The Good, the Bad, and the Ugly in Markov Chains

Ad Ridder

April 12, 2019

Abstract

My main research interest concerns rare-event simulation. In this paper I explain a few concepts of this research field, and illustrate it by two toy examples.

1 Introduction

Consider a discrete-time Markov chain $\{X(t), t = 0, 1, \ldots\}$ on a state space S with transition probabilities $p(x, y) \doteq \mathbb{P}\big(X(t+1) = y | X(t) = x\big)$. The state space is partitioned into three (disjoint) sets, $S = G \cup B \cup U$, where G is the set of *good states*, B is the set of *bad states*, and U is the set of *ugly states*. It is assumed that the chain is irreducible, which means that all states are attainable from any state in the state space. Suppose that the chain starts in an ugly state, say u, and that we are interested in the probability that the chain reaches a bad state before a good state, while this probability is extremely small. We say that the event is a *rare event*. The rare-event probability is denoted by γ, which is supposed to be $\gamma \approx 0$. In the analysis we denote sometimes $\gamma(u)$ for specifying the initial state. Two examples will illustrate these concepts.

Example 1 ($M/M/1$ Queue)**.** Recall the $M/M/1$ queue with arrival rate λ and service rate μ, where $\lambda < \mu$. A busy period is the period of time during which the server is continuously busy. We are interested in the probability that the queue length exceeds n ever during a busy period, where n is large. This problem is cast easily into our formulation, by considering the discrete-time Markov chain embedded at the arrival and departure times of the queueing process. The state space $S = \{0, 1, \ldots\}$ represents the number of customers in the queue and service, with transition probabilities

$$p(x, x+1) = \lambda/(\lambda + \mu),\ p(x, x-1) = \mu/(\lambda + \mu) = 1 - p \quad (x = 1, 2, \ldots).$$

The good, the bad and the ugly states are,

$$G = \{0\},\ B = \{n, n+1, \ldots\},\ U = \{1, \ldots, n-1\}.$$

When we denote the rare event probability by $\gamma^{(n)}$, then $\gamma^{(n)} \to 0$ as $n \to \infty$. □

Example 2 (Repair Sytem)**.** Consider a factory with n identical machines. Each machine fails after an exponentially distributed random time with mean $1/\epsilon$. The machines are highly reliable, meaning $\epsilon > 0$ but small. Immediately after failure, a handyman starts to work on it to repair the machine, which takes an exponentially distributed random time with mean $1/\mu$. After repair the machine is as good as new. There are enough handymen for all machines because the factory is down when all machines have failed. Suppose that at time 0 a machine has failed, then the interest is in the probability that the factory goes down before all machines are working. This problem is cast easily into our formulation, by considering the discrete-time

Markov chain embedded at the failure and repaired times of the machines. The state space $S = \{0, 1, \ldots, n\}$ represents the number of failed machines, with transition probabilities

$$\begin{aligned} p(x, x+1) &= \frac{(n-x)\epsilon}{(n-x)\epsilon + x\mu} & (x \leq n-1), \\ p(x, x-1) &= \frac{x\mu}{(n-x)\epsilon + x\mu} & (x \geq 1). \end{aligned}$$

The good, the bad and the ugly states are,

$$G = \{0\},\ B = \{n\},\ U = \{1, \ldots, n-1\}.$$

When we denote the rare event probability by $\gamma^{(\epsilon)}$, then $\gamma^{(\epsilon)} \to 0$ as $\epsilon \to 0$. □

2 Monte Carlo Simulation

The rare-event probability is computed by running a simulation of the Markov chain. This is fairly easy to model and to program. The program generates N sample paths of the Markov chain, all starting from the same ugly state, say u, and ending either in a good state (score $Z_i = 0$ of the i-th path) or in a bad state (score $Z_i = 1$). The average of the Z_i scores is an unbiased estimator of the target probability,

$$\widehat{\gamma} = \overline{Z}_N \doteq \frac{1}{N} \sum_{i=1}^{N} Z_i.$$

Commonly this is called the *Monte Carlo estimator*. Using the sample variance we compute the standard error of the estimator, and construct the associated confidence intervals. In our numerical experiments we set the sample sizes so large that the 95% confidence intervals have relative widths of 20% (in both directions 10%). An important aspect in simulation projects is the computation time needed for accurate estimates. In these Monte Carlo simulations, the computation times are proportional to the sample sizes, and these are proportional to $1/\gamma$ due to our criterion.

Example 3. In the $M/M/1$ queue we set arrival rate $\lambda = 0.8$, service rate $\mu = 1$, and let overflow level n be increasing from $20, 30, \ldots$. In the repair system we set repair rate $\mu = 1$, $n = 10$ machines, and let the failure rate ϵ decreasing from $0.5, 0.4, \ldots$.

Table 1: Simulation results for the $M/M/1$ queue and the repair system.

