

Advanced Stochastic Processes.

David Gamarnik

LECTURE 23

Fluid model of a G/G/1 queueing system

Lecture outline

- Fluid model of G/G/1 queueing system (continued)
- G/G/1 in heavy-traffic

23.1. Fluid model of a G/G/1 queueing system

We continue from previous lecture and prove Theorem 21.6.

Proof. Recall that $Z(t) = V(Q(0) + A(t)) - B(t)$ which we again rewrite as

$$\begin{aligned} Z(t) &= V(Q(0) + A(t)) - t + Y(t) \\ &= X(t) + Y(t), \end{aligned}$$

where $X(t) = V(Q(0) + A(t)) - t$. Then, after rescaling, we obtain

$$\bar{Z}^n(t) = \frac{Z(nt)}{n} = \frac{X(nt)}{n} + \frac{I(nt)}{n}$$

Let $\bar{X}^n(t) = \frac{X(nt)}{n}$, $\bar{I}^n(t) = \frac{I(nt)}{n}$. Then $\bar{X}^n, \bar{I}^n, \bar{Z}^n$ are the solutions to the Skorohod mapping problem from Theorem 21.8. Indeed $\bar{X}^n \in D$, and does not have downward jumps, $d\bar{I}^n(t) \geq 0$, $\bar{I}^n(0) = 0$ and $\bar{Z}^n(t)d\bar{I}^n(t) = 0$ as the same holds for Z and I . We claim that $\bar{X}^n(t)$ has a very simple limit:

Lemma 23.1. *The following converges holds almost surely u.o.c.*

$$\bar{X}^n(t) \rightarrow \frac{q}{\mu} + (\rho - 1)t.$$

Before we establish this, let us see how this implies the result of the theorem. We use Lemma 22.4 with $\theta_n = \theta = 0, x_n = \bar{X}^n(t)$ to conclude that $\bar{I}^n(t) \rightarrow \Psi(\frac{q}{\mu} + (\rho - 1)t)$ a.s. u.o.c. and

$\bar{Z}^n(t) \rightarrow \frac{q}{\mu} + (\rho - 1)t + \Psi(\frac{q}{\mu} + (\rho - 1)t)$ also a.s. u.o.c. But

$$\begin{aligned}\Psi(\frac{q}{\mu} + (\rho - 1)t) &= \sup_{0 \leq s \leq t} \left(-\frac{q}{\mu} - (\rho - 1)s \right)^+ \\ &= \left(-\frac{q}{\mu} - (\rho - 1)t \right)^+, \end{aligned}$$

implying

$$\bar{I}^n(t) \rightarrow \left(-\frac{q}{\mu} - (\rho - 1)t \right)^+$$

a.s. u.o.c. As for \bar{Z}^n , we obtain

$$\bar{Z}^n(t) \rightarrow \frac{q}{\mu} + (\rho - 1)t + \Psi(\frac{q}{\mu} + (\rho - 1)t) = \left(\frac{q}{\mu} + (\rho - 1)t \right)^+$$

also a.s. u.o.c. (we used a simple identity $x + (-x)^+ = x^+$). This concludes the derivation (22.7) and (22.8). In order to derive the corresponding limit (22.9) for the queue length process, we recall that $Q(t) = Q(0) + A(t) - S(B(t)) = Q(0) + A(t) - S(t - I(t))$. Since $I(nt)/n \rightarrow \left(-\frac{q}{\mu} - (\rho - 1)t \right)^+$ a.s. u.o.c. then

$$B(nt)/n \rightarrow t - \left(-\frac{q}{\mu} - (\rho - 1)t \right)^+.$$

By direct inspection, we can see that the expression on the right-hand side is simply equal to t for all $t \geq 0$ when $\rho \geq 1$ and equal to t for $t \leq q/(\mu(1 - \rho))$ and $q/\mu + \rho t$ for $t \geq q/(\mu(1 - \rho))$, when $\rho < 1$. Using the result of a problem 1 from HW 6, this implies that when $\rho \geq 1$, $S(nt - I(nt))/n \rightarrow \mu t$ a.s. u.o.c., and $S(nt - I(nt))/n \rightarrow \mu t$ for $t \leq q/(\mu(1 - \rho))$ and $S(nt - I(nt))/n \rightarrow \mu(q/\mu + \rho t) = q + \lambda t$ for $t \geq q/(\mu(1 - \rho))$. On the other hand $A(nt)/n \rightarrow \lambda t$ a.s. u.o.c. Combining, we obtain that

$$\bar{Q}^n(t) = \frac{Q(nt)}{n} \rightarrow q + \lambda t - \mu t$$

when $\rho \geq 1$ and

$$\begin{aligned}\bar{Q}^n(t) &= \frac{Q(nt)}{n} \rightarrow q + \lambda t - \mu t, & \text{when } t \leq \frac{q}{\mu(1 - \rho)} \\ \bar{Q}^n(t) &= \frac{Q(nt)}{n} \rightarrow q + \lambda t - (q + \lambda t) = 0, & \text{when } t > \frac{q}{\mu(1 - \rho)}.\end{aligned}$$

