IEOR 8100: Topics in OR: Asymptotic Methods in Queueing Theory

Fall 2009, Professor Whitt

Class Lecture Notes: Wednesday, September 9.

Heavy-Traffic Limits for the GI/G/1 Queue

1. The GI/G/1 Queue

We will start by considering this basic model, assuming unlimited waiting space and the first-come first-served service discipline (all customers served in order of arrival). This model is covered in almost all queueing textbooks. See Chapters III, VIII and X in Asmussen (2003), *Applied Probability and Queues*, second ed. A major treatise is Cohen (1982), *The Single-Server Queue*, second ed. Actually, one or two full courses could be devoted to this model and its special cases. However, since we consider the case of general service-time and interarrival-time distributions, the story is quite complicated. The material here might be treated at the very end of a first queueing course, or not until a second queueing course.

It is good to have some understanding about the strong results that are possible when we make extra assumptions about the service-time and interarrival-time distributions. Here are some things to learn: the Kendall notation. the elementary $M/M/1$ model - a birth-and-death process. the Pollaczek-Khintchine formulas for the mean and the Laplace transform of the steady-state waiting time in the $M/G/1$ model. the special techniques for other special cases, such as $GI/M/1$, $GI/PH/1$, $PH/G/1$, etc. For example, see Chapter III and Section VIII.5 of Asmussen.

The general model $GI/G/1$ is difficult to analyze. A key relation connects the sequence of waiting times of successive customers, $\{W_n : n \geq 0\}$ to a random walk, i.e., to the sequence of partial sums of i.i.d. random variables. That key relation for the waiting times of successive customers is:

$$W_n = S_n - \min_{0 \leq k \leq n} S_k = \max_{0 \leq k \leq n} \{S_n - S_k\}, \quad n \geq 0, \tag{1}$$

where $S_n \equiv X_1 + \cdots + X_n$ and $X_n \equiv V_{n-1} - U_n$, $n \geq 1$, $W_0 \equiv S_0 \equiv 0$, V_n is the service time of the n^{th} customer and U_n is the interarrival time between the $(n-1)^{st}$ and n^{th} arrival, with a 0^{th} arrival coming to an empty system at time 0. As a consequence (because the random variables X_n being added are i.i.d), for each $n \geq 1$,

$$W_n \stackrel{\mathrm{d}}{=} M_n \equiv \max_{0 \leq k \leq n} S_k, \quad n \geq 0. \tag{2}$$

That is, for each $n \geq 1$, W_n is distributed exactly as the maximum of the first n partial sums. To obtain this simple relation, we chose special initial conditions. It is not difficult to treat the more general case; the limiting steady-state distribution, when well defined, will not be altered by the initial conditions. It is important to note that we do not have equality in distributions of the stochastic processes $\{W_n : n \geq 0\}$ and $\{M_n : n \geq 0\}$. That is easy to see, because $\{M_n : n \geq 0\}$ has nondecreasing sample paths, while $\{W_n : n \geq 0\}$ does not.

The theory of general random walks $\{S_n\}$ and their successive maxima $\{M_n\}$ is well developed; see Chapter VIII of Asmussen (2003) and Chapter 8 of Chung (1974), *A Course in*

Probability Theory. The process $\{M_n\}$ is considered in §8.5 of Chung. Theorem 8.5.1 in Chung gives the *Spitzer identity*

$$\sum_{n=0}^{\infty} r^n E[exp(itM_n)] = exp(\sum_{n=1}^{\infty}(r^n/n)E[exp(itS_n^+)], \tag{3}$$

where $a^+ \equiv \max\{a, 0\}$. As a consequence, we have the following expression for the steady-state random variable $M \equiv \max\{S_n\}$. This is going to be finite if $E[X_1] < 0$. However, it is known that M is finite if and only if

$$\sum_{n=1}^{\infty}(1/n)P(S_n > 0) < \infty. \tag{4}$$

If condition (4) is satisfied, then the characteristic function of the random variable M is given by

$$E[exp(itM)] = exp(\sum_{n=1}^{\infty}(1/n)(E[exp(itS_n^+)] - 1) \tag{5}$$

The mean and variance of M are given by

$$E[M] = \sum_{n=1}^{\infty}(1/n)E[S_n^+] \quad \text{and} \quad Var[M] = \sum_{n=1}^{\infty}(1/n)E[(S_n^+)^2]. \tag{6}$$

