Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC-PapersOnLine 49-26 (2016) 089–094

Preconditioned Metropolis sampling as a strategy to improve efficiency in Posterior exploration ⋆

Stefan Engblom * **Vikram Sunkara** **

* *Division of Scientific Computing, Department of Information Technology, Uppsala University, SE-751 05 Uppsala, Sweden.(email:stefane@it.uu.se)*

** *Computational Systems Biology, Konrad-Zuse-Zentrum for Informationstechnik, D-10557 Berlin, Germany. (email: sunkara@zib.de)*

Abstract: In the low copy number regime, the dynamics of chemically reacting systems is accurately modeled as a continuous-time Markov chain and the associated probability density obeys the chemical master equation. Parameter inference in such models is very challenging for various reasons: large levels of noise implies that large amount of data is required for identification, the presence of transient phases may shadow subsets of the parameters, and accurate likelihood estimation requires the solutions to master equations. The latter is itself a computational very challenging problem and although many approximate computational methods have been proposed previously, the final implied accuracy in estimated rate parameters is difficult to assess.

In this paper we look at the problem from the perspective of the Markov chain Monte Carlo method. Assuming the existence of a practically exact, but expensive, master equations solver, together with a cheaper, approximate alternative, we pick up the idea of *preconditioned Metropolis sampling*. Here the solutions of full master equations almost always imply an accepted step in the Markov chain, and consequently, step rejections are much cheaper. We investigate the properties of this technique theoretically and via illustrative examples. Whenever a suitable preconditioner is available, large savings in computational times are possible while the accuracy in deduced parameters is identical to using the exact likelihood.

Keywords: Chemical Master Equation, Preconditioning Markov Chain Monte Carlo, Metropolis Hastings.

1. INTRODUCTION

Systems biology is working towards describing complex interactions and processes in biology by Bio-Chemical Reaction Networks (Biological Networks for short). Through advances in imaging and sequencing technologies, biologists are able to scope deeper and describe critical biological processes as paths on a complex biological network. Current systems biology has been able to represent metabolic processes of cells Kholodenko (2000), stem cell fate paths MacArthur et al. (2009), gene transcriptions Srivastava et al. (2002) and translation regulation processes as biological networks. Blake et al. (2003) In the last three decades it was shown by experimentalists that the observed variation in the data can be attributed to the inherent stochasticity in parts of the network where low copy numbers are present.

The modeling of biological networks with intrinsic stochasticity gives realistic forecasts of the system behavior. However, to go further and use experimental data to infer the reaction rates of the underlying model is a computationally and mathematically challenging task. The critical aspect is the computation of the *likelihood function*. The likelihood function describes the conditional probability of observing the data for a given parameter value. In engineering settings where the noise is understood to be Gaussian and additive, the likelihood function can be easily approximated by a mean dynamics of the system and some fixed variance. However, if the system has intrinsic stochasticity, then computing the likelihood requires the solutions to equations such as the *Chemical Master Equation* (CME) or the *Fokker Planck equation*. These equations are difficult to solve numerically as they are prone to the *curse of dimensionality* Higham (2008); Engblom (2009a).

While many numerical methods have been investigated to solve the CME for given set of parameters, the accuracy of these methods for the purpose of inferring parameters from some given data is unclear. Let us see why: let $\mathcal{L}_{CME}(O|\sigma)$ be the likelihood function, given by the solutions of the CME, that data O is observed for the parameter σ. Similarly, let $\mathcal{L}_\star(O|\sigma)$ denote an approximate likelihood constructed using some approximation of the CME. We

⋆ The work was supported by the Swedish Research Council within the UPMARC Linnaeus center of Excellence (S. Engblom) and by BMBF (Germany) project PrevOp-OVERLOAD, grant number 01EC1408H (V. Sunkara).

10.1016/j.ifacol.2016.12.108

say that the likelihood $\mathcal{L}_\star$ is biased if

$$\|\mathcal{L}_{CME}(O|\sigma) - \mathcal{L}_\star(O|\sigma)\| = f(\sigma), \tag{1}$$

where $f(\cdot)$ is not a constant function. For many numerical approximations of the CME the function f is unfeasible to compute. With this in mind, when we use a biased likelihood for finding a posterior distribution, there is no simple rule on how the error in the likelihood translates across into the error of the posterior. For scenarios when we have a biased likelihood and computing f in (1) is unfeasible, one must get samples from the true posterior to verify the accuracy of the posterior constructed by the biased likelihood. Computing approximations of the CME which have a uniform error over the parameter space is computationally demanding. This motivates the proposal for a *preconditioning* MCMC algorithm (pcMCMC) Efendiev et al. (2006) [1] . In summary, the pcMCMC has two proposal steps in series for the same proposed state. The first proposal step uses the biased likelihood to accept or reject the new state. If the state is accepted in the first proposal step, then that state needs to be accepted in the second proposal step; this then involves the computation of a likelihood made up of unbiased CME approximations. The state which is accepted by both proposals in series is a true sample of the posterior distribution of the unbiased likelihood function.