$M/M/1$ queue			repair system		
n	N	$\overline{Z}_N$	ϵ	N	$\overline{Z}_N$
20	150 K	2.8067e-03	0.5	200 k	1.9100e-03
30	1.3 M	3.4000e-04	0.4	1.6 M	2.4500e-04
40	12 M	3.1833e-05	0.3	21 M	1.8095e-05
50	112 M	3.1875e-06	0.25	108 M	3.611e-06
60	1050 M	3.7810e-07	0.2	800 M	4.900e-07

The last row of the table took about 4 minutes for each of the systems, executed by programs coded in C on a MacBook Pro laptop with 2.4GHz processor, and 8GB 1333Mhz RAM. Larger overflow levels in the $M/M/1$, and smaller failure rates in the repair system become problematic, for instance $n = 100$ would take a sample size of about 7800 G which would take about 225 days, and $\epsilon = 0.05$ would take a sample size of about 20 T which would take about 18 days. □

3 Complexity of Estimators

Consider any unbiased estimator $\overline{\zeta}_N$ of the rare-event probability γ based on N i.i.d. replicas of a single-run estimator ζ. The target is to get 95% confidence intervals having relative widths of 20%, which leads to (with 95%-quantile 2.0 in stead of 1.96),

$$4\sqrt{\frac{\mathbb{V}ar(\zeta)}{N}} \leq 0.2\gamma \quad \Leftrightarrow \quad N \geq 400\frac{\mathbb{V}ar(\zeta)}{\gamma^2}.$$

The single-run Monte Carlo estimator Z is a Bernoulli random variable with variance $\gamma(1-\gamma) \approx \gamma$ in the rare-event regime ($\gamma \approx 0$). Thus, indeed we need sample size N proportionally to $1/\gamma$, and typically, the rare-event probability decays exponentially fast. For instance in the $M/M/1$ queue,

$$\gamma = O\big(\exp(-\alpha n)\big), \ n \to \infty.$$

Thus, the required sample sizes have an *exponential complexity*, see also Table 1. The research field of rare-event simulation is about finding estimators for which the required sample sizes have a *polynomial complexity*, or even better, a *bounded complexity*, or optimally, a *zero complexity*. The latter is obtained iff $\mathbb{V}ar(\zeta) = 0$. In the next section we shall show how we can construct an estimator with this property.

4 Importance Sampling Simulation

Simulation of the Markov chain is implemented using other transition probabilities, say $\{q(x, y) : x, y \in S\}$. This is called a *change of measure*. Again, we run it until either a good state or a bad state is attained, which gives scores 0 and 1, respectively for estimating the rare-event probability. However, to get an unbiased estimator we need to multiply the score with the likelihood ratio of the generated sample path.

Consider an arbitrary sample path $\boldsymbol{X} = \big(X(0) = u, X(1), \ldots, X(\tau)\big)$ of the Markov chain until either a good or bad state is reached. The *likelihood ratio* of this path is defined by

$$L(\boldsymbol{X}) \doteq \prod_{t=0}^{\tau-1} \frac{p(X(t), X(t+1))}{q(X(t), X(t+1))}.$$

The new single-run estimator becomes

$$\zeta \doteq L(\boldsymbol{X})\mathbb{1}\{X(\tau) \in B\},$$

and is called an *importance sampling estimator* of the rare-event probability. It is easy to see that it is unbiased. Note that expectations are taken with respect to the change of measure, which will be indicated by subscript COM. The second moment of the importance sampling estimator is

$$\mathbb{E}_{\mathrm{COM}}\big[\zeta^2\big] = \mathbb{E}_{\mathrm{COM}}\big[L^2(\boldsymbol{X})\mathbb{1}\{X(\tau) \in B\}\big] = \mathbb{E}\big[L(\boldsymbol{X})\mathbb{1}\{X(\tau) \in B\}\big]. \tag{1}$$

The last equality can be seen by an equivalent shorthand

$$\sum_{\omega\in\Omega}\Big(\frac{P(\omega)}{Q(\omega)}\Big)^2 X(\omega)\,Q(\omega)=\sum_{\omega\in\Omega}\frac{P(\omega)}{Q(\omega)}X(\omega)\,P(\omega).$$

Now, consider a specific change of measure given by

$$q(x,y)=p(x,y)\,\frac{\gamma(y)}{\gamma(x)},\quad (x\in U). \tag{2}$$

Recall that $\gamma(x)$ is the probability of reaching the bad state before the good state when the chain would start in the ugly state x.

Theorem 1. *The importance sampling estimator associated with the change of measure* (2) *has zero variance.*

Proof. Work out the likelihood ratio $L(\boldsymbol{X})$ of a sample path:

$$\prod_{t=0}^{\tau-1}\frac{p(X(t),X(t+1))}{q(X(t),X(t+1))}=\prod_{t=0}^{\tau-1}\frac{\gamma(X(t))}{\gamma(X(t+1))}=\frac{\gamma(X(0))}{\gamma(X(\tau))}=\frac{\gamma(u)}{\gamma(X(\tau))}.$$

Now substitute this expression in the second moment (1):

$$\mathbb{E}_{\mathrm{COM}}\big[(\zeta)^2\big]=\mathbb{E}\Big[\frac{\gamma(u)}{\gamma(X(\tau))}\,\mathbb{1}\{X(\tau)\in B\}\Big]\overset{(i)}{=}\gamma(u)\mathbb{E}\big[\mathbb{1}\{X(\tau)\in B\}\big]=\gamma^2(u).$$