We also remark that a CLT-type limit theorem for M_n was established in the seminal paper by Erdös and Kac (1946), cited below and distributed. That CLT is now regarded as an elementary consequence of Donsker's FCLT. Since the waiting times in the $GI/G/1$ queue are closely linked to the successive maxima, the waiting times W_n in the $GI/G/1$ queue are intimately connected to the origins of Donsker's theorem.

We will be studying the heavy-traffic limit for W_n. But it is important to realize that a lot is known about this process and its steady-state distribution. In general, the steady-state distribution was characterized in terms of its Laplace transform by Pollaczek in 1952. (The content of Pollaczek (1952) is closely related to §8.5 of Chung (1974).) Given that the Laplace transforms of the interarrival time and service time are known, that transform can be numerically inverted to calculate the steady-state waiting time distribution and its moments; see

Abate, J., G. L. Choudhury and W. Whitt, Calculation of the $GI/G/1$ waiting-time distribution and its cumulants from Pollaczek's formulas. *Arkiv für Elektronik und Übertragungstechnik* (AËU, *International Journal of Electronics and Communications*), 47 (1993) 311-321.

The Pollaczek expression does not look like the Spitzer identity, but they can be (have been) linked. The Pollaczek representation is a contour integral representation for the transform of the steady-state waiting time W or, equivalently, for M. In particular,

$$E[exp(-sW)] = E[exp(-sM)] = exp\left(-\frac{1}{2\pi i}\int_C \frac{s}{z(s-z)}\log[1-\phi(-z)]dz\right) , \tag{7}$$

where s is a complex number with real part $Re(s) \geq 0$, C is a contour to the left of, and parallel to the imaginary axis, and to the right of any singularities of $\log[1-\phi(-z)]$ in the left half plane. Of course, $i \equiv \sqrt{-1}$. The function ϕ is the transform of $X \equiv V - U$, where V is a generic service time and U is a generic interarrival time, i.e.,

$$\phi(z) \equiv E[exp(s(V-U))] \equiv \int_{-\infty}^{\infty} e^{zt} dG(t) = E[exp(sV)]E[exp(-sU)], \tag{8}$$

where G is the cdf of $V-U$, which we assume is analytic for complex z in the strip $Re(z) < \delta$ for some positive δ. All this is reviewed in the ACW93 paper above.

Clearly, these expressions are complicated. They greatly simplify if we make additional assumptions, such as assuming that either V or U has an exponential distribution. Then we get the $M/G/1$ and $GI/M/1$ models.

2. **Donsker's Theorem and the Continuous Mapping Theorem**

We now present Donsker's theorem. For that purpose, let $\lfloor a \rfloor$ be the *floor function*, i.e., the greatest integer less than a.

Theorem 0.1 (*Donsker's FCLT or invariance principle*) *Let $\{X_n : n \geq 1\}$ be a sequence of i.i.d random variables with $E[X_1] = m$ and $Var(X_1) = \sigma^2 < \infty$. Let $S_n \equiv X_1 + \cdots + X_n$, $n \geq 1$, and $S_0 \equiv 0$. Let*

$$\mathbf{S}_n(t) \equiv \frac{S_{\lfloor nt \rfloor} - mnt}{\sigma\sqrt{n}} \quad for \quad 0 \leq t \leq 1. \tag{9}$$

so that $\mathbf{S_n}$ a random element of the function space $D \equiv D([0,1], \mathbb{R})$. Then

$$\mathbf{S}_n \Rightarrow \mathbf{B} \quad in \quad D, \tag{10}$$

where $\mathbf{B}$ is standard Brownian motion (BM) on the interval $[0,1]$.