The principle idea being that if the biased likelihood is close to the unbiased likelihood (locally), then the acceptance rate of the second proposal should be higher than the first. Since the CMEs are only being computed in the second proposal, having a higher acceptance rate would imply that we are minimizing the amount of CMEs we need to compute to find true samples of the posterior distribution. The acceptance rate of the second proposal is also an indicator of how close the biased likelihood is to the unbiased likelihood in some local region of the parameter space. In the sections below, we introduce the CME and the parameter inference problem. Then, we give an overview of the preconditioning MCMC method followed by some illustrative examples.

2. CHEMICAL MASTER EQUATION

The population of $N_s \in \mathbb{N}$ species undergoing $N_r \in \mathbb{N}$ reactions is described by the following sum of inhomogeneous Poisson processes,

$$X_\sigma(t) := X(0) + \sum_{r=1}^{N_r} \mathcal{P}_r\left(\int_0^t \alpha_r(X_\sigma(s),\sigma)\right)\rho_r. \tag{2}$$

The state space of $X_\sigma(t)$ is denoted by Ω and is a subset of $\mathbb{N}_0^{N_s}$. The variable σ is an element of our parameter space $\Sigma \subset \mathbb{R}_+^{N_r}$. The function $\alpha_r : \Omega \times \Sigma \to [0,\infty)$ describes the propensity/intensity at which the r^{th} reaction occurs. The stoichiometric vector, $\rho_r \in \mathbb{M}_{N_s\times 1}$, gives the change in population induced by the r^{th} reaction. Many biological processes' populations are described by (2). Historically, Thomas Kurtz investigated the convergence and analysis of (2) applied to the field of stochastic epidemiology, and for this reason we refer to (2) as the *Kurtz process* Ethier and Kurtz (2009).

[1] also refereed to as delayed acceptance by the statistics community Golightly et al. (2015)

To find the probability of observing $X_\sigma(t)$ in a state $x \in \Omega$ at a time point t, we need to substitute the Kurtz process into the Chapman–Kolmogorov equation. This will lead to the evolution of the probability over the state space being governed by the *Chemical Master Equation* (CME),

$$\frac{\partial P_\sigma(x;t)}{\partial t} = \sum_{r=1}^{N_r} \alpha_r(x-\rho_r,\sigma)P_\sigma(x-\rho_r;t) - \sum_{r=1}^{N_r}\alpha_r(x,\sigma)P_\sigma(x;t). \tag{3}$$

Verbosely, the change in probability of observing a state x at time t is equal to the transition probability of coming from an adjacent state into x, minus the transition probability of leaving the state x . The CME can be solved by formulating a linear initial value problem (IVP):

$$\frac{dp_\sigma(t)}{dt} = \mathcal{A}_\sigma p_\sigma(t) \quad \text{i.c.} \quad p(0), \tag{4}$$

where $p_\sigma(t)$ is a vector indexed by states in the state space and $\mathcal{A}_\sigma$ is an infinitesimal generator with columns summing to zero.

Broadly speaking, there are three major strategies for numerically approximating the solution to (4): Domain reduction, Galerkin methods and Tensor decomposition. An example of a domain reduction method is aggregation, where the idea is to aggregate states where the distribution has a shallow spatial gradient reducing the number of equations to solve Munsky and Khammash (2006); Sunkara and Hegland (2010). In Galerkin methods, the distribution is projected on a finite dimensional Hilbert space spanned by a chosen set of basis functions. This changes the IVP of the probability distribution to an IVP of the weights of the basis representation of the probability distribution Engblom (2009b); Jahnke and Udrescu (2010). Like in the continuous Galerkin methods, if the distribution has some inherent regularity, choosing the right basis functions can give a significantly smaller IVP to solve. Lastly, a Tensor decomposition method exploits possible tensor structure of the infinitesimal generator $\mathcal{A}_\sigma$ Kazeev et al. (2014). It has been shown that representing the generators of particular systems in a tensor format can partially overcome the curse of dimensionality. All methods exploit some inherent structure to achieve a significant speed-up over constructing an empirical distribution using trajectory based methods. Since our focus is not on any particular solver, we use the notational convention of writing a $\star$ when we are referring to an arbitrary approximation method of the CME. Similarly, we denote $P_\sigma^\star(x;t)$ as the corresponding solution to the approximate CME.

