Lietuvos matematikos rinkinys. LMD darbai
51 tomas, 2010, 285–290

ISSN 0132-2818
www.mii.lt/LMR/

MCMC approach to modelling queuing systems

Mantas Landauskas, Eimutis Valakevičius

Kaunas University of Technology, Faculty of Fundamental Sciences
K. Donelaičio St. 73, LT-44029 Kaunas
E-mail: mantas.landauskas@ktu.lt; eimval@ktu.lt

Abstract. The paper presents some numerical results on modelling a $M/G/1/\infty$ queuing system using Markov chain Monte Carlo (MCMC). This particular technique was chosen in order to draw samples from the service time distribution from which it is assumed complicated to sample random numbers. The software for experiments was created using C++ Builder development environment.

Keywords: Queuing system, Markov chain Monte Carlo.

Introduction

Consider a single channel queuing system depicted in Fig.1. λ represents the arrival rate (number of arriving customers per time unit) and μ is the service rate (number of customers being serviced per time unit). When the service time has a complicated or unknown distribution, it is difficult (or even impossible) to draw samples from that distribution. In real world this can present a situation, in which the process of arriving customers is a Poisson process and the service time has an unknown distribution.

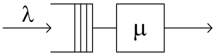


Fig. 1. $M/G/1/\infty$ queuing system.

In this situation it is possible, for example, to construct a kernel density estimate for service time distribution if there are some empirical data given. But the latter estimate does not indicate how to sample from the service time distribution. Consequently, the necessity for special sampling technique emerges. MCMC was chosen to deal with this question.

1 MCMC and Metropolis–Hastings algorithm

If a service time density function $\pi(\cdot)$ is taken, $Y_i \sim \pi(y)$, $i = \overline{1, n}$, have to be generated to run the simulation of the queuing system. The idea of MCMC is to construct a Markov chain $\{X_i\}_{i=0}^{\infty}$ such that $\lim_{i\to\infty} P(X_i = x) = \pi(x)$,

$$P(X_0 = x) = g(x), \tag{1}$$

$$P(y|x) = P(X_{i+1} = y|X_i = x). \tag{2}$$

Every Markov chain can be determined through an initial state (1) and a transition kernel (2). It is known that the stationary distribution is unique if Markov chain is ergodic:

$$\pi(y) = \sum_{x \in \Omega} \pi(x) P(y|x), \quad \forall y \in \Omega. \tag{3}$$

Suppose the discrete stationary probabilities $\pi(X_i)$ had been given. Then, having ergodic and discrete Markov chain, the equation (3) holds. Total number of $(n-1)$ equations and $n(n-1)$ unknown transition kernel probabilities are apparent. Thus there exist an infinite number of transition kernels representing the stacionary distribution $\pi(x)$. Any of those transition kernels can be constructed and used for generating X_i. One of the most widely used methods for constructing such Markov chain is Metropolis–Hastings algorithm [2].

Metropolis–Hastings algorithm is implemented as follows. At first an optional transition kernel $Q(y|x)$ is chosen. Then there exists a probability α for chosen kernel Q being equal to transition kernel P:

$$P(y|x) = Q(y|x)\alpha(y|x), \quad y \neq x. \tag{4}$$

Considering the detailed balance condition of a time-homogeneous Markov chain we have:

$$\pi(x)Q(y|x)\alpha(y|x) = \pi(y)Q(x|y)\alpha(x|y), \quad \forall x \neq y. \tag{5}$$

General solution for (5) is $\alpha(y|x) = r(x|y)\pi(y)Q(x|y)$. It is neccessary to have a higher acceptance ratio when sampling random numbers [3], therefore by adjusting $r(x,y)$ it is shown that:

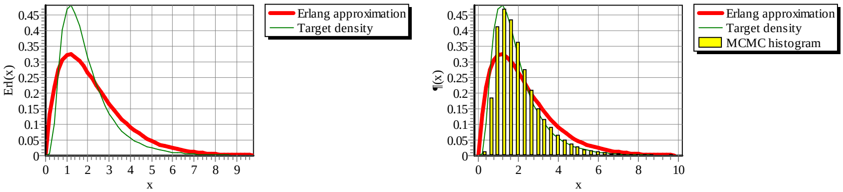
$$\alpha(y|x) = \min\left(1, \frac{\pi(y)Q(x|y)}{\pi(x)Q(y|x)}\right). \tag{6}$$

Sampling of each X_i is performed in 4 steps. Firstly a candidate point X_i is drawn from proposal distribution. Then the probability α_i, that this point is also distributed by the target density, is calculated. The next step is to draw $u_i \sim U(0;1)$ and compare it to α_i. Finally, X_i is accepted to the sample if $u_i < \alpha_i$. Otherwise $X_i = X_{i-1}$.

