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 sequence of discrete-time Markov chains $(X^{(\epsilon)}(t))_{t=0}^{\infty}$ parameterized by $\epsilon > 0$. The ϵ-chain is defined on a state space $S^{(\epsilon)}$, and has transition probabilities

$$p^{(\epsilon)}(x,y) \doteq \mathbb{P}\big(X^{(\epsilon)}(t+1) = y | X^{(\epsilon)}(t) = x\big).$$

The state space is partitioned into three (disjoint) sets, $S^{(\epsilon)} = G^{(\epsilon)} \cup B^{(\epsilon)} \cup U^{(\epsilon)}$, where $G^{(\epsilon)}$ is the set of *good states*, $B^{(\epsilon)}$ is the set of *bad states*, and $U^{(\epsilon)}$ 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 $X^{(\epsilon)}(0) = u \in U^{(\epsilon)}$, then we are interested in the probability that the chain reaches a bad state before a good state. In formal notation: define stopping time

$$\tau_G^{(\epsilon)} \doteq \inf\big\{t \geq 0 : X^{(\epsilon)}(t) \in G^{(\epsilon)}\big\},$$

and similarly for $\tau_B^{(\epsilon)}$. The probability of interest is

$$\gamma^{(\epsilon)}(u) \doteq \mathbb{P}\big(\tau_B^{(\epsilon)} < \tau_G^{(\epsilon)} | X^{(\epsilon)}(0) = u\big),\ u \in U^{(\epsilon)}. \tag{1}$$

Specifically, we are interested in computing this probability for initial states u that are 'close' to the good set, and for which

$$\lim_{\epsilon \to 0} \gamma^{(\epsilon)}(u) = 0,$$

assuming that the limit exists. This property says that the event $\big\{\tau_B^{(\epsilon)} < \tau_G^{(\epsilon)}\big\}$ is a *rare event* for initial state u. Two examples will illustrate the context and the concepts.

Example 1 ($M/M/1$ Queue)**.** Recall the well-known $M/M/1$ 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 chains embedded at the arrival and departure times of the queueing process. The parameterization is $n = [1/\epsilon]$. All n-chains have state space $S = \{0, 1, \ldots\}$ representing the number of customers in the queue and service, with transition probabilities

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

The partition sets are parameterized:

$$G^{(n)} = \{0\},\ B^{(n)} = \{n, n+1, \ldots\},\ U^{(n)} = \{1, \ldots, n-1\},\ n = 1, 2, \ldots.$$

The initial ugly state is $X^{(n)}(0) = 1$. One can show that we deal with a rare event probability $\gamma^{(n)}(1) \to 0$ as $n \to \infty$. □

Example 2 (Repair System)**.** 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 chains embedded at the failure and repaired times of the machines. All ϵ-chains have state space $S = \{0, 1, \ldots, n\}$ representing the number of failed machines. The state space is partitioned by

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

The transition probabilities are parameterized

$$p^{(\epsilon)}(x, x+1) = \frac{(n-x)\epsilon}{(n-x)\epsilon + x\mu} \qquad (x \le n-1),$$
$$p^{(\epsilon)}(x, x-1) = \frac{x\mu}{(n-x)\epsilon + x\mu} \qquad (x \ge 1).$$

The initial ugly state is $X^{(\epsilon)}(0) = 1$. One can show that we deal with a rare event probability $\gamma^{(\epsilon)}(1) \to 0$ as $\epsilon \to 0$. □

2 Monte Carlo Simulation

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

Algorithm 1 Monte Carlo Simulation

```
Initialize a vector (Z_1, ..., Z_N) of zeros.
for i = 1 to N do
    X ← u
    while X ∈ U^(ε) do
        Generate Y ~ p^(ε)(X, ·)
        if Y ∈ U^(ε) then
            X ← Y
        else
            Z_i ← 𝟙{Y ∈ B^(ε)}
            return
        end if
    end while
end for
```

The average of the Z_i scores is an unbiased estimator of the target probability,

$$\widehat{\gamma}^{(\epsilon)}(u) = \overline{Z}_N^{(\epsilon)} \doteq \frac{1}{N}\sum_{i=1}^{N} Z_i. \tag{2}$$

Commonly this is called the *Monte Carlo estimator*. By the strong law of large numbers it holds that $\overline{Z}_N^{(\epsilon)}$ converges almost surely to the target probability. This is actually the essence of stochastic simulation, and why simulation works to compute estimates. Using the sample variance we compute the standard error of the estimator, and construct the associated confidence intervals. We have applied Algorithm 1 to the two examples. The sample sizes are taken so large that the 95% confidence intervals have relative widths of 20% (in both directions 10%). It is a first year Statistics exercise to derive that the sample sizes should be about 400 times the reciprocal of the probability that we want to estimate. 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.