2.1 Random Variables and Likelihood

In this section we familiarize ourselves with the notational convention that will be used through this paper describing random variables and their likelihoods. We begin with the notational convention for the parameters we wish to infer. As in earlier sections, we denote the parameter to be inferred as $\sigma \in \Sigma$. The set Σ is a closed subset of $\mathbb{R}_0^{N_p}$, where N_p is the number of parameters we are to infer.

Let O_n^t (the data) be the n^{th} *random variable* governed by the stochastic process $X_\sigma(t)$ with probability $p_\sigma(t)$. For simplicity we assume data from only two time points, that is, snapshot data. The term O_0^0 denotes the initial

value. We denote the likelihood of observing O_n^t, for $n = 1, \dots, N_o$, at time t for a parameter set σ to be,

$$\mathcal{L}_{CME}(O_1^t, \dots, O_{N_o}^t \mid \sigma) := \prod_{n=1}^{N_o} P_\sigma\left((O_n^t; t)|\ O_0^0\right). \quad (5)$$

Similarly the approximate likelihood function constructed with the approximate CME is given by,

$$\mathcal{L}_\star(O_1^t, \dots, O_{N_o}^t \mid \sigma) := \prod_{n=1}^{N_o} P_\sigma^\star\left((O_n^t; t)|\ O_0^0\right). \quad (6)$$

Given a likelihood, using Bayes' theorem, we can compute the posterior distributions, that is, the probability of the parameter given the data. We assume our prior is from a uniform distribution. We denote

$$\pi_{CME}(\sigma \mid O_1^t, \dots, O_{N_o}^t) = \frac{\mathcal{L}_{CME}(O_1^t, \dots, O_{N_o}^t \mid \sigma)}{C_{CME}}, \quad (7)$$

$$\pi_\star(\sigma \mid O_1^t, \dots, O_{N_o}^t) = \frac{\mathcal{L}_\star(O_1^t, \dots, O_{N_o}^t \mid \sigma)}{C_\star}, \quad (8)$$

to be the posterior distributions of the CME based likelihood and the approximate likelihood, respectively. Since the normalizing constants C_{CME} and $C_\star$ are not known, we use the Metropolis–Hastings algorithm (MH) to sample the posteriors Metropolis et al. (1953); Golightly and Wilkinson (2005). An overview of the basic MH algorithm step can be described as following: Let σ_1 and σ_2 be two parameters in Σ. Let $Q(\cdot|\cdot)$ be the proposal distribution over $\Sigma \times \Sigma$ and $\mathcal{L}(\cdot)$ the likelihood function. Given we start at σ_1, the acceptance probability of moving to σ_2 is given by

$$\min\left(1.0, \frac{Q(\sigma_1|\sigma_2)\mathcal{L}(\sigma_2)}{Q(\sigma_2|\sigma_1)\mathcal{L}(\sigma_1)}\right).$$

Now we introduce the pcMCMC method, for an in-depth study into the method please refer to Efendiev et al. (2006). The pcMCMC algorithm has two proposal steps in series. The new proposed state has to be accepted in both steps before it is accepted as a new state in the chain. In summary, the first acceptance probability is computed using a biased likelihood. Once the new state is accepted here, then in the second step the acceptance probability is computed using an unbiased likelihood.

A single acceptance step of the pcMCMC can be written down as follows: Let σ_1 be the current state of the Markov chain and σ_2 the proposed transition state. Let $\tau_1, \tau_2 \sim \texttt{Uniform}(0, 1)$.

(P1) *Accept* σ_2 in the first proposal step if

$$\tau_1 < \min\left(1.0, \frac{Q(\sigma_1|\sigma_2)\mathcal{L}_\star(\cdot|\sigma_2)}{Q(\sigma_2|\sigma_1)\mathcal{L}_\star(\cdot|\sigma_1)}\right).$$

(P2) *Accept* σ_2 in the second proposal step if

$$\tau_2 < \min\left(1.0, \frac{\mathcal{L}_{CME}(\cdot|\sigma_2)\mathcal{L}_\star(\cdot|\sigma_1)}{\mathcal{L}_{CME}(\cdot|\sigma_1)\mathcal{L}_\star(\cdot|\sigma_2)}\right).$$

We speculate the following: If $\mathcal{L}_\star$ is close to $\mathcal{L}_{CME}$ in shape, then the much cheaper $\mathcal{L}_\star$ step should search through and filter the parameter space and propose states which we will later accept with a high probability. In turn, a good preconditioner would minimize the number of CME computations which then end in the rejection of a state. If a good preconditioner is found, then we are accepting good sample points with a very high acceptance rate Efendiev et al. (2006).