Equality (i) follows from $\mathbb{1}\{X(\tau)\in B\}\ \Rightarrow\ \gamma(X(\tau))=1$. Conclude that the variance of the importance sampling estimator ζ has zero variance. □

An implementation of the zero-variance importance sampling estimator needs the unknown rare-event probabilities $\gamma(x)$, see (2)! Considering these probabilities as absorption probabilities, they satisfy the linear equations,

$$\gamma(x)=\sum_{y\in S}p(x,y)\gamma(y),\ x\in U, \tag{3}$$

with boundary conditions $\gamma(x)=0$ for $x\in G$, and $\gamma(x)=1$ for $x\in B$. In other words, the probabilities form a *Lyapunov function*. Generally, no analytic expressions exist for the solution. However, for specific cases there are ways to express the solution, like in our examples.

Example 4. The $M/M/1$ queue and the repair system are random walk models on a state space $S=\{0,1,\ldots\}$ with ±1 jumps and state-dependent jump probabilities $p(x,x\pm1)$. Let f be a Lyapunov function satisfying (3), and define $\Delta f(x)\doteq f(x+1)-f(x)$. For our random walks we get that

$$p(x,x+1)\Delta f(x+1)=p(x,x-1)\Delta f(x)\ \Leftrightarrow\ \Delta f(x+1)=\frac{p(x,x-1)}{p(x,x+1)}\Delta f(x).$$

Iterating, this leads to Lyapunov solutions

$$f(x)=\big(\Delta f(1)\big)\sum_{k=1}^{x}\prod_{y=1}^{k-1}\frac{p(y,y-1)}{p(y,y+1)},\ \ x=1,2,\ldots$$

The boundary conditions $f(0) = 0$ and $f(n) = 1$ would result in the rare-event probabilities $\gamma(x)$.

Importance sampling of the $M/M/1$ queue and the repair sytem is implemented with the zero-variance transition probabilities in (2), obtained by the Lyapunov solutions. Then it suffices to generate just a *single sample path* with these transition probabilities, and compute the associated likelihood ratio. The path ends in the bad state with probability 1, and the likelihood ratio is equal to the correct estimate! It is possible to do this for any rare event. For instance, the probability of more than $n = 1000$ customers in a busy cycle in our $M/M/1$ queue is equal to 3.0756e-98 which is a ridiculous small number, but it is obtained immediately from the zero-variance estimator. Similarly, suppose that the failure rate is ϵ =1.0e-10 in the repair system, then our rare-event probability is 1.0000e-90, again obtained immediately. □

For more complicated models and for realistic problems, it is not possible to implement exactly the zero-variance change of measure. The common approach is to approximate it. In stead of constructing the Lyapunov function exactly, we might try to approximate it, or simplify the Lyapunov equations, or solve it with an inequality. The approximated function is denoted by $v(x)$, and the transition probabilities for the change of measure become

$$q(x,y) = \frac{p(x,y)\,v(y)}{\sum_{y\in S} p(x,y)\,v(y)}, \quad (x \in S).$$

This *zero-variance approximation* may result in estimators with bounded or with polynomial complexity.

Example 5. Consider again our $M/M/1$ queue and the repair system. We simplify the Lyaponov equation to

$$p(x-1,x)v(x) = p(x,x-1)v(x-1).$$

Which leads to

$$v(x) = v(1)\prod_{y=1}^{x-1}\frac{p(y,y-1)}{p(y-1,y)}, \quad x = 1,2,\ldots.$$

Executing importance sampling simulation with this change of meaure in the $M/M/1$ queue, we found a bounded (constant) complexity of the estimator: it suffices a sample size of $N = 1500$ to obtain 95% confidence intervals with relative width of 20%, whatever the overflow level n might be. In the repair system, the results show decreasing sample size! For $\epsilon = 0.5, 0.4, \ldots, 0.1$ the required sample sizes were 4M, 70K, 50K, 500, 100, respectively. The complexity is said to be *vanishing*. □

Conclusion

The take away is that when you do a simulation, do first some analysis of the model and problem. This may give you a more efficient simulation algorithm that saves you a lot of time.

References

Most problems, results, algorithms and analysis on rare-event simuation can be found only in papers that appeared (and still appear) in scientific journals. There are a few books and review papers that could serve as starting points.

Bucklew, J.A. (2004). *Introduction to Rare Event Simulation*. Springer.

Heidelberger, P. (1995). Fast simulations of rare events in queueing and reliability models. *ACM Transactions on Modeling and Computer Simulations* 5, 43–85.

Juneja, S. and Shahabuddin, P. (2006). Rare-event simulation techniques: an introduction and recent advances. In *Handbooks in Operations Research and Management Science, Vol. 13: Simulation*. S. Henderson and B. Nelson (Eds.), Elsevier, Amsterdam, 291–350.

Rubino, G. and Tuffin, B. (Eds.) (2009). *Rare Event Simulation Using Monte Carlo Methods*, Wiley.