Advanced Stochastic Processes.

David Gamarnik

LECTURE 22
Fluid model of a G/G/1 queueing system

Lecture outline

- Stationary distribution of RBM
- Fluid model of G/G/1 queueing system

22.1. Stationary distribution of RBM

In the previous lecture we showed that the distribution of $RBM(\theta, \sigma^2)$ is exponential with parameter $-2\theta/\sigma^2$, when $\theta < 0$. In fact, as is not surprising, this distribution is also *stationary* distribution of the RBM. First we define stationarity for any stochastic process.

Definition 22.1. A stochastic process $X(t)$ is defined to be stationary if for any $t_1 < t_2 < \cdots < t_k$ and t, the joint distribution of $(X(t_1+t), \ldots, (X(t_k+t))$ is the same as of $(X(t_1), \ldots, (X(t_k))$.

Observe that when the process is Markovian, it suffices to require that $X(t+s) \overset{d}{=} X(t)$ for any t, s. The Proposition states that the RBM is a Markovian process.

Proposition 1. The exponential distribution π with parameter $2\theta/\sigma^2$ is the unique stationary distribution of the RBM when $\theta < 0$.

Proof. Since, as we said, the RBM process is Markovian, it suffices to establish that the distribution of $Z(0)$ and $Z(t)$ is the same under π. We fix t. Portmentau Theorem establishes that weak convergence of measures occurs if and only if convergence of expectations holds for every bounded continuous function. This also implies that two measure are the same iff for every bounded continuous function f, its expectation with respect to two measures is the same. Thus it suffices to show that $\mathbb{E}_\pi[f(Z(0))] = \mathbb{E}_\pi[f(Z(t))]$ for every bounded continuous f. We have established that π is the limiting distribution of RBM. Namely, $\lim_{s\to\infty} \mathbb{E}[f(Z(s))] = \mathbb{E}_\pi[f(Z(0))]$. But

$$\lim_{s\to\infty} \mathbb{E}[f(Z(s))] = \lim_{s\to\infty} \mathbb{E}[f(Z(s+t))] = \lim_{s\to\infty} \mathbb{E}[\mathbb{E}_{Z(s)}[f(Z(s+t))]]$$

Thus

$$\lim_{s\to\infty} \mathbb{E}[\mathbb{E}_{Z(s)}[f(Z(s+t))]] = \mathbb{E}_\pi[f(Z(0))].$$

Let $g(x) = \mathbb{E}_x[f(Z(t))]$. We just showed

$$\lim_{s\to\infty} \mathbb{E}[g(Z(s))] = \mathbb{E}_\pi[f(Z(0))]. \tag{22.2}$$

We claim that $g(x)$ is a bounded continuous function. It is bounded since f is bounded. It is continuous follows from the following fact given any function $x \in D$ if we change its starting point $x(0)$ to a different point $\hat{x}(0) = \hat{x}$ by value $\|x(0) - \hat{x}(0)| \le \delta$, but otherwise leave the function intact, then the function $\hat{x}$ satisfies $\|\hat{x} - x\|_T = |x(0) - \hat{x}(0)| < \delta$. The reflected process $z = \Phi(x)$ is a Lipshitz continuous (with constant 2) image of x. Thus changing the starting point by amount δ changes creates a new process $\hat{z}$ such that $\|\hat{z} - z\|_T \le 2|x(0) - \hat{x}(0)| = 2\delta$. This means that the expected value $\mathbb{E}_x[f(Z(s+t))]$ changes also by at most δ and indeed $g(x)$ is a continuous function. But then Portmentau Theorem, the fact $Z(s) \Rightarrow \pi$ imply that

$$\lim_{s\to\infty} \mathbb{E}[g(Z(s))] = \mathbb{E}_\pi[g] = \mathbb{E}_\pi[Z(t)]. \tag{22.3}$$

Combining with (22.3) we complete the proof of stationarity. □

There is an alternative derivation of the fact that exponential distribution with parameter $-2\theta/\sigma^2$ is stationary distribution of the RBM. It uses Ito formula, see Section 6 of Chen and Yao [**1**] in the course packet.

22.2. Convergence of reflected processes

For the following discussion we will be considering convergence of functions in $D = D[0,\infty)$ uniformly on compact sets (u.o.c.). Recall, that in the context of this space $x_n \in D$ converges u.o.c. to $x \in D$ if for every $T > 0$, we have $\|x_n - x\|_T \to 0$.