3. NUMERICAL EXPERIMENTS

The rules of choosing a good preconditioner are still unclear. We will use the following examples to investigate some different preconditioners for basic biochemical models. We set up the experiments to understand the acceptance rate of the second proposal step and construct the chains in the following manner. Multiple pcMCMC chains are run with different starting states, sampled from a uniform distribution over the parameter space. The scaling parameter is chosen such that the first proposal step has an acceptance rate of $\approx 40\%$. Each chain is allowed to take 20 acceptance steps (accepted by second proposal). Then the average acceptance rate is given by the number of accepted states in the second proposal divided by the number of accepted states in the first proposal. The averages of acceptance rates of each of the chains are then used to construct an empirical distribution of the acceptance rate. To compare two preconditioners, we compare the two empirical distributions of the acceptance rates.

3.1 Birth–Death Process

The Birth–Death process is a simple jump Markov process given by

$$X(t) = x_0 + \mathcal{P}_1(\lambda t) - \mathcal{P}_2\left(\int_0^t \beta X(s) ds\right).$$

The births are driven by a homogeneous Poisson process at rate λ and the deaths driven by an inhomogeneous Poisson processes with its rate dependent on β and the state $X(\cdot)$. Given a starting population of $x_0 \in \mathbb{N}$, the analytical solution of the probability density of observing a state $x \in \mathbb{N}_0$ at some time $t > 0$ is given by Jahnke and Huisinga (2007),

$$P_{(\lambda,\beta)}(x; t) = \sum_{k=0}^{\min(x,x_0)} \binom{x_0}{k} g(t)^k (1 - g(t))^{x_0 - k} \frac{h^{x-k}(t) e^{-h(t)}}{x - k!},$$

where $g(t) = \exp(-\beta t)$ and $h(t) = \lambda(1 - \exp(-t\beta))/\beta$. The expectation and the variance of the Birth–Death process is given by, $\mathbb{E}_{\lambda,\beta}(t|x_0) = x_0 g(t) - h(t)$ and $\mathbb{V}_{\lambda,\beta}(t|x_0) = x_0 g(t)(1 - g(t)) + h(t)$ respectively.

To demonstrate how the acceptance rate distribution changes as we vary the degree of the approximation of the likelihood, let us consider the following approximation:

$$P^\tau_{(\lambda,\beta)}(x; t) := \frac{1}{\sqrt{2\pi \mathbb{V}_{\lambda,\beta}(t|x_0)}} \exp\left(-\frac{(x - \tau \mathbb{E}_{\lambda,\beta}(t|x_0))^2}{2\mathbb{V}_{\lambda,\beta}(t|x_0)}\right).$$

Here $\tau \in \mathbb{R}_+$ is an experimental parameter we can choose to change the quality of the approximation. To produce data to infer over we used the parameters $\lambda = 40$ and $\beta = 2$ such that the steady-state distribution of the birth-death process is Poissonian with mean 20. Using SSA we generated a total of 50 independent trajectories starting from initial conditions chosen uniformly in $[0, 40]$ and observed at 10 equidistant points in time in $[0, 1]$. In Fig. 1, we see that at $\tau = 1$ we have the highest acceptance rate. Here the preconditioner does a very good job in that the number of samples from the true posterior is very nearly equal to the number of true likelihoods computed — despite the overall $\approx 40\%$ acceptance rate.

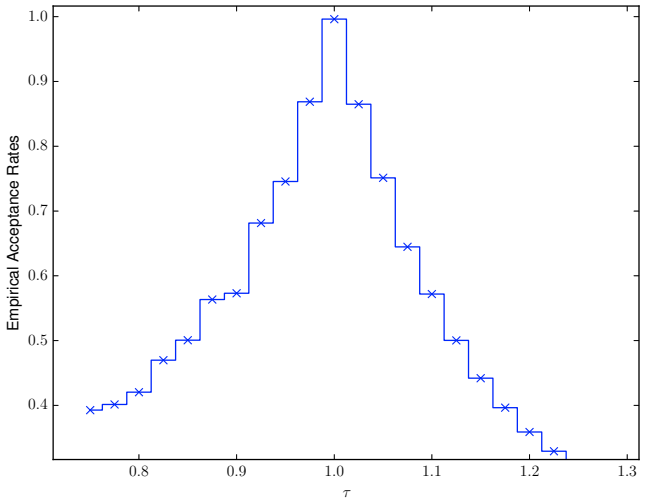

Fig. 1. The empirical acceptance rate as a function of the parameter τ. Close to $\tau = 1$ the preconditioner ensures that almost all steps proposed by the preconditioner will later be accepted by the unbiased likelihood.