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 the Estimator

The research objective is to construct an unbiased estimator of the rare-event probability (1) having a much smaller variance than the Monte Carlo estimator (2). The Monte Carlo estimator $\overline{Z}_N^{(\epsilon)}$ is the average of N i.i.d. replicas of the single-run estimator

$$Z^{(\epsilon)} \doteq \mathbb{1}\{\tau_B^{(\epsilon)} < \tau_G^{(\epsilon)}\}.$$

Let $\overline{\zeta}_N^{(\epsilon)}$ be some unbiased estimator of the rare-event probability $\gamma^{(\epsilon)}(u)$ based on N i.i.d. replicas of a single-run estimator $\zeta^{(\epsilon)}$. Again, the target is to get 95% confidence intervals

having relative widths of 20%. Thus (with 95%-quantile 2.0 in stead of 1.96),

$$4\sqrt{\frac{\mathbb{V}ar\big(\zeta^{(\epsilon)}\big)}{N}} \le 0.2\gamma^{(\epsilon)}(u) \quad \Leftrightarrow \quad N \ge 400\frac{\mathbb{V}ar\big(\zeta^{(\epsilon)}\big)}{\big(\gamma^{(\epsilon)}(u)\big)^2}.$$

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

$$\gamma^{(\epsilon)}(u) = O\big(\exp(-\alpha/\epsilon)\big),\ \epsilon \to 0.$$

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 $\mathbb{V}ar\big(\zeta^{(\epsilon)}\big) = 0$. In the next section we shall show how we can construct an estimator with this property.

4 Importance Sampling Simulation

Suppose that we simulate the Markov chain with other transition probabilities, say $q^{(\epsilon)}(x,y), x, y \in S^{(\epsilon)}$. 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.

Let $\boldsymbol{X}^{(\epsilon)} \doteq \big(X(0), X(1), \ldots, X(\tau)\big)$ be a sample path generated by the new transition probabilities, where

$$\tau = \min\big(\tau_G^{(\epsilon)}, \tau_B^{(\epsilon)}\big)$$

signals the end of the sample path in either a good state or a bad state. The *likelihood ratio* of this path is defined by

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

The new single-run estimator becomes

$$\zeta^{(\epsilon)} \doteq L\big(\boldsymbol{X}^{(\epsilon)}\big)\, \mathbb{1}\{X(\tau) \in B^{(\epsilon)}\}, \tag{3}$$

and is called an *importance sampling estimator* of the rare-event probability. It is easy to see that it is unbiased. What about its variance?

$$\mathbb{V}ar_{\mathrm{COM}}\big(\zeta^{(\epsilon)}\big) = \mathbb{E}_{\mathrm{COM}}\big[\big(\zeta^{(\epsilon)}\big)^2\big] - \big(\gamma^{(\epsilon)}(u)\big)^2.$$

The subscript COM indicates that expectations are taken with respect to the probability measure that is defined by the change of measure. Therefore we get

$$\begin{aligned} &\mathbb{E}_{\mathrm{COM}}\big[\big(\zeta^{(\epsilon)}\big)^2\big] = \mathbb{E}_{\mathrm{COM}}\big[L^2\big(\boldsymbol{X}^{(\epsilon)}\big)\, \mathbb{1}\{X(\tau) \in B^{(\epsilon)}\}\big] \\ &= \mathbb{E}\big[L\big(\boldsymbol{X}^{(\epsilon)}\big)\, \mathbb{1}\{X(\tau) \in B^{(\epsilon)}\}\big]. \end{aligned} \tag{4}$$

