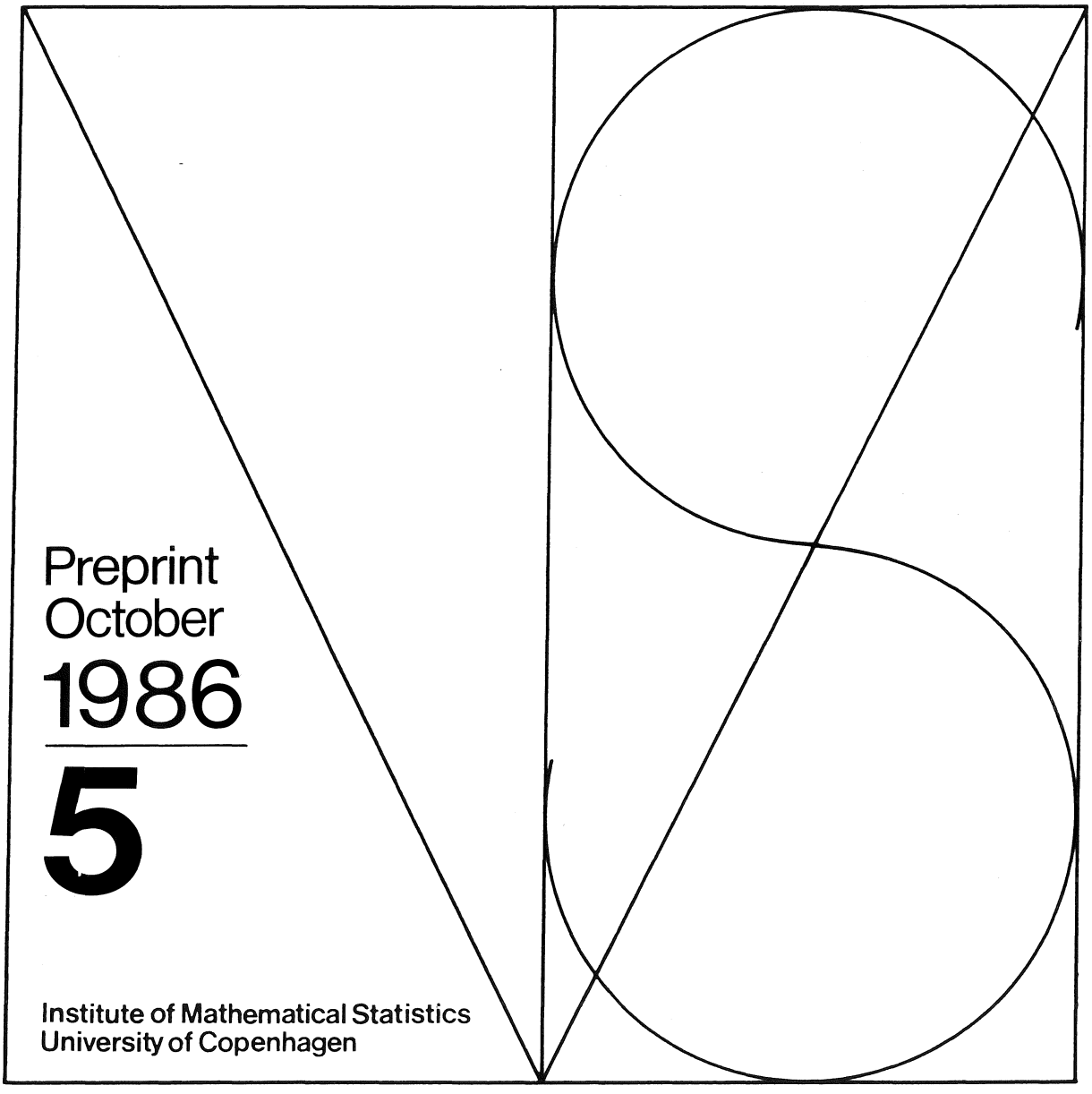
Søren Asmussen

The Heavy Traffic Limit of a Class of Markovian Queueing Models



Preprint
October
1986

5

Institute of Mathematical Statistics
University of Copenhagen

Søren Asmussen

THE HEAVY TRAFFIC LIMIT OF A CLASS OF MARKOVIAN QUEUEING MODELS

Preprint 1986 No. 5

INSTITUTE OF MATHEMATICAL STATISTICS
UNIVERSITY OF COPENHAGEN

October 1986

THE HEAVY TRAFFIC LIMIT OF A CLASS OF MARKOVIAN QUEUEING MODELS*

Søren ASMUSSEN

Institute of Mathematical Statistics, University of Copenhagen, Denmark

Reflected Brownian motion is obtained as the heavy traffic limit of the level component $\{Q_n\}$ of a class of bivariate Markov chains $\{(J_n,Q_n)\}$ incorporating those having a matrix-geometric stationary distribution. Approximations for both transient and ergodic behaviour are obtained as a corollary.

heavy traffic limit theorem * Markov chain * matrix-geometric stationary distributions * Markov-modulation

1. Introduction and statement of results

We are concerned with Markov chains $\{(J_n,Q_n)\}$ having transition matrices of the block form

$$P = \begin{pmatrix} K(0) & H(1) & H(2) & H(3) & \cdots \\ K(1) & G(0) & G(1) & G(2) & \\ K(2) & G(-1) & G(0) & G(1) & \\ K(3) & G(-2) & G(-1) & G(0) & \\ \vdots & & & & \ddots \end{pmatrix} \qquad (1)$$

where the dimensions are G(n) : p×p, K(n) : p×m, n>0, K(0) : m×m and H(n) : m×p. We write E_0 for the set of the m boundary states and E = {1,...,p}. Thus the state space is $E_0\times\{0\}\cup E\times\{1,2,...\}$ (E_0 and E may be disjoint).

In most examples, $\{Q_n\}$ is the process of main intrinsic interest (typically a queue length process) and J_n a supplementary variable needed for the Markov property. As a special case (G(n) = H(n) = 0, n>2, H(1) = G(1)) the setting incorporates Markov chains of the GI/M/1 type (Neuts [5]) having a matrix-geometric stationary distribution. Already this class of models is extremely versatile and has become a popular tool in applications, but also further examples like the M/G/1 type briefly discussed in [5] are included in (1). In fact, apart from the discrete nature of the variables, the only restriction inherent in (1) is the spatial homogeneity in levels ≠ 0. These facts indicate that results on the