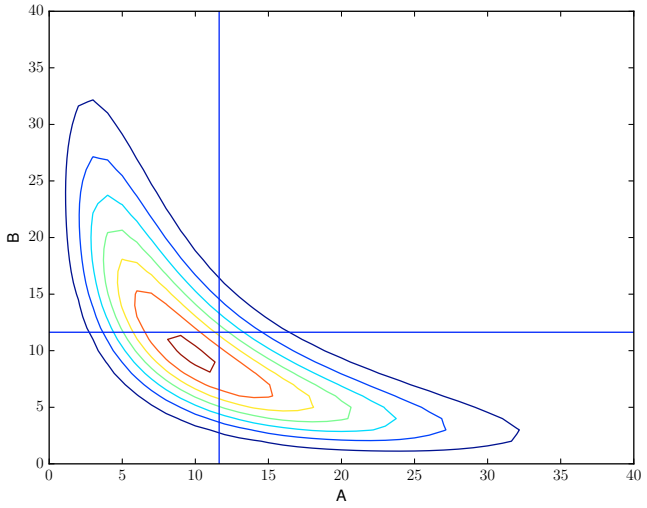

Fig. 2. The solution of the CME for $t = 100$ for the process (9). The intersection of the blue lines indicates the expectation.

3.2 Boomerang

We next consider a two dimensional jump Markov process where its CME does not have known analytical solutions. We approximate the CME of this model using the Optimal Finite State Projection method (OFSP). The OFSP methods error is independent of the model parameters Sunkara and Hegland (2010). The stochastic process of the biological system is given by the following:

$$\begin{pmatrix} a(t) \\ b(t) \end{pmatrix}_{\lambda,\beta} = \begin{pmatrix} a_0 \\ b_0 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \end{pmatrix} \times \mathcal{P}\left(\int_0^t \beta a(s) b(s) ds \right) + \begin{pmatrix} 1 \\ 0 \end{pmatrix} \times \mathcal{P}(\lambda) + \begin{pmatrix} 0 \\ 1 \end{pmatrix} \times \mathcal{P}(\lambda). \quad (9)$$

The system consists of two species (species a and b), each being born at a rate λ. When the two species come in contact, they annihilate each other at rate β. The steady state distribution of this system is shaped like a boomerang, with its mean and mode not coinciding (see Fig. 2).

In this example we will consider using two different preconditioners and investigate their enhancements to the acceptance rates. The first preconditioner will be a Gaussian approximation with the mean and the covariance computed via the moment equations given in Engblom (2006); Golightly and Wilkinson (2005); Sotiropoulos and Kaznessis (2011); Frhlich et al. (2016). The second preconditioner will also be a Gaussian approximation, however, the mean will be computed with all higher order moments set to zero and the covariance will be approximated by the sample covariance of the data. If we denote $\mathbb{E}_{\lambda,\beta}(t)$ to be the expectation and $\mathbb{C}_{\lambda,\beta}(t)$ to be the covariance at time t calculated via the first two moment equations, then the first preconditioner is given as

$$\hat{P}_{(\lambda,\beta)}(O;t) := (1/\sqrt{2\pi \mathbb{V}_{\lambda,\beta}(t|x_0)}) \times e^{-1/2(O-\mathbb{E}_{\lambda,\beta}(t))\mathbb{C}_{\lambda,\beta}(t)^{-1}(O-\mathbb{E}_{\lambda,\beta}(t))^T} \quad (10)$$

with O being the data. For the second preconditioner, we denote the expected value computed by setting all higher moments to zero as $E_{\lambda,\beta}(t)$ and the sample covariance of the data by Cov. Then the approximate probability of observing a data point O at time t for a given parameter set λ, β is:

$$\tilde{P}_{(\lambda,\beta)}(O;t) := (1/\sqrt{2\pi Cov}) \times e^{-1/2(O-E_{\lambda,\beta}(t))Cov^{-1}(O-E_{\lambda,\beta}(t))^T} \quad (11)$$

To construct the empirical distribution of the second proposal's acceptance rate, we perform the following steps. Firstly, we find the scaling parameter such that the MCMC with the likelihood $\mathcal{L}_{CME}$ has an acceptance rate of approximately 40%. This means that if no preconditioner is applied and a full likelihood is used, then we expect to find that roughly sixty percent of CMEs computed would lead to a rejection step. The principle is that if the preconditioner is good, then the same scaling parameter should give a higher acceptance rate. For the experiment conducted here, the scaling parameter was found to be 0.4.

