **23**, 287–301. (doi:10.1080/08982112.2011.576203)
4. Chaloner K, Verdinelli I. 1995 Bayesian experimental design: a review. *Stat. Sci.* **10**, 273–304. (doi:10.1214/ss/1177009939)
5. Ryan EG, Drovandi CC, McGree JM, Pettitt AN. 2016 A review of modern computational algorithms for Bayesian optimal design. *Int. Stat. Rev.* **84**, 128–154. (doi:10.1111/insr.12107)
6. Bandi FM, Russell JR. 2008 Microstructure noise, realized variance, and optimal sampling. *Rev. Econ. Stud.* **75**, 339–369. (doi:10.1111/j.1467-937X.2008.00474.x)
7. Ehrenfeld S. 1962 Some experimental design problems in attribute life testing. *J. Am. Stat. Assoc.* **57**, 668–679. (doi:10.1080/01621459.1962.10500555)
8. Huan X, Marzouk YM. 2013 Simulation-based optimal Bayesian experimental design for nonlinear systems. *J. Comput. Phys.* **232**, 288–317. (doi:10.1016/j.jcp.2012.08.013)
9. Dushenko S, Ambal K, McMichael RD. 2020 Sequential Bayesian experiment design for optically detected magnetic resonance of nitrogen-vacancy centers. *Phys. Rev. Appl.* **14**, 054036. (doi:10.1103/PhysRevApplied.14.054036)
10. Myung JI, Cavagnaro DR, Pitt MA. 2013 A tutorial on adaptive design optimization. *J. Math. Psychol.* **57**, 53–67. (doi:10.1016/j.jmp.2013.05.005)

11. Cavagnaro DR, Myung JI, Pitt MA, Kujala JV. 2010 Adaptive design optimization: a mutual information-based approach to model discrimination in cognitive science. *Neural Comput.* **22**, 887–905. (doi:10.1162/neco.2009.02-09-959)
12. Cook AR, Gibson GJ, Gilligan CA. 2008 Optimal observation times in experimental epidemic processes. *Biometrics* **64**, 860–868. (doi:10.1111/j.1541-0420.2007.00931.x)
13. Ross J, Pagendam D, Pollett P. 2009 On parameter estimation in population models II: multi-dimensional processes and transient dynamics. *Theor. Popul. Biol.* **75**, 123–132. (doi:10.1016/j.tpb.2008.12.002)
14. Ferguson JM, Langebrake JB, Cannataro VL, Garcia AJ, Hamman EA, Martcheva M, Osenberg CW. 2014 Optimal sampling strategies for detecting zoonotic disease epidemics. *PLoS Comput. Biol.* **10**, e1003668. (doi:10.1371/journal.pcbi.1003668)
15. Becker G, Kersting G. 1983 Design problems for the pure birth process. *Adv. Appl. Probab.* **15**, 255–273. (doi:10.2307/1426436)
16. Pagendam D, Pollett P. 2010 Locally optimal designs for the simple death process. *J. Stat. Plann. Inference* **140**, 3096–3105. (doi:10.1016/j.jspi.2010.04.017)
17. Bhatt DL, Mehta C. 2016 Adaptive designs for clinical trials. *N. Engl. J. Med.* **375**, 65–74. (doi:10.1056/NEJMra1510061)
18. Mehtälä J, Auranen K, Kulathinal S. 2015 Optimal designs for epidemiologic longitudinal studies with binary outcomes. *Stat. Methods Med. Res.* **24**, 803–818.
19. Chernoff H. 1959 Sequential design of experiments. *Ann. Math. Stat.* **30**, 755–770. (doi:10.1214/aoms/1177706205)
20. Michel J. 2020 Optimal adaptive sampling for a symmetric two-state continuous time Markov chain. *Econometric Rev.* **39**, 602–611. (doi:10.1080/07474938.2019.1701808)
21. Pagendam D, Pollett P. 2009 Optimal sampling and problematic likelihood functions in a simple population model. *Environ. Model. Assessm.* **14**, 759–767. (doi:10.1007/s10666-008-9159-1)
22. Huan X, Marzouk Y. 2014 Gradient-based stochastic optimization methods in Bayesian experimental design. *Int. J. Uncertain. Quantification* **4**, 479–510. (doi:10.1615/Int.J.UncertaintyQuantification.2014006730)
23. Drovandi CC, Pettitt AN. 2013 Bayesian experimental design for models with intractable likelihoods. *Biometrics* **69**, 937–948. (doi:10.1111/biom.12081)
24. Drovandi CC, McGree JM, Pettitt AN. 2014 A sequential Monte Carlo algorithm to incorporate model uncertainty in Bayesian sequential design. *J. Comput. Graph. Stat.* **23**, 3–24. (doi:10.1080/10618600.2012.730083)
25. Rainforth TWG. 2017 *Automating inference, learning, and design using probabilistic programming*. PhD thesis. Oxford: University of Oxford.