The continuous mapping theorem then gives us the following corollary. Much depends on the topological structure on the spcae D, but we do not specify that here.

Corollary 0.1 (*continuous image*) *Let f be a continuous function mapping D into another space, e.g., into D or $\mathbb{R}$. Under the assumptions of Theorem* 0.1*,*

$$f(\mathbf{S}_n) \Rightarrow f(\mathbf{B}) \quad as \quad n \to \infty. \tag{11}$$

A specific example of a continuous function is $f : D \to \mathbb{R}$, where $f(x) \equiv \sup\{x(t) : 0 \leq t \leq 1\}$. When we apply this function with Donsker's FCLT, we get a CLT for M_n. Let $\stackrel{\mathrm{d}}{=}$ mean equal in distribution.

Corollary 0.2 (*CLT for the successive maxima*) *Let $M_n \equiv \max\{S_k - mk : 0 \leq k \leq n\}$ in the framework of Theorem* 0.1*. Then*

$$\frac{M_n}{\sigma\sqrt{n}} \Rightarrow \sup_{0 \leq t \leq 1} B(t) \stackrel{\mathrm{d}}{=} |N(0,1)| \quad in \quad \mathbb{R}, \tag{12}$$

where $N(0,1)$ is a standard normal random variable, so that

$$P(|N(0,1)| > x) = 2P(N(0,1) > x) = \sqrt{(2/\pi)} \int_x^\infty exp(-y^2/2)dy, \quad x \geq 0. \tag{13}$$

We can combine equation (2) and Corollary 0.2 to obtain a heavy-traffic limit theorem for the waiting time W_n in the $GI/G/1$ queue when the traffic intensity if $\rho = 1$ (the critically loaded case).

Corollary 0.3 (*heavy-traffic limit for W_n in the $GI/G/1$ queue with $\rho = 1$*) *Let W_n be the waiting time of the n^{th} arrival in the $GI/G/1$ queue with finite mean service time $E[V]$ and interarrival time $E[U]$, and traffic intensity $\rho \equiv E[V]/E[U] = 1$ (so that $m \equiv E[X] = 0$). Then*

$$\frac{W_n}{\sigma\sqrt{n}} \Rightarrow |N(0,1)| \quad in \quad \mathbb{R}, \tag{14}$$

where $N(0,1)$ is a standard normal random variable, so that

$$P(|N(0,1)| > x) = 2P(N(0,1) > x) = \sqrt{(2/\pi)} \int_x^\infty exp(-y^2/2)dy, \quad x \geq 0. \tag{15}$$

Readings:

1. Introduction (pp. 1-6) of [B], the 1999 Billingsley book (explains the motivation for studying stochastic-process limits)

2. Erdös, P. and M. Kac, On certain limit theorems of the theory of probability. *Bull. Amer. Math. Soc.* 52 (1946) 292-302. (a seminal paper)

3. Donsker, M., An invariance principle for certain probability limit theorems. *Mem. Amer. Math. Soc.* 6 (1951), 12 pages. (a seminal paper)

4. papers by Prohorov and Skorohod in *Theor. Prob. Appl.* 1 (1956). (seminal papers)

5. the class textbooks [B] and [W]; my book (*Stochastic-Process Limits*, Springer 2002, available online, link on course web page.)

From [W] on GENERAL THEORY

(i) continuous mapping theorem, §3.4, pp. 84-86

(ii) Donsker's FCLT, §4.3, pp. 101-103

(iii) one-dimensional reflection map, formula (5.4) on p. 87; plus §§13.4-13.5, pp. 435-441

6. Background on the classical limit theorems and their proofs:

(a) W lecture notes on transforms from IEOR 6711, 09/11/08. Proof of WLLN and CLT on pages 10-11.