Once the scaling parameter is fixed, we choose 400 random start points from a uniform distribution over the interval, $[0.7, 1.4] \times [0.007, 0.014]$, in the parameter space. With each starting point, we evolve a pcMCMC chain to a length of 20. Then for each chain, we divide the number of successful samples (20 in this case), by the total number of states accepted in the first proposal step. This gives us a single sample from the empirical distribution of the second proposal step. Then we plot the acceptance rates samples as a histogram (see Fig. 3).

We used the following data to fit: At $t = 0.0$ $(a_0, b_0) = (5, 5)$. Then at $t = 15$ we fit five *i.i.d* data points: $(13, 8)$, $(14, 7)$, $(7, 17)$, $(8, 10)$ and $(8, 11)$. Fig. 3 shows the empirical densities of the acceptance rate (P2) of using the preconditioner $\hat{P}$ and $\tilde{P}$. Since $\hat{P}$ has information of the mean and the variance we expect this method to be a very good preconditioner. $\tilde{P}$ has lower acceptance rates since it only uses the information of the mean and the covariance was the sample covariance of the data.

This preconditioner still resulted in a 30% increase in the acceptance rates when compared to no preconditioning. Both preconditioners have substantially decreased the number of rejection steps.

We investigate the quality of the samples gained from the pcMCMC chains by analyzing their respective autocorrelation functions (ACF). We use the SSA to generate 25 data points at $t = 15$. We start the process (9) with an initial population of $(a_0 = 5, b_0 = 5)$ and the parameter values $(\lambda = 1.0, \beta = 0.01)$. This data is the same for all chains. Each chain: preconditioned with $\hat{P}$; preconditioned with $\tilde{P}$ and no preconditioning, is run for a length of 1000 states. We use the same scaling parameter for all three chains $(S = 0.15)$, to compare the decay of the ACFs and the acceptance rates of the second proposal. In Figure 4 we see the ACFs of the three chains for the parameter β. The ACFs for the parameter λ decay similarly to β (see Fig. 5). The three chains have a similar decay to zero. Hence, we thin the three chains by a lag of 50. In this particular example, preconditioning gave us the same quality of points at a higher acceptance rate, that is, less CMEs were computed for the same quality of samples. We see this further in Figure 6, the samples of the respective target distributions all fall within the same region and have a similar overall shape.

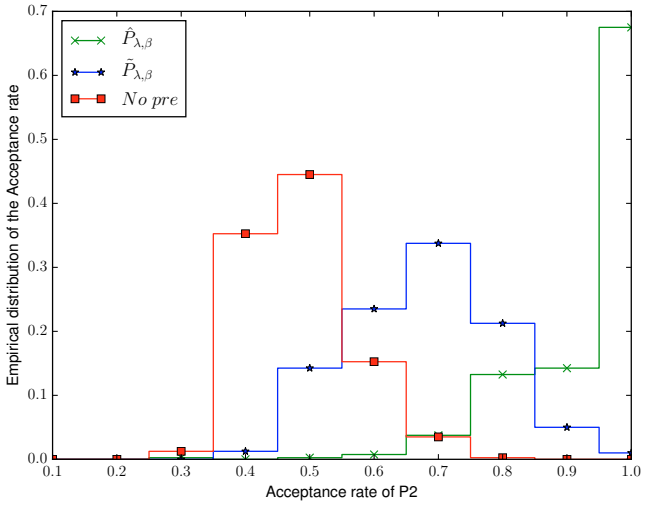

Fig. 3. Empirical distribution of the acceptance rate generated from 400 different starts from a uniformly distribution over, $[0.7, 1.4] \times [0.007, 0.014]$, in the parameter space.

