HEAVY TRAFFIC LIMITS VIA BROWNIAN EMBEDDINGS

Erol A. Peköz
Boston University School of Management
Boston, MA 02215
E-mail: pekoz@bu.edu

Jose Blanchet
Harvard University
Statistics Department
Cambridge, MA 02138
E-mail: blanchet@stat.harvard.edu

For the $GI/GI/1$ queue we show that the scaled queue size converges to reflected Brownian motion in a critical queue and converges to reflected Brownian motion with drift for a sequence of subcritical queuing models that approach a critical model. Instead of invoking the topological argument of the usual continuous-mapping approach, we give a probabilistic argument using Skorokhod embeddings in Brownian motion.

1. INTRODUCTION

Consider a $GI/GI/1$ queue where the mean service time is equal or nearly equal to the mean interarrival time and at least one is nondeterministic. The asymptotic behavior of the queue length Q_n just prior to the nth arrival is usually studied by the "continuous mapping approach" (see, e.g., Whitt [9] and Robert [6]). This approach starts with the observation that

$$Q_n = \left(\sup_{m \le n} \sum_{i=m}^{n} X_i \right)^+,$$

where X_i is the difference between the ith service time and the $(i+1)$st interarrival time. As this is a continuous mapping of a random walk process, which, suitably

© 2006 Cambridge University Press 0269-9648/06 $16.00

https://doi.org/10.1017/S0269964806060360 Published online by Cambridge University Press

scaled, converges to Brownian motion, the continuous-mapping theorem of weak convergence theory (see Billingsley [1]) states that a suitably scaled Q_n converges to the same mapping of a Brownian motion. This argument involves invoking the topology of uniform convergence on the space of continuous functions.

In this article we give a somewhat more direct argument for convergence to reflected Brownian motion using Skorokhod embeddings. This type of argument is one of the ingredients in functional central limit theorems (see Durrett [3]) and has been applied to the $GI/GI/1$ queue in Rosenkrantz [7] to assess the accuracy of Brownian approximations. It is also a key ingredient in what are called "strong approximation" theorems of queuing theory; see Chen and Yao [2]. Although the argument we give here is therefore not new and is straightforward to put together, it seems to have previously appeared in the literature only as an ingredient in other somewhat more complex results. Our contribution is to record and publicize this direct and simple argument as a method to prove convergence to reflected Brownian motion for a $GI/GI/1$ queue. In Section 2 we present the main result and give its proof.

2. MAIN RESULTS

Here we show how a suitably scaled queue length converges to reflected Brownian motion. This type of limit theorem goes back to Kingman [4] and there has been a very large literature developed from this (see, e.g., Whitt [9]). Below we use the notation $R_a(t)$ to denote reflected Brownian motion with downward drift a, defined as

$$R_a(t) = B(t) - at - \inf_{s \le t}(B(s) - as),$$

for standard Brownian motion $B(t)$. We also use $\to_d$ to denote convergence in distribution.

Theorem 1: *Suppose $X_1, X_2, \ldots$ are independent and identically distributed (i.i.d.) random variables with mean zero and variance $\sigma^2 > 0$, and $Q_n = (\sup_{m \le n} \sum_{i=m}^{n}(X_i - c/\sqrt{n}))^+$ for a given c. Then for any $t \ge 0$, we have $Q_{nt}/(\sigma\sqrt{n}) \to_d R_{c/\sigma}(t)$.*

This theorem applied with $c = 0$ corresponds to the setting of a critical $GI/GI/1$ queue where U_n is the service time of the nth customer, V_n is the interarrival time between the nth and $(n+1)$st customer, and $X_n = U_n - V_n$. In this setting, the variable Q_n is the amount of work in the queue just prior to the arrival of the nth customer.

The theorem applied with $c = 1$ corresponds to a setting where we are looking at a sequence of increasingly congested $GI/GI/1$ queuing models with $X_i - 1/\sqrt{n}$ representing the difference between the $(i+1)$st interarrival time and the ith service time in the nth model. In this case each model will have a stable queue, but as $n \to \infty$, these models approach a critical queuing model. The variable Q_n in this case will represent the amount of work in the nth queuing model immediately before the arrival there of the nth customer.

https://doi.org/10.1017/S0269964806060360 Published online by Cambridge University Press

The next result gives convergence in the space of continuous sample paths.

THEOREM 2: *With the above definitions and $Q^*_{nt} = (nt - [nt])Q_{nt+1} + (1 - nt + [nt])Q_{nt}$, we have $Q^*_{n\cdot}/(\sigma\sqrt{n}) \to_d R_{c/\sigma}(\cdot)$, where $\to_d$ denotes weak convergence of the sample paths in the space of continuous sample paths $C[0,\infty)$.*

To prove these results we will make use of the following representation result, due to Skorokhod [8] (see also Obłój [5] or Durrett [3]).

PROPOSITION 3: *If $E[X] = 0$ and $E[X^2] < \infty$, then there is a stopping time T for Brownian motion so that $B(T) =_d X$ and $E[T] = E[X^2]$.*

We also make use of the following well-known scaling lemma (see, e.g., Durrett [3]).