From (6) it is evident that $\pi(x)$ can be determined up to a multiplicative constant c, i.e., $\pi(x) = c \cdot h(x)$, where $h(x)$ is a probability density function. Having chosen $Q(x|y) \equiv Q(x)$, an MCMC independence sampler is implemented. In that case there is no neccesity for Markov chain to loose its memory because each X_i depends on ratio $\frac{\pi(\cdot)}{Q(\cdot)}$ and not on X_{i-1}.

2 Proposal selection and convergence of MCMC

Suppose we have a $M/G/1/\infty$ queue in which service time X is a random number having lognormal distribution $f_t(x, \mu, \sigma)$ with parameters $\mu = 0.5$ and $\sigma = 0.6$. Nevertheless this density could be sampled using normal distribution, say it is complicated to sample from it. Here comes the necessity for special sampling technique. A proposal density Q must be chosen now. There are many techniques to achieve this, because the problem is to find probability distribution similar in shape to the



a) approximation

b) MCMC results

Fig. 2. Using Erlang distribution as proposal density.

Table 1. MCMC convergence.

Sample size	$\triangle_{te}$	$\triangle_{t\pi}$
500	0.0069	0.0340
1 000	0.0069	0.0258
5 000	0.0069	0.0106
10 000	0.0069	0.0095
50 000	0.0069	0.0057
500 000	0.0069	0.0050
1 000 000	0.0069	0.0046

target distribution. One way of achieving this is to take Erlang distribution $f_e(x,k,\lambda)$ and compare its first 3 moments according to moments m_1, m_2 and m_3 of the target distribution.

$$b = \frac{m_3 + m_1 - 3m_2}{m_2 - m_1}, \qquad k = b, \qquad \lambda = \sqrt[b]{\frac{m_2 - m_1}{b^2}}. \tag{7}$$

By doing this, equations (7) for k and λ are obtained. Substituting these values into $f_e(x,k,\lambda)$ make it target density approximation, that is, proposal density. Difficulties arises when performing such approximation. k calculates as a floating point number and needs to be integer. There are cases when rounding k gives undesirable results, in such cases approximation is more precise if 1 or 2 is added to k. The second drawback is the non-versatility of this method. For particular distributions custom k and λ are found, but approximation is undesirable.

In this particular case $k = 2$, $\lambda = 0.882$ and $f_e(x,k,\lambda)$ is quite similar in shape to $f_t(x,\mu,\sigma)$. Let us denote $f_\pi(\cdot)$ to be an empirical probability density obtained by MCMC. The histogram, consisting of 100 bars, of sampled numbers X_i, was taken as $f_\pi(\cdot)$. The differences between the target and the proposal densities or between the target and empirical densities are evaluated as:

$$\triangle_{te} = \frac{1}{m}\sum_{j=1}^{m}\bigl(\bigl|f_t(X_j) - f_e(X_j)\bigr|\bigr), \quad m < n. \tag{8}$$

$$\triangle_{t\pi} = \frac{1}{m}\sum_{j=1}^{m}\bigl(\bigl|f_t(X_j) - f_\pi(X_j)\bigr|\bigr), \quad m < n. \tag{9}$$

According to Table 1, the convergence of MCMC independence sampler to the target distribution is rather slow.

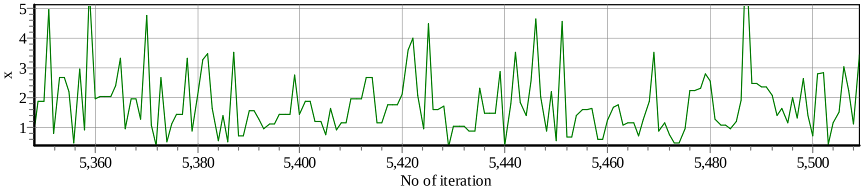


Fig. 3. Dependent MCMC process run.

3 Numerical model of $M/G/1/\infty$ queuing system

Consider a queuing system with $\lambda = 0.4$ when service time distribution is $f_t(x, 0.5, 0.6)$. The software created represents a queuing system run as a dynamic chain of objects in computer memory. Each object has attributes just as a particular customer. The first attribute is arrival time to the system. The second attribute is service time. The time in queue, the time in system and etc are calculated recursively during modelling process. Arrival time is generated using inverse cumulative exponential distribution, although service times is said to have difficult distribution and, therefore, it is sampled using MCMC. The approximation of target density with Erlang distribution function was used as a technique for constructing a proposal density.