Lastly, we do a simple comparison of the run times of the two preconditioners and the non preconditioned case. It must be noted that the underlying code is not optimized to give a rigorous time comparison. To get a general understanding of the computation time, we conduct a numerical experiment where we compute 100 i.i.d. chains of length 50 for the two preconditioners and with no preconditioner. We time how long each chain takes to accept 50 states in the second proposal. The data and scaling factors are similar to the experiment for finding the empirical distribution of the acceptance rates. We see in Figure 7 that the preconditioner using the mean and the variance, $\hat{P}$, has on average twice the computation speed as the no preconditioning method. The fixed variance

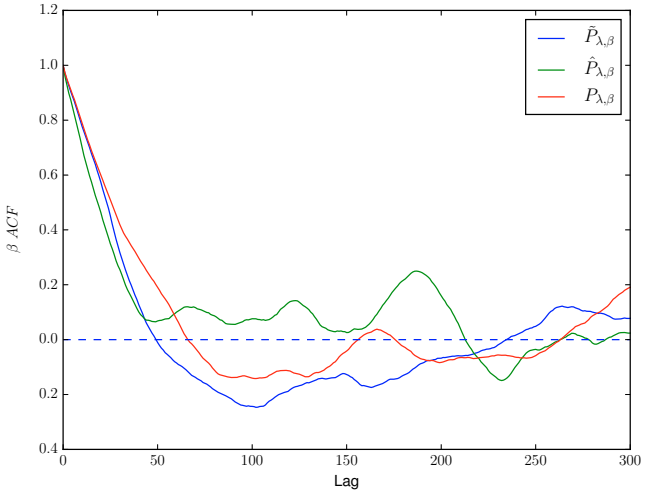

Fig. 4. Autocorrelation functions of the chains in the parameter β.

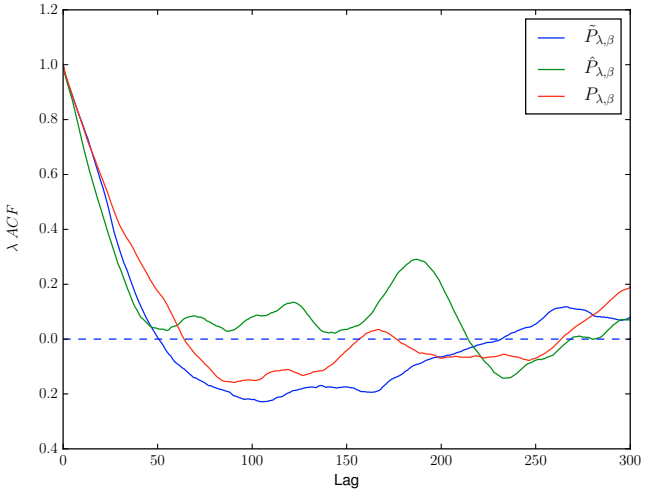

Fig. 5. Autocorrelation functions of the chains in the parameter λ.

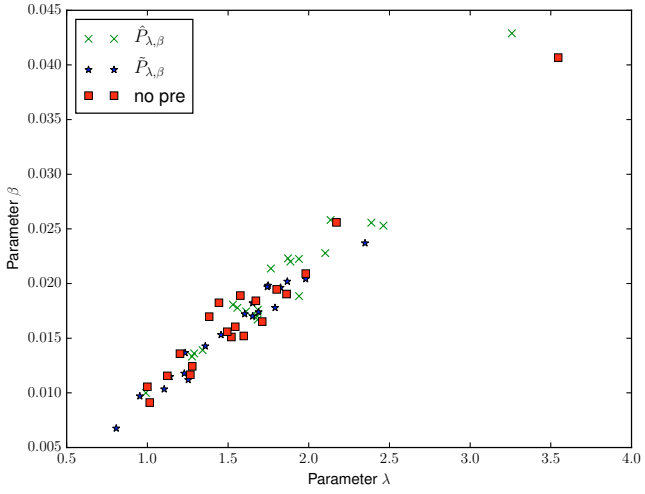

Fig. 6. Sample points of the target distribution gained after thinning the chains of the respective different likelihood approximations. The chains were thinned with a lag of 50 states.

preconditioner, $\tilde{P}$, has on average one and a half times the speed as the no preconditioning method.

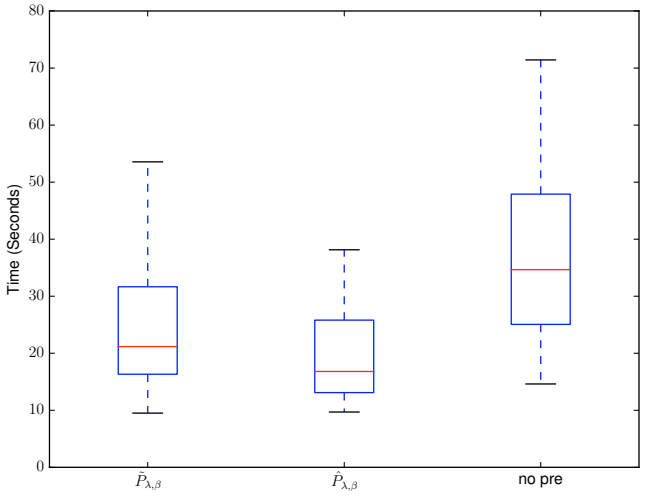

Fig. 7. Computation time of 100 i.i.d chains of length 50 with the respective preconditioner method.