We now pose the following question: if a sequence of functions (processes) x_n converges to x does the same apply to the reflected processes z_n and z? The answer is yes, as we will establish soon. The importance of this property stems from the fact that we will be able to approximate process $X(t)$ corresponding to the queueing process by a Brownian motion. As a result we can approximate the workload process $Z(t)$ by an RBM.

Lemma 22.4. *Given sequences $x_n \in D$, suppose $x_n \to x$ u.o.c. Let*

$$\begin{aligned} y_n &= \Psi(x_n) \\ z_n &= x_n + y_n \\ y &= \Psi(x) \\ z &= x + y \end{aligned}$$

Then $y_n \to y$ and $z_n \to z$ u.o.c.

Proof. The proof follows from reflection mapping Theorem 21.8 from Lecture 21, in particular, the Lipschitz continuity of the mappings Ψ, Φ. Fix any $T > 0$. We have from Theorem 21.8 that

$$\|y_n - y\|_T = \|\Psi(x_n) - \Psi(x)\|_T \le \|x_n - x\|_T \le \|x_n - x\|_T.$$

It follows that $z_n \to z$ u.o.c. Similar result follows for z_n and z. □

22.3. Fluid model of a G/G/1 queueing network

Our next goal is to establish essentially FSLLN for a queueing system. SLLN says that averages converges to a mean a.s. FSLLN says that functions interpolating averages converge *uniformly* to a mean process which is just a linear function. We now establish a similar result for counting (renewal) processes and "pass" it through the Skorohod mapping. On the output we obtain the following process which approximates the workload process but has a very simple piece-wise linear form. Let $\rho = \lambda/\mu$. Parameter ρ is called average workload or utilization of the queueing system. $Z(t) = Z(0) + (\rho - 1)t$ for $t \leq Z(0)/(1-\rho)$ and $Z(t)$ for larger t, when $\rho < 1$ and $Z(t) = Z(0) + (\lambda/\mu - 1)t$ for all t when $\lambda \geq \mu$. We call this *fluid model* of a queueing system. It does have the following indeed fluid model interpretation. Imagine that you have a queue with a large number n of jobs at initial time 0. As time goes on jobs arrive at rate λ and depart at rate μ. When $\rho < 1$, namely $\lambda < \mu$ the queue length will be decreasing with rate $\lambda - \mu < 0$ and workload will be decreasing with rate $\rho - 1$ and roughly at time $Z(0)/(1-\rho)$, the queue will be closed to empty. Starting from this time the workload will be fluctuating between zero and non-zero, but typically will be much smaller than n. Thus we approximate it by $Z(t) = 0$. When $\rho \geq 1$, however, the queue length will stay roughly the same when $\rho = 1$ and will be linearly increasing when $\rho > 1$. We can draw the following analogy with the fluid system. Imagine water goes through a pipe, which has processing rate μ. If water arrives with rate $\lambda < \mu$ and initially there is some amount x of water in the tank, then the water in the tank will decrease to zero at time $x/(1 - \lambda/\mu)$ and will stay at zero level ever after. But if water rate $\lambda \geq \mu$ then the amount in the tank will stay the same ($\lambda = \mu$) or will start increasing ($\lambda > \mu$).

We begin by strengthening FSLLN. In particular, we now show that SLLN implies FSLLN as well as FSLLN for the corresponding renewal process. Thus, consider a sequence X_n of r.v. Let $S_n = \sum_{1\leq k\leq n} X_k$ and let $N(t) = \max\{n : \sum_{1\leq j\leq n} X_j \leq t\}$ be the corresponding counting process.

Theorem 22.5 (FSLLN for renewal processes). *Suppose $S_n/n \to \bar{m}$ almost surely, for some constant $\bar{m}$. Then $S(\lfloor nt\rfloor)/n \to \bar{m}t$ and $N(nt)/n \to \bar{m}^{-1}t$ almost surely u.o.c.*