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

$$q^{(\epsilon)}(x,y) = p^{(\epsilon)}(x,y)\,\frac{\gamma^{(\epsilon)}(y)}{\gamma^{(\epsilon)}(x)}, \quad (x \in U^{(\epsilon)}). \tag{5}$$

Theorem 1. *The importance sampling estimator given in* (3) *using the change of measure* (5) *has zero variance.*

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

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

Now substitute this expression in the second moment (4):

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

Equality (i) follows from $\mathbb{1}\{X(\tau)\in B^{(\epsilon)}\} \;\Rightarrow\; \gamma^{(\epsilon)}\big(X(\tau)\big) = 1$. □

When you look closely to the expressions of the zero-variance transition probabilities in (5), you observe that they implement the rare-event probabilities $\gamma^{(\epsilon)}(x)$ in the ugly states. This is knowledge that we actually do not know and were going to estimate! However, this observation may lead to efficient estimators as we shall see later.

Based on well-known Markov chain theory on absorption probabilities, the rare-event probabilities $\{\gamma^{\epsilon)}(x), x \in S\}$ satisfy the equation

$$\gamma^{(\epsilon)}(x) = \sum_{y\in S^{(\epsilon)}} p^{(\epsilon)}(x,y)\gamma^{(\epsilon)}(y),\ x \in U^{(\epsilon)}, \tag{6}$$

with boundary conditions $\gamma^{(\epsilon)}(x) = 0$ for $x \in G^{(\epsilon)}$, and $\gamma^{(\epsilon)}(x) = 1$ for $x \in B^{(\epsilon)}$. 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 $S^{(\epsilon)} = \{0,1,\ldots\}$ with ± 1 jumps and state-dependent jump probabilities $p^{(\epsilon)}(x, x\pm 1)$. Let f be a Lyapunov function satisfying (6), and define $\Delta f(x) \doteq f(x+1) - f(x)$. For the random walk we get (in the first equivalence it is used that $p^{(\epsilon)}(x,x+1) + p^{(\epsilon)}(x,x-1) = 1$),

$$\begin{aligned} & f(x) = p^{(\epsilon)}(x,x+1)f(x+1) + p^{(\epsilon)}(x,x-1)f(x-1)\\ \Leftrightarrow\ & p^{(\epsilon)}(x,x+1)\Delta f(x+1) = p^{(\epsilon)}(x,x-1)\Delta f(x)\\ \Leftrightarrow\ & \Delta f(x+1) = \frac{p^{(\epsilon)}(x,x-1)}{p^{(\epsilon)}(x,x+1)}\Delta f(x).\end{aligned}$$

Iterating, this leads to Lyapunov solutions

$$f(x) = \Delta f(1)\sum_{k=1}^{x}\prod_{y=1}^{k-1}\frac{p^{(\epsilon)}(y,y-1)}{p^{(\epsilon)}(y,y+1)},\ \ x = 1,2,\ldots \tag{7}$$

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

Importance sampling of the $M/M/1$ queue and the repair sytem is implented with the zero-variance transition probabilities in (5), using 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 the zero-variance change of measure. The approach could be to approximate it. A successful approach goes as follows. In stead of constructing the Lyapunov function exactly, we might try to approximate it, or simplify the Lyapunov equation (6), or solve it with an inequality. The approximated function is denoted by $v(x)$, and the transition probabilities for the change of measure become

$$q^{(\epsilon)}(x, y) = \frac{p^{(\epsilon)}(x, y)\, v(y)}{\sum_{y \in S^{(\epsilon)}} p^{(\epsilon)}(x, y)\, v(y)}, \quad (x \in S^{(\epsilon)}). \tag{8}$$

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

$$p^{(\epsilon)}(x-1, x)v(x) = p^{(\epsilon)}(x, x-1)v(x-1).$$

Which leads to

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

With this approximation in the transition probabilities (8) for the $M/M/1$ queue, we get the change of measure

$$q^{(\epsilon)}(x, x+1) = 1 - p, \; q^{(\epsilon)}(x+1, x) = p.$$

Executing importance sampling simulation with this change of meaure we find 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.

Also importance sampling simulation of the repair system has been executed with this zero-variance approximation. Here 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.