26. Foster A, Ivanova DR, Malik I, Rainforth T. 2021 Deep adaptive design: amortizing sequential Bayesian experimental design. In *Proc. of the 38th International Conference on Machine Learning*, Online, 18-24 July 2021. (http://arxiv.org/abs/2103.02438)
27. Webb EG. 2021 Bayesian inference of Markov transition rates. Master's thesis. Boulder, CO: University of Colorado Boulder.
28. Wimmer K, Nykamp DQ, Constantinidis C, Compte A. 2014 Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nat. Neurosci.* **17**, 431–439. (doi:10.1038/nn.3645)
29. Panichello MF, DePasquale B, Pillow JW, Buschman TJ. 2019 Error-correcting dynamics in visual working memory. *Nat. Commun.* **10**, 1–11. (doi:10.1038/s41467-019-11298-3)
30. Jones B, Allen-Moyer K, Goos P. 2020 A-optimal versus D-optimal design of screening experiments. *J. Qual. Technol.* **53**, 1–14.
31. de Aguiar PF, Bourguignon B, Khots M, Massart D, Phan-Than-Luu R. 1995 D-optimal designs. *Chemom. Intell. Lab. Syst.* **30**, 199–210. (doi:10.1016/0169-7439(94)00076-X)
32. Kiefer J. 1953 Sequential minimax search for a maximum. *Proc. Am. Math. Soc.* **4**, 502–506. (doi:10.1090/S0002-9939-1953-0055639-3)
33. Smid SC, McNeish D, Miočević M, van de Schoot R. 2020 Bayesian versus Frequentist estimation for structural equation models in small sample contexts: a systematic review. *Struct. Equ. Model.: A Multidiscipl. J.* **27**, 131–161. (doi:10.1080/10705511.2019.1577140)
34. Kass RE, Raftery AE. 1995 Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795. (doi:10.1080/01621459.1995.10476572)
35. Nadarajah S, Gupta AK. 2006 Some bivariate gamma distributions. *Appl. Math. Lett.* **19**, 767–774. (doi:10.1016/j.aml.2005.10.007)

36. Gross D, Harris CM. 1998 *Fundamentals of queueing theory*, 3rd edn. New York, NY: John Weiley and Sons.
37. Taylor H, Karlin S. 1984 *An Introduction to Stochastic Modeling, Aca*. Orlando, FL: Academic Press, Inc.
38. MacKay DJ. 1992 Information-based objective functions for active data selection. *Neural Comput.* **4**, 590–604. (doi:10.1162/neco.1992.4.4.590)
39. Shahriari B, Swersky K, Wang Z, Adams RP, De Freitas N. 2015 Taking the human out of the loop: a review of Bayesian optimization. *Proc. IEEE* **104**, 148–175. (doi:10.1109/JPROC.2015.2494218)
40. Anderson DF, Kurtz TG. 2015 *Stochastic analysis of biochemical systems*, vol. 674. Berlin, Germany: Springer.
41. Pu S, Thomas PJ. 2020 Fast and accurate Langevin simulations of stochastic Hodgkin-Huxley dynamics. *Neural Comput.* **32**, 1775–1835. (doi:10.1162/neco_a_01312)
42. Pu S. 2021 *Noise decomposition for stochastic Hodgkin-Huxley Models*. PhD thesis. Cleveland, OH: Case Western Reserve University.
43. Zhang W, Zhang C, Bi Y, Yuan L, Jiang Y, Hasi C, Zhang X, Kong X. 2021 Analysis of COVID-19 epidemic and clinical risk factors of patients under epidemiological Markov model. *Results Phys.* **22**, 103881. (doi:10.1016/j.rinp.2021.103881)
44. Ilker E, Chiel J, Deffner S, Hinczewski M. 2021 Shortcuts in Stochastic Systems and Control of Biophysical Processes. *Phys. Rev. X* **12**, 021048. (http://arxiv.org/abs/2106.07130)