(b) Chung, K. L., *A Course in Probability Theory*, Chapters 4-7: Ch. 4, convergence concepts; Ch. 5, LLN's; Ch. 6, characteristic functions; Ch. 7, CLT.

3. **Heavy-Traffic limit for the waiting time in the GI/G/1 Queue**

We actually are more interested in obtaining heavy-traffic limits for stable queues with traffic intensities strictly less than 1. To do so, we consider a sequence of $GI/G/1$ models indexed by n and let the traffic intensities approach 1 from below as $n \to \infty$. Then the queueing model is stable for each n, and yet we still have a heavy-traffic limit. But we must be careful to let ρ_n approach 1 at just the right rate. It is easy to see what will work by considering Donsker's theorem.

First, note that we introduce a technical complication when we consider a sequence of queueing models. We now have a double sequence or triangular array of random variables. We have a sequence of random variables $\{X_{n,k} : k \geq 1\}$ for each $n \geq 1$. This is a standard framework for the CLT, but it needs to be handled properly. We need to impose regularity conditions on these sequences as $n \to \infty$. A convenient sufficient condition is to assume that the means and variances of the random variable $X_{n,1}$ satisfy $m_n \to m$, $-\infty < m < \infty$, and $\sigma_n^2 \to \sigma^2 < \infty$, plus a condition on a higher absolute moment. In addition, it suffices to have

$$\sup\{E[|X_{n,1}|^{2+\delta}]\} < \infty \tag{16}$$

for some $\delta > 0$. This implies the classical Lindeberg condition. Donsker's theorem is known to extend to this more general setting, e.g., see p. 219 of Parthasarathy (1967), *Probability*

Measures on Metric Spaces, Academic Press. In the queueing setting, we will also assume that $U_{n,1} \Rightarrow U$ and $V_{n,1} \Rightarrow V$ as $n \to \infty$ in $\mathbb{R}$ with $E[U_{n,1}] \to E[U]$ and $E[V_{n,1}] \to E[V]$.

It is easy to see what is needed by considering the translation term in Donsker's theorem, in the double sequence framework That is we can rewrite $\mathbf{S}_n$ as

$$\mathbf{S}_n(t) \equiv \frac{S_{n,[nt]} - m_n nt}{\sigma_n \sqrt{n}} \quad \text{for} \quad 0 \le t \le 1, \tag{17}$$

where $S_{n,k} \equiv X_{n,1} + \cdots X_{n,k}$, $k \ge 1$. the translation term is

$$\mathbf{T}_n(t) \equiv \frac{m_n nt}{\sigma_n \sqrt{n}} \quad \text{for} \quad 0 \le t \le 1. \tag{18}$$

We will want to have this translation term converge to a constant function $c\mathbf{e}$, where $-\infty < c < 0$ and $\mathbf{e}$ is the identity function in D, i.e., $\mathbf{e}(t) = t$, $0 \le t \le 1$. (Asymptotically, there should be negative constant drift.) That will be achieved if and only if

$$m_n \sqrt{n} \to b, \quad -\infty < b < 0 \quad \text{as} \quad n \to \infty, \tag{19}$$

in which case $c = b/\sigma$. Note that this in turn holds if and only if we have the heavy-traffic condition

$$\rho_n \uparrow 1 \quad \text{and} \quad \sqrt{n}(1 - \rho_n) \to \beta \equiv |b/E[U]| \quad \text{as} \quad n \to \infty. \tag{20}$$

The main heavy-traffic limit for W_n establishes convergence to reflected Brownian motion (RBM) for the sequence of $GI/G/1$ models with $\rho_n \uparrow 1$ as $n \to \infty$ under the regularity conditions in (19) or (20). We go beyond Corollary 0.3 in two ways: first, by letting the traffic intensities approach 1 from below according to (20) and, second by establishing a functional central limit theorem for $W_{n,k}$ in D.

