A Heavy Traffic Diffusion Limit for Dummies

by Christos Zacharias, NYU Stern
http://people.stern.nyu.edu/cz464/
August 2013

1. About

This article demonstrates a comprehensive, step-by-step construction of a heavy traffic diffusion limit for stochastic processes. The analysis is based on the random evolution of the appointment backlog of a service system (for example doctor’s appointments).

2. The Appointment Queue

Consider a service system where requests for appointments are fully backlogged up to a finite buffer $K$. There is a fixed supply of $s$ available time slots per day. There is a random demand of $V_k$ appointment requests for day $k$, iid with finite mean $\lambda$ and variance $\sigma^2$, $k \geq 1$. Let $W_k$ be the appointment backlog (workload) at the end of period $k$, $k \geq 0$. Assuming that the appointment system starts empty, the successive workloads can be defined recursively by the Lindley recursion

$$W_k = \min \{ K, \max\{0, W_{k-1} + V_k - s\} \}, \text{ for } k \geq 1,$$

and $W_0 = 0$. \hfill (1)

The maximum term in (1) is induced by the fact that the workload is never allowed to become negative and at most $s$ customers are scheduled each period. The minimum term in (1) restricts the workload in $\{0,1,\ldots,K\}$; not more than $K$ customers are allowed to be backlogged.

3. The Diffusion (Brownian) Limit

As in Whitt (2002), we consider a sequence of appointment systems indexed by $n$, with buffer size $K_n$, capacity of $s_n$ appointment slots and a fixed input process $\{V_k : k \geq 1\}$. For model $n$ we have

$$W^n_k = \min \{ K_n, \max\{0, W^n_{k-1} + V_k - s_n\} \}, \text{ for } k \geq 1.$$

Let $\{S^n_k : k \geq 0\}$ be the random walk with step size the appointment requests, i.e. $S^n_k = \sum_{i=1}^k V_i$ for $k \geq 1$ and $S^n_0 = 0$. For model $n$, let $S^n_k = \sum_{i=1}^k (V_i - s_n)$ for $k \geq 1$ and $S^n_0 = 0$, random walk with steps $V - s_n$. Consider the scaled stochastic processes

$$S^n(t) := \frac{S^n_{\lfloor nt \rfloor} - \lambda nt}{\sqrt{n}},$$

$$S(t) := \frac{S_{\lfloor nt \rfloor}}{\sqrt{n}},$$

and

$$W^n(t) := \frac{W^n_{\lfloor nt \rfloor}}{\sqrt{n}},$$

where $S^n(t)$ is the scaled random walk, $S(t)$ is the scaled random walk, and $W^n(t)$ is the scaled workload process.
and note that
\[ S_n(t) = S^n_v(t) + \sqrt{n}(\lambda - s_n)t. \]

In order to establish a heavy-traffic diffusion limit for the workload process, we assume the following heavy traffic requirements:

**Assumption 1.** (a) \( \sqrt{n}(\lambda - s_n) \to \eta \in \mathbb{R} \) as \( n \to \infty \).

(b) \( K_n = \sqrt{n}K, \quad K \in \mathbb{R}^+ \).

(c) \( \sigma^2 < \infty \).

Under Assumption 1(a) and 1(c), and from the Functional Central Limit Theorem (FCLT) we get
\[ S^n_v(t) := S^n_v(\lfloor nt \rfloor) - \lambda nt \to \sigma B(t), \]
and \( S_n(t) = S^n_v(t) + \sqrt{n}(\lambda - s_n)t \to \sigma B(t) + \eta t, \)

where \( B(t) \) is a standard Brownian motion.

Now we define two more processes of interest. For system \( n \), let \( U^n_k \) be the cumulative number of customers lost (blocked) up to period \( k \) and \( L^n_k \) be the number of unutilized slots up to period \( k \). The workload at the end of period \( k \) satisfies
\[ W^n_k = \left( \sum_{i=1}^k V_i - U^n_k \right) - (ks_n - L^n_k) \]
\[ = \sum_{i=1}^k (V_i - s_n) + L^n_k - U^n_k \]
\[ = S^n_k + L^n_k - U^n_k. \]

Then, we define the associated scaled stochastic processes
\[ L^n(t) := \frac{L^n(\lfloor nt \rfloor)}{\sqrt{n}}, \]
and \( U^n(t) := \frac{U^n(\lfloor nt \rfloor)}{\sqrt{n}}. \)

Consequently we note that the triplet \((W^n, L^n, U^n)\) satisfies the following three conditions:

(a) \( W^n(t) = S^n(t) + L^n(t) - U^n(t) \in [0, K]. \)

(b) \( L^n(t) \) and \( U^n(t) \) are non-decreasing with \( L^n(0) = U^n(0) = 0. \)

(c) \( L^n(t) \) and \( U^n(t) \) increase only when \( W^n(t) = 0 \) and \( W^n(t) = K \) respectively.

The triplet \((W^n, L^n, U^n)\) is said to solve the Skorokhod problem for \( S_n \) on \([0, K]\). Such a triplet exists and it is unique, see Harrison (1985). One explicit solution to the Skorokhod problem is provided by Andersen and Mandjes (2008):
\[ W^n(t) = R(S^n)(t) := \sup_{s \leq t} \left( S^n(t) - S^n(s) \wedge \left( \inf_{u \in [s,t]} (K + S^n(t) - S^n(u)) \right) \right). \]
The mapping $R$ is often referred to as the “two-sided reflection map”. The Continuous Mapping Theorem (CMT) provides the desired diffusion limit for the workload

$$W_n(t) = R(S_n)(t) \Rightarrow R(\sigma B(t) + \eta t) =: W(t).$$

4. Concluding Remarks

The limiting workload (backlog) process $W(t)$ is a two-sided reflected Brownian motion (RBM) with drift $\eta$, infinitesimal variance $\sigma^2$, with reflective barriers 0 and $K$. As in Whitt (2004), $W(t) \Rightarrow W(\infty)$ as $t \to \infty$ with density

$$f_{W(\infty)}(x) = \frac{2\eta e^{\frac{2\eta x}{\sigma^2}}}{\sigma^2(e^{\frac{2\eta K}{\sigma^2}} - 1)}, \quad 0 \leq x \leq K$$

when $\eta \neq 0$, and the uniform density on $[0, K]$ when $\eta = 0$.

References