We may summarize this as $\bar{Q}^n(t) \rightarrow (q + \lambda t - \mu t)^+$ a.s. u.o.c. This concludes the proof of the theorem.

The unfinished business is proof of Lemma 23.1

Proof of Lemma 23.1. Since interarrival and service times are i.i.d. then, applying Theorem 22.5, $A(nt)/n \rightarrow \lambda t$, $V(nt)/n \rightarrow \mu^{-1}t$ a.s. u.o.c. This means in particular that $(Q(0) + A(nt))/n \rightarrow q + \lambda t$ a.s. u.o.c. In HW 6 you are asked to establish that this implies $V(Q(0) + A(nt))/n \rightarrow (q + \lambda t)\mu^{-1} = q\mu^{-1} + \rho t$ (HW problem covers the case $Q(0) = 0$ but the extension is immediate). We conclude that $\bar{X}^n(t) \rightarrow q\mu^{-1} + (\rho - 1)t$. \square

\square

23.2. G/G/1 queue in heavy-traffic

23.2.1. Preliminary discussion

The fluid model analysis given by Theorem 22.6 is essentially FSSLN applied to a queueing system. Unfortunately, at this SLLN type scaling $X(nt)/n$ of various underlying processes, we cannot say much about how the workload $Z(t)$ and queue length $Q(t)$ behave in the limit t , since in this scaling they simply converge to zero when $\rho < 1$. We would like to obtain a more refined understanding of the behavior of $Z(t), Q(t)$ in the limit when t is large (that is in steady-state), and for this we resort to CLT type of rescaling. Basically everywhere instead of considering $X(nt)/n$ for various processes X we will consider instead $X(nt)/\sqrt{n}$.

First, we need an analogue of FCLT for counting processes. We have established in Lecture 20, that, roughly speaking, if $X_n, n \geq 1$ is an i.i.d. sequence with zero mean and variance σ^2 , then for $S_n = \sum_{1 \leq k \leq n} X_k$ the process $S_{[nt]}/n$ converges weakly to a Brownian motion $W(t)$ with zero mean and standard deviation σ^2 . We have derived this for the continuous interpolation function rather than $S_{[nt]}$, but with a little extra technicalities, it can be shown that this also applies to $S_{[nt]}$. In case X_n are not zero mean, the same result applies by simply subtracting the mean from this i.i.d. sequence. What about the corresponding counting process $N(t) = \max\{n : S_n \leq t\}$, when X_n is non-negative? It turns out that the following FCLT holds. The proof of this result can be found in Chen & Yao [1].

Theorem 23.2 (FCLT for renewal processes). *For every $T > 0$, the process $N(t)$ as defined on $D[0, T]$ satisfies the following weak convergence with respect to the uniform metric $\|\cdot\|_T$:*

$$\frac{N(nt) - \mu nt}{n} \Rightarrow \mu^{\frac{3}{2}} \sigma W(t),$$

where W is the standard Brownian motion.

Armed with this result, we can try to say something about the queueing system. Let σ_a^2, c_a denote respectively the variance of the interarrival times and the coefficient of variation: $c_a = \sigma_a / \mathbb{E}[u_1] = \lambda \sigma_a$. Similarly, let σ_s^2, c_s denote respectively the variance of the service times and the coefficient of variation $c_s = \mu \sigma_s$.

Unfortunately, in order to obtain meaningful result we also need to consider a queueing system where ρ is not a constant but also "scales" like $1 - \frac{\theta}{\sqrt{n}}$. In order to appreciate why this is the case, let us do analysis for the case $\rho < 1$ is a constant. For simplicity assume that $Q(0) = Z(0) = 0$. We recall that $Z(t) = V(Q(0) + A(t)) - B(t) = X(t) + I(t)$, where $X(t) = V(Q(0) + A(t)) - t$. In our special case $Q(0) = 0$ we have

$$X(t) = V(A(t)) - t = V(A(t)) - \mu^{-1}A(t) + \mu^{-1}(A(t) - \lambda t) + \rho t - t.$$

From FSSLN $A(nt)/n \rightarrow \lambda t$ a.s. u.o.c. Applying Donsker's Theorem for cumulative service times $V(n)$:

$$\frac{V([nt]) - \mu^{-1}[nt]}{\sqrt{n}} \Rightarrow \sigma_s W(t) = \mu^{-1} c_s W_1(t),$$

where W_1 is a standard Brownian motion.