Theorem 0.2 (*heavy-traffic limit for the waiting times in the $GI/G/1$ queue for $\rho_n \uparrow 1$*) *Consider a sequence of $GI/G/1$ models. Let $W_{n,k}$ be the waiting time of the k^{th} arrival in the n^{th} $GI/G/1$ queueing system with finite mean service time $E[V_n]$ and finite mean interarrival time $E[U_n]$. Assume the regularity conditions above, including* (19) *or, equivalently,* (20)*. Then*

$$\mathbf{W}_n \Rightarrow \mathbf{W} \equiv \phi(\mathbf{B} + c\mathbf{e}) \quad \textit{in} \quad D, \tag{21}$$

where

$$\mathbf{W}_n(t) \equiv \frac{W_{n,\lfloor nt \rfloor}}{\sigma_n \sqrt{n}}, \quad 0 \le t \le 1 \tag{22}$$

$\mathbf{B} + c\mathbf{e}$ *is BM with drift c, $-\infty < c < 0$, $\phi : D \to D$ is the reflection map*

$$\phi(x) = x + (-x \vee 0\mathbf{e})^{\uparrow} \quad \textit{with} \quad x^{\uparrow}(t) \equiv \sup_{0 \le s \le t} x(s), \quad 0 \le t \le 1. \tag{23}$$

and $\phi(\mathbf{B} + c\mathbf{e})$ is reflected Brownian motion (RBM).

The stronger FCLT result in Theorem 0.2 is useful to establish limits for functionals of the waiting times. For example, we could apply the supremum functional, just as we did in Corollary 0.2.

Corollary 0.4 (*heavy-traffic limit for the maximum waiting time in $GI/G/1$ queues with $\rho_n \uparrow 1$*) *Consider a sequence of $GI/G/1$ models under the conditions of Theorem* 0.2*. Let $W_{n,k}$ be the waiting time of the k^{th} arrival in the n^{th} $GI/G/1$ queueing system. Then*

$$\frac{\max\{W_{n,k} : 0 \le k \le n\}}{\sigma \sqrt{n}} \Rightarrow \sup_{0 \le t \le 1} \phi(\mathbf{B} + c\mathbf{e})(t) \quad \textit{in} \quad \mathbb{R}. \tag{24}$$

It is of course useful to be able to determine the probability distribution of the limit. I have not done that yet here. I am not sure how hard it is.

Theorem 0.2 supports approximating the steady-state random variable, say $W_{n,\infty}$, where $W_{n,k} \Rightarrow W_{n,\infty}$ in $\mathbb{R}$ as $k \to \infty$ for each n, by the steady-state limit of the RBM, Z, where $\phi(\mathbf{B}+c\mathbf{e})(t) \to Z$ in $\mathbb{R}$. It is known that the random variable Z has an exponential distribution. This leads to the heavy-traffic approximation

$$P(W_{n,\infty} > x) \approx e^{-x/m_n}, \quad x \geq 0, \tag{25}$$

where

$$m_n \equiv E[W_{n,\infty}] \approx \frac{\rho E[V_n](c_a^2 + c_s^2)}{2(1-\rho_n)}, \tag{26}$$

with c_a^2 and c_s^2 being the squared coefficient of variation (SCV, variance divided by the square of the mean) of an interarrival time and service time, respectively. Unfortunately, actually proving a limit directly supporting formulas (25) and (26) is not straightforward from Theorem 0.2. See §5.7.2 of [W] for further discussion. For recent work, see Gamarnik, D. and A. Zeevi, "Validity of heavy traffic steady-state approximations in generalized Jackson networks," *Annals of Applied Probability*, 16 (2006) 56-90.

More Readings:

From [W] on HEAVY-TRAFFIC LIMITS FOR W_n

(i) basic idea via fluid queue model, §§5.1-5.4, pp. 137-157

(ii) Brownian approximations, §§5.7.1-5.7.2, pp. 165-170

(ii) single server queue, §§9.2-9.2.1, pp. 288-289, and §§9.3-9.3.2, pp. 292-297

(ii) Brownian approximations, §§5.7.1-5.7.2, pp. 165-170