*The present work was done while the author was visiting Stanford University. I gratefully acknowledge the hospitality of the Department of Operations Research and a grant from The Danish Natural Science Research Council. Thanks also to Mike Harrison for a conversation from which I learned a lot about diffusion approximations.

behaviour of processes like $\{(J_n,Q_n)\}$ are of very general nature.

We are here concerned with one of the classical areas of queueing theory, viz. the heavy traffic limit theorem. That is, we are looking for the limiting behaviour of $\{Q_n\}$ under conditions corresponding to $\rho\uparrow 1$ in the traditional queueing setting (also the case $\rho\downarrow 1$ has been considered, Iglehart and Whitt [3] or Whitt [6], but the stable case $\rho<1$ seems more interesting). We first need to introduce some notation. Define $G = \sum_{-\infty}^{\infty} G(k)$. If G is an irreducible transition matrix, as will typically be the case, an invariant probability (row) vector ν exists and we let

$$M = \sum_{k=-\infty}^{\infty} kG(k), \quad \mu = \nu Me, \quad \sigma^2 = \nu \sum_{k=-\infty}^{\infty} k^2 G(k)e - 3\mu^2 + 2\nu M(I+e\nu-G)^{-1}Me$$

where e is the (column) vector of ones. Further $B_\xi = \{B_\xi(t)\}_{t\geq 0}$ denotes Brownian motion with unit variance and drift ξ, and $B_\xi^{(R)}$ the zero-reflected version

$$B_\xi^{(R)}(t) = B_\xi(t) - \min_{0\leq s\leq t} B_\xi(s) \tag{2}$$

and we let $B'(t) = |\mu|Q_{[t\sigma^2/\mu^2]}/\sigma^2$ where $[\cdot]$ denotes integer part. Finally Φ denotes the standard normal distribution function and (J,Q) a pair of random variables having the limiting stationary distribution of (J_n,Q_n). In the heavy traffic situation we are thinking of the given transition matrix P as imbedded in a sequence $\{P^{(m)}\}$ with limit $P^{(0)}$, and limit theorems as $m\to\infty$ thus provide approximations for the given process. For notational convenience, we most often suppres indices $m\neq 0$ and thus e.g. B' really depends on m, $\mu\to 0$ means $\mu_m\to 0$ and so on.