Proof. We first prove that $S(\lfloor nt\rfloor)/n \to \bar{m}t$ a.s. point-wise for each fixed t. For this we simply observe that

$$\frac{S(\lfloor nt\rfloor)}{n} = \frac{S(\lfloor nt\rfloor)}{\lfloor nt\rfloor}\frac{\lfloor nt\rfloor}{n} \to \bar{m}t$$

a.s. as $n \to \infty$ since $S_n/n \to \bar{m}$ a.s. The proof of u.o.c. convergence follows exactly the same argument as the proof of FSLLN – Theorem 20.5 from Lecture 20. (There we considered a continuous interpolation of S_n instead of simply $S_{\lfloor nt\rfloor}$ but the argument is the same). We now prove that for every t there holds $\frac{N(nt)}{n} \to \bar{m}^{-1}t$ a.s. The proof of u.o.c. again follows the same line as the proof of Theorem 20.5. We have

$$S_{N(nt)} \leq nt \leq S_{N(t)+1},$$

implying

$$\frac{S_{N(nt)}}{N(nt)} \leq \frac{nt}{N(nt)} \leq \frac{S_{N(nt)+1}}{N(nt)+1}\frac{N(nt)+1}{N(nt)}.$$

The assumption $S_n/n \rightarrow \bar{m}$ a.s. implies that $N(nt) \rightarrow \infty$ a.s. as $n \rightarrow \infty$ and therefore $\frac{N(nt)+1}{N(nt)} \rightarrow 1$ a.s. Then we obtain that a.s. $\frac{S_{N(nt)}}{N(nt)} \rightarrow \bar{m}$ a.s. and $\frac{S_{N(nt)+1}}{N(nt)+1} \rightarrow \bar{m}$ a.s. Combining, we obtain that $\frac{nt}{N(nt)} \rightarrow \bar{m}$ a.s. or $\frac{N(nt)}{n} \rightarrow \bar{m}^{-1}t$. □

We now return to the queueing processes. We consider the following setting. Initial queue length is assumed to be large: $Q(0) = \lfloor qn \rfloor$ for some non-negative real q and large positive integer n.

We would like to analyze the queue length, workload and idling time processes at time scale comparable to n: $Q(nt), Z(nt), I(nt)$. To obtain some limiting statement we also rescale these values by n and thus introduce

$$\bar{Q}^n(t) = \frac{Q(nt)}{n}, \ \bar{Z}^n(t) = \frac{Z(nt)}{n}, \ \bar{I}^n(t) = \frac{I(nt)}{n}$$

Theorem 22.6 (Fluid model of a G/G/1 queueing system)**.** *Given a G/G/1 queueing system with arrival rate λ and service rate μ, Consider the sequence of processes $(Q(t), Z(t), B(t))$ corresponding to $Q^n(0) = \lfloor qn \rfloor$, for some $q \in \mathbb{R}_+$. Then*

$$\lim_{n\rightarrow\infty} \bar{Z}^n(t) = (\frac{q}{\mu} + (\rho - 1)t)^+, \tag{22.7}$$

$$\lim_{n\rightarrow\infty} \bar{I}^n(t) = (-\frac{q}{\mu} + (1 - \rho)t)^+, \tag{22.8}$$

$$\lim_{n\rightarrow\infty} \bar{Q}^n(t) = (q + (\lambda - \mu)t)^+ \tag{22.9}$$

a.s. u.o.c.

Before we prove the result, let us interpret it. First, while we assumed that our G/G/1 queueing system has i.i.d. interarrival and service times, as we will see in the proof the result of the theorem holds under any conditions provided that arrival process $A(t)$ and service process $S(t)$ satisfy FSLLN: $A(nt)/n \rightarrow \lambda, S(nt)/n \rightarrow \mu t$ for some constant $\lambda, \mu > 0$. This covers the far larger class of arrival and service processes than covered by i.i.d. case.

To obtain the limiting behavior we essentially rescale the time and space by the same factor n by which we rescale the initial queue length. In this case the theorem says the following: when $\rho < 1$, the rescaled queue length and workload processes drop to zero at the same time $q/(\mu - \lambda)$ and from then on stay at zero level. The idle time process is zero up until this time and from then on has rate $1 - \rho$. Namely, the server works ρ percent of time and idles the rest of the time. On the other hand, when $\rho \geq 1$, the rescaled queue length and workload processes become linear processes with non-negative rate $\lambda - \mu$ and $\rho - 1$ respectively. Also the idling process is always zero - the server works "almost" all the time. We say "almost" because we only know that the limit is zero.

We will prove this theorem in the next lecture.

22.4. Additional reading materials

- Chapter 6 of Chen & Yao book [**1**] from the course packet.

BIBLIOGRAPHY

1. H. Chen and D. Yao, *Fundamentals of queueing networks: Performance, asymptotics and optimization*, Springer-Verlag, 2001.