4. CONCLUSION

It is well known that solving CMEs is computationally demanding. Using the preconditioning MCMC method we can maximize the ratio between the number of target samples and the number of CME computations. We saw through our numerical experiments that even simple preconditioners can give an increase in the acceptance rate in the second proposals (where the CMEs are computed). We saw even with a 90% acceptance rate in the second proposal, the target samples were uncorrelated. Furthermore, the acceptance rate of the second proposal can be seen as a *goodness of fit* estimate of how close the likelihood in the preconditioning step is to the likelihood in the second proposal step. The ingredients for constructing a good preconditioner is a question for future research.

REFERENCES

Blake, W.J., Kærn, M., Cantor, C.R., and Collins, J.J. (2003). Noise in eukaryotic gene expression. *Nature*, 422(6932), 633–637.

Efendiev, Y., Hou, T., and Luo, W. (2006). Preconditioning markov chain monte carlo simulations using coarse-scale models. *SIAM Journal on Scientific Computing*, 28(2), 776–803.

Engblom, S. (2006). Computing the moments of high dimensional solutions of the master equation. *Applied Mathematics and Computation*, 180(2), 498–515.

Engblom, S. (2009a). Galerkin spectral method applied to the chemical master equation. *Communications in Computational Physics*, 5(5), 871–896.

Engblom, S. (2009b). Spectral approximation of solutions to the chemical master equation. *Journal of Computational and Applied Mathematics*, 229(1), 208–221.

Ethier, S.N. and Kurtz, T.G. (2009). *Markov processes: characterization and convergence*, volume 282. John Wiley & Sons.

Fröhlich, F., Thomas, P., Kazeroonian, A., Theis, F.J., Grima, R., and Hasenauer, J. (2016). Inference for stochastic chemical kinetics using moment equations and system size expansion. *PLoS Comput Biol*, 12(7), 1–28. doi:10.1371/journal.pcbi.1005030.

Golightly, A., Henderson, D.A., and Sherlock, C. (2015). Delayed acceptance particle mcmc for exact inference in stochastic kinetic models. *Statistics and Computing*, 25(5), 1039–1055.

Golightly, A. and Wilkinson, D.J. (2005). Bayesian inference for stochastic kinetic models using a diffusion approximation. *Biometrics*, 61(3), 781–788.

Higham, D.J. (2008). Modeling and simulating chemical reactions. *SIAM review*, 50(2), 347–368.

Jahnke, T. and Huisinga, W. (2007). Solving the chemical master equation for monomolecular reaction systems analytically. *Journal of mathematical biology*, 54(1), 1–26.

Jahnke, T. and Udrescu, T. (2010). Solving chemical master equations by adaptive wavelet compression. *Journal of Computational Physics*, 229(16), 5724–5741.

Kazeev, V., Khammash, M., Nip, M., and Schwab, C. (2014). Direct solution of the chemical master equation using quantized tensor trains. *PLoS Comput Biol*, 10(3), e1003359.

Kholodenko, B.N. (2000). Negative feedback and ultrasensitivity can bring about oscillations in the mitogen-activated protein kinase cascades. *European Journal of Biochemistry*, 267(6), 1583–1588.

MacArthur, B.D., Ma'ayan, A., and Lemischka, I.R. (2009). Systems biology of stem cell fate and cellular reprogramming. *Nature Reviews Molecular Cell Biology*, 10(10), 672–681.

Metropolis, N., Rosenbluth, A.W., Teller, A.H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21, 1087–1092.

Munsky, B. and Khammash, M. (2006). The finite state projection algorithm for the solution of the chemical master equation. *The Journal of chemical physics*, 124(4), 044104.

Sotiropoulos, V. and Kaznessis, Y.N. (2011). Analytical derivation of moment equations in stochastic chemical kinetics. *Chemical engineering science*, 66(3), 268–277.

Srivastava, R., You, L., Summers, J., and Yin, J. (2002). Stochastic vs. deterministic modeling of intracellular viral kinetics. *Journal of Theoretical Biology*, 218(3), 309–321.

Sunkara, V. and Hegland, M. (2010). An optimal finite state projection method. *Procedia Computer Science*, 1(1), 1579–1586.