Although MCMC convergence to the target distribution is a matter-of-course, the result of this sampling technique is still a Markov chain, i.e., numbers are dependent. Fig. 3 shows that there are moments when $X_{i+1} = X_i$. Before using these numbers for modelling a queuing system it is essential to scramble the sample to prevent from being dependent. A simpler way is to accept every second (third, etc.) random number to the sample while running MCMC. The more $f_t(\cdot)$ differs in shape from the proposal density, the more numbers should be missed. If MCMC process run is dependent, there is a probability for two or more service times in a row being equal and relatively high. This results in the queue being much higher in length than the average queue for some time. According to the experiments carried out, time and queue characteristics were higher than the theoretical ones if dependency was not eliminated from the sample.

Having generated a single process run of independent random values, the average value of customers in queue or system and the mean value of time they were in queue or system can be evaluated. It is advisable to generate several process runs and calculate the average estimates. By doing this the dispersion of an estimate is being reduced. The system's empirical characteristics are compared to theoretical ones in order to evaluate the accuracy of the model. $M/G/1/\infty$ queuing system's theoretical characteristics are calculated in accordance with Little's and Pollaczek-Khinchin formulas [1]. Figs. 4 and 5 show the mean absolute value of the average relative error of each of the queue's characteristics for specified λ. These results were obtained by modelling 50000 arriving customers for 10 times for each λ.

Any characteristic is calculated more precisely if λ is small. This dependence is reasonable because the less customers enter the system, the shorter the queue is. But the purpose of this experiment is to obtain a numerical value of what the average error will be for this particular system. Considering this relationship, one can decide

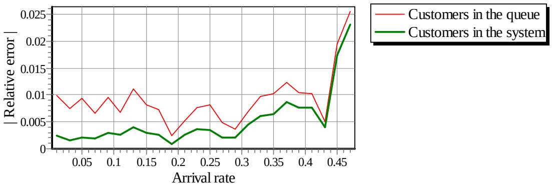


Fig. 4. The relative error of characteristics of the queue as the function of λ.

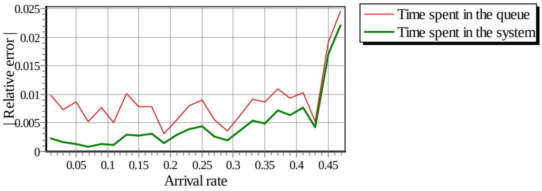


Fig. 5. The relative error of characteristics of the queue as the function of λ.

whether modelling a system with particular parameters will give desirable results. It is also noticeable from Figs. 4 and 5 that system's characteristics are evaluated more precise than the queue characteristics.

When modelling the system by usual techniques (e.g., inverse cumulative distribution function), the precision of its characteristics has similar trends. In several cases it is possible to achieve a smaller rate of error with MCMC, when random numbers sampled with MCMC have a better quality (e.g., if scrambled they can be less dependent).

4 Conclusions

1. The higher the ratio $\frac{\lambda}{\mu}$, the higher the dispersion of estimated system characteristics. The modelling also shows that errors for each system characteristics are dependent (Figs. 4 and 5).
2. Heavier tails of the service time distribution leads to the higher dispersion of characteristics modelled.
3. The results obtained enables us to choose how many arrivals of customers we have to sample in order to get desired precision of system characteristics while knowing its parameters.
4. When using MCMC for queuing systems or other type of logical aggregates, the sample $X_i \sim \pi(x)$ must be scrambled to prevent X_i from being dependent.

References

[1] G. Bolch, S. Greiner, H. de Meer and S. K. Trivedi. *Queuing Networks and Markov Chains*. John Willey & Sons Ltd., 2006.

[2] J.S. Dagpunar. *Simulation and Monte Carlo*. John Willey & Sons Ltd., Great Britain, Chippenham, Willtshire, 2007.

[3] V. Prokaj. Proposal selection for mcmc simulation. *Applied Stochastic Models and Data Analysis*, pp. 61–65, 2009.

REZIUMĖ

Eilių sistemų modeliavimas MCMC metodu

M. Landauskas, E. Valakevičius

Straipsnyje nagrinėjama eilių sistema, kurios paraiškų atėjimo srautas yra puasoninis, o aptarnavimas pasiskirstęs neeksponentiškai. Aptarnavimo srautui generuoti naudotas MCMC metodas. Siūlomo skirstinio pasirinkimas atliekamas sulyginant tikslinio ir Erlango pasiskirstymų 3 pradinius momentus. Sukurta programinė įranga ir tyrinėtos MCMC metodo subtilybės leido sumodeliuoti konkrečią eilių sistemą. Modeliavimo metu pastebėtas ženklus paklaidų išaugimas, kai aptarnavimo ir paraiškų atėjimo į sistemą srautai yra panašūs. Iš atliktų tyrimų darytina išvada, kad modeliuojant eilių sistemas verta modeliavimą kartoti daug kartų ir skaičiuoti sistemos charakteristikų vidurkius.

Raktiniai žodžiai: eilių sistema, Markovo grandinių Monte Karlo metodas.