Combining the two convergence statements as well as the fact that the interarrival times and service times are independent from each other, we obtain that

$$\frac{V(A(nt)) - \mu^{-1}A(nt)}{\sqrt{n}} = \frac{V(A(nt)) - \mu^{-1}A(nt)}{\sqrt{A(nt)}} \sqrt{\frac{A(nt)}{n}} \Rightarrow \lambda^{\frac{1}{2}} \mu^{-1} c_s W_1(t),$$

Applying Theorem 23.2 for the arrival process $A(t)$ we obtain

$$\frac{A(nt) - \lambda nt}{n} \Rightarrow \lambda^{\frac{1}{2}} c_a W_2(t),$$

where $W_2(t)$ is a standard Brownian motion. Note that W_1, W_2 are independent, since the first is obtained from the service times, second is obtained from the interarrival times, and the two sequences are independent. Putting this together we obtain that $\frac{X(nt)}{\sqrt{n}}$ behaves like

$$\lambda^{\frac{1}{2}} \mu^{-1} c_s W_1(t) + \mu^{-1} \lambda^{\frac{1}{2}} c_a W_2(t) - (1 - \rho) \frac{nt}{\sqrt{n}}.$$

So when $\rho < 1$ is constant we have a term $-(1 - \rho)\sqrt{nt}$ which diverges to $-\infty$ and "kills" the Brownian term.

But if we at the same time make ρ approach unity at the rate such that $(1 - \rho) \approx \frac{\theta}{\sqrt{n}}$, then we obtain a non-trivial scaling behavior: Brownian motion with drift $-\theta$ and variance $\lambda\mu^{-2}(c_s^2 + c_a^2)$.

23.2.2. G/G/1 heavy-traffic theorem

How can we arrange for $\rho \approx 1 - \theta/\sqrt{n}$. One natural way to do this is as follows. Assume we have a queueing system such that $\lambda = \mu, \rho = 1$, namely, it is critically loaded. Fix a parameter $\theta > 0$ and let $n \in \mathbb{N}$ be large. Let $\rho_n = 1 - \frac{\theta}{\sqrt{n}}$ and consider a rescaled sequence of interarrival times $u_k^n = \rho_n^{-1} u_k, k \geq 1$. Denote by $A^n(t)$ the corresponding arrival process. Let also $S^n(t) = S(t)$ denote the service counting process (unaffected by n), and let $Q^n(t), Z^n(t), B^n(t), I^n(t)$ be the corresponding queue length, workload, busy time and idle time processes. In effect we are considering a sequence of G/G/1 type queueing systems. For this system we obtain that the arrival rate $\lambda_n = \rho_n \lambda$ and average utilization is ρ_n .

We now state formally the result we are about to derive. Recall that $RBM(\theta, \sigma^2)$ stands for Reflected Brownian Motion with drift θ and variance σ^2 . We also let $BM(\theta, \sigma^2)$ denote a Brownian motion with the same parameters.

Theorem 23.3. *Consider a sequence of G/G/1 queueing systems with $\rho_n = 1 - \theta/\sqrt{n}$. Suppose $\frac{Q(0)}{\sqrt{n}} \rightarrow q$ for some $q \geq 0$. Then the following weak convergence takes place in $D[0, T]$ with respect to $\|\cdot\|_T$ for every $T > 0$*

$$\begin{aligned} \frac{Z(nt)}{\sqrt{n}} &\Rightarrow RBM(-\theta, \lambda^{-1}(c_a^2 + c_s^2)), \\ \frac{Q(nt)}{\sqrt{n}} &\Rightarrow RBM(-\theta, c_a^2 + c_s^2), \\ \frac{I(nt)}{\sqrt{n}} &\Rightarrow \sup_{0 \leq s \leq t} (-W(s))^+, \end{aligned}$$

where $W = BM(-\theta, \lambda(c_a^2 + c_s^2))$.

We recall that $\lambda = \mu$ in this setting, so we could have used μ in the above expressions. We will prove this result in the next lecture.