Theorem 1 *Suppose that* $G(n)\to G^{(0)}(n)$, $K(n)\to K^{(0)}(n)$, $H(n)\to H^{(0)}(n)$ *for all* n *in such a way that the elements of* $\Sigma|n|^3G(n)$ *and* $\Sigma n^2H(n)$ *remain bounded, that* $\mu<0$ *and that the limit matrices* $P^{(0)}$, $G^{(0)}$ *are irreducible with* $\mu_0=0$, $\sigma_0^2>0$. *Then* B' *converges weakly to* $B_{-1}^{(R)}$ *in* $D[0,\infty)$. *In particular*

$$P(|\mu|Q_{[t\sigma^2/\mu^2]}/\sigma^2 > x,\ J_{[t\sigma^2/\mu^2]}=i)/\nu_i \to$$
$$1-\Phi(xt^{-1/2}+t^{1/2}) + e^{-2x}\Phi(-xt^{-1/2}+t^{1/2}) \tag{3}$$
$$P(|\mu|Q/\sigma^2 > x,\ J=i)/\nu_i \to 1-e^{-2x} \tag{4}$$

2. Proofs

We let $\{J_n^*\}$ be a Markov chain on E governed by G and $S_n^* = X_0^*+...+X_n^*$ the corresponding Markov-modulated random walk. That is, $\{(J_n^*,X_n^*)\}$ is a Markov chain on $E\times\{0,\pm1,\pm2,...\}$ which goes from state is to jt with probability $g_{ij}(t)$ (as initial condition we take $X_0^*=0$ throughout). Define further the corresponding Markov-modulated Lindley process by $Q_0^*=0$,

$$Q_n^* = (Q_{n-1}^*+X_n^*)^+ = S_n^* - \min_{0\le k\le n} S_k^* \qquad (5)$$

The following intuitive description provides the key for the proofs. The Markov-modulated Lindley process has again a transition matrix P^* of the form (1) corresponding to $G^*(k) = H^*(k) = G(k)$, $K^*(n) = I - \Sigma_{n+1}^{\infty} G(k)$. The transitions of $\{(J_n^*,X_n^*)\}$ and $\{(J_n,X_n)\}$ from levels s>0 to levels s+t>0 are governed by the same probabilities, viz. the elements of G(t), and also

$$P_{is}(Q_1=0) = 1 - \Sigma_{n=1-s}^{\infty} \Sigma_{j=1}^{P} g_{ij}(n) = \Sigma_{j=1}^{P} k_{ij}^*(s) = P_{is}(Q_1^*=0) .$$

However, when $\{Q_n^*\}$ hits zero, then $\{(J_n^*,X_n^*)\}$ just continues according to the Markov property, whereas at the hitting time τ (say) $\{J_n\}$ is reset to a value in the set of boundary states E_0. The exit from zero, i.e. the value of $(J_{\tau+1},Q_{\tau+1})$, is then chosen according to the atypical first row of P, and first when $Q_{\tau+s}>0$ (which may require more than s=1 steps), the G(k) take over to govern the transitions of $\{J_n\}$ again. The idea is now first to obtain reflected Brownian motion $B_{-1}^{(R)}$ as limit of $\{Q_n^*\}$ (this is the easy step), and next to show that $\{(J_n^*,Q_n^*)\}$ and $\{(J_n,Q_n)\}$ asymptotically behave the same way.

To carry out the details, one needs to extend a number of known estimates for random walks to the Markov-modulated case. It is frequently convenient to do this by studying $S_n^i = S_{\lambda(n;i)}$ where $\lambda(n;i)$ is the time of the n^{th} visit of $\{J_n^*\}$ to state i. If $J_0^*=i$, then $\{S_n^i\}$ is a usual random walk, and letting $\lambda(1)=\lambda(1;i)$, we have $E_i\lambda(i) = 1/\nu_i$ and:

Lemma 1 (a) *There exist* $\eta<1$ *and* N *such that* $P_i(\lambda(i)>n) < \eta^n$