LEMMA 4: *If $B(0) = 0$, then for any $n > 0$,*

$$\{B(s/n), s \geq 0\} =_d \left\{ \frac{1}{\sqrt{n}} B(s), s \geq 0 \right\}.$$

PROOF (of Theorem 1): Proposition 3 states we can create an i.i.d. sequence of stopping times t_i with $T_n = \sum_{i=1}^n t_i$ so we have $B(T_n) = \sum_{i=1}^n X_i$. If we let $W_n(t) = B(n\sigma^2 t)/(\sigma\sqrt{n}) =_d B(t)$ by the scaling lemma (Lemma 4), we have

$$\begin{aligned}\frac{Q_{nt}}{\sigma\sqrt{n}} &= \frac{1}{\sigma\sqrt{n}}\left(B(T_{nt}) - \frac{c[nt]}{\sqrt{n}} - \inf_{0\leq k\leq nt}\left(B(T_k) - \frac{c[k]}{\sqrt{n}}\right)\right)\\ &= W_n\left(\frac{T_{nt}}{n\sigma^2}\right) - \frac{c[nt]}{\sigma n} - \inf_{0\leq k\leq nt}\left(W_n\left(\frac{T_k}{n\sigma^2}\right) - \frac{c[k]}{\sigma n}\right).\end{aligned}$$

For a given $t, \varepsilon > 0$, since $B(t)$ is uniformly continuous on $[0, t]$, we can pick δ small enough so that

$$P(|W_n(x) - W_n(y)| > \varepsilon \quad \text{for some } 0 \leq x \leq y \leq t \text{ with } |x - y| < \delta) < \varepsilon \tag{1}$$

and then we can pick n large enough so that

$$\frac{|c|(k - [k])}{\sigma n} < \varepsilon \quad \text{for all } 0 \leq k \leq nt, \tag{2}$$

and, because $T_k/(k\sigma^2) \to_{\text{a.s.}} 1$, also

$$P\left(\left|\frac{T_k}{k\sigma^2} - 1\right| > \frac{\delta}{t} \text{ for some } k > \varepsilon\delta nt, \text{ and } \frac{T_{\varepsilon\delta nt}}{n\sigma^2} < \delta\right) < \varepsilon$$

and thus

$$P\left(\left|\frac{T_k}{n\sigma^2} - \frac{[k]}{n}\right| > \delta \text{ for some } 0 \leq k \leq nt\right) < \varepsilon. \tag{3}$$

https://doi.org/10.1017/S0269964806060360 Published online by Cambridge University Press

Together these give

$$P\left(\left|W_n(t)-\frac{ct}{\sigma}-\inf_{s\le t}\left(W_n(s)-\frac{cs}{\sigma}\right)-\frac{Q_{nt}}{\sigma\sqrt{n}}\right|>4\varepsilon\right)<2\varepsilon,$$

and so

$$\frac{Q_{nt}}{\sigma\sqrt{n}}\rightarrow_d R_{c/\sigma}(t).$$

■

Proof (of Theorem 2): It suffices (see Durrett [3, lemmas (6.8) and (6.9)] and the discussion following them) to show

$$\sup_{0\le t\le 1}|Q^*_{nt}/(\sigma\sqrt{n})-R_{c/\sigma}(t)|\rightarrow_p 0.$$

Since for $0\le t\le 1$ the function $Q^*_{nt}/(\sigma\sqrt{n})$ is uniformly continuous, we can pick δ small enough and n large enough so that

$$P(|Q^*_{nx}-Q^*_{ny}|/(\sigma\sqrt{n})>\varepsilon \text{ for some } 0\le x\le y\le 1 \text{ with } |x-y|<\delta)<\varepsilon$$

holds in addition to conditions (1)–(3) for all $0<t<1$. Together these give

$$P\left(\sup_{0\le t\le 1}\left|W_n(t)-\frac{ct}{\sigma}-\inf_{s\le t}\left(W_n(s)-\frac{cs}{\sigma}\right)-\frac{Q^*_{nt}}{\sigma\sqrt{n}}\right|>5\varepsilon\right)<2\varepsilon,$$

and so the theorem is proved. ■

References

1. Billingsley P. (1968). *Convergence of probability measures.* New York: Wiley.
2. Chen, H. & Yao, D.D. (2001). *Fundamentals of queueing networks. Performance, asymptotics, and optimization*. New York: Springer-Verlag.
3. Durrett, R. (2005). *Probability: Theory and examples*, 3rd ed. Belmont, CA: Duxbury Press.
4. Kingman, J.F.C. *The heavy traffic approximation in the theory of queues.* 1965 Proceedings Symposium on Congestion Theory (Chapel Hill, N.C., 1964), pp. 137–169. Chapel Hill: University of North Carolina Press.
5. Obłój, J. (2004). The Skorokhod embedding problem and its offspring. *Probability Surveys* 1: 321–390.
6. Robert, P. (2003). *Stochastic networks and queues.* Berlin: Springer-Verlag.
7. Rosenkrantz, W.A. (1980). On the accuracy of Kingman's heavy traffic approximation in the theory of queues. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwund Gebiete* 51(1): 115–121.
8. Skorokhod, A.V. (1965). *Studies in the theory of random processes.* Reading, MA: Addison-Wesley Publishing.
9. Whitt, W. (2002). *Stochastic-process limits. An introduction to stochastic-process limits and their application to queues.* New York: Springer-Verlag.

https://doi.org/10.1017/S0269964806060360 Published online by Cambridge University Press