In the remainder of this lecture we state without a proof the following fundamental implication of Theorem 23.3. Our setting is again the same as in Theorem 23.3: we have sequence of queueing systems G/G/1 with traffic intensities $\rho_n = 1 - \theta/\sqrt{n}$, which is achieved by considering a queueing system with $\rho = 1$, and rescaling interarrival times by ρ_n^{-1} , and leaving service times intact. We call this sequence of queueing systems simply G/G/1 queueing system in heavy-traffic. We let $\text{Exp}(z)$ denote exponential distribution with parameter z .

Theorem 23.4 (Heavy-traffic formula for G/G/1 queueing system). *Consider a G/G/1 queueing system in heavy traffic with traffic intensity $\rho_n = 1 - \theta/\sqrt{n}$. Then the corresponding queue length and workload processes $Q^n(t), Z^n(t)$ converge in distribution to a steady state distribution as $t \rightarrow \infty$. Moreover, if $Q^n(\infty), Z^n(\infty)$ denote r.v. selected according to this distribution, then*

$$\begin{aligned} \frac{Z^n(\infty)}{\sqrt{n}} &\Rightarrow \text{Exp}\left(\frac{2\lambda\theta}{c_a^2 + c_s^2}\right) \\ \frac{Q^n(\infty)}{\sqrt{n}} &\Rightarrow \text{Exp}\left(\frac{2\theta}{c_a^2 + c_s^2}\right), \end{aligned}$$

as $n \rightarrow \infty$.

We will not prove this result, just mention some underlying ideas. First it is not obvious that the processes $Z(t), Q(t)$ converge to a stationary regime when $\rho < 1$. This, however can be established using various ways, for example using so-called Lindeley's recursion. Provided that convergence to stationarity holds, the main building block is to use Theorem 23.3 which states that the processes $Z(t), Q(t)$ look "increasingly" like RBM as $\rho_n = 1 - \theta/\sqrt{n} \rightarrow 1$ and Theorem 21.12 which states that stationary distribution of an RBM is exponential.

Although, as we know, convergence in distribution does not necessarily imply convergence of expectations (or higher moments) it turns out that it does in this case, and, in particular

$$(23.5) \quad \lim_n \frac{\mathbb{E}[Z^n(\infty)]}{\sqrt{n}} = \frac{c_a^2 + c_s^2}{2\lambda\theta},$$

$$(23.6) \quad \lim_n \frac{\mathbb{E}[Q^n(\infty)]}{\sqrt{n}} = \frac{c_a^2 + c_s^2}{2\theta},$$

These two formulas have a great practical use, as they predict a performance of a queueing system using only first and second moments of interarrival and service times. Let us do a quick example of applying these formulas. Suppose we have a queueing system where arrivals occur on average every 3.4 seconds and the variance of interarrival times is estimated to be 2. Suppose the service time takes deterministically 3.3 seconds. What is our prediction of average queue length and average workload? It seems at first that we are in trouble here: the formulas give predictions for $\mathbb{E}[Z^n(\infty)]/\sqrt{n}, \mathbb{E}[Q^n(\infty)]/\sqrt{n}$ and n is an artifact of the analysis here, but not a parameter of the model. But the approximations (23.5) and (23.6) suggest

$$\begin{aligned} \mathbb{E}[Z^n(\infty)] &= \sqrt{n} \frac{c_a^2 + c_s^2}{2\lambda\theta} = \frac{c_a^2 + c_s^2}{2\lambda(1 - \rho_n)} \\ \mathbb{E}[Q^n(\infty)] &= \sqrt{n} \frac{c_a^2 + c_s^2}{2\theta} = \frac{c_a^2 + c_s^2}{2(1 - \rho_n)} \end{aligned}$$

Thus, we have expressions involving only the data of the model. In particular, for our case $\lambda = 1/3.4 = .2941$, $\rho = 3.3/3.4 = .9706$, $c_a^2 = 2/3.4^2 = .1953$ and $c_s^2 = 0$ (since service time is deterministic). Thus we obtain estimates

$$\mathbb{E}[Z(\infty)] \approx \frac{.1953}{2 \cdot .2941 \cdot (1 - .9706)} = 11.29,$$

$$\mathbb{E}[Q(\infty)] \approx \lambda \mathbb{E}[Z(\infty)] = 3.32.$$

Even though, this does not look much as a heavy-traffic system, these types of approximations are often remarkably accurate. Refer to Chen & Yao [1] for a summary of simulations and computations experiments done for a more general class of queueing systems, called Generalized Jackson Networks, to see that even when traffic intensities are around .60, the heavy-traffic theory predictions are quite accurate.

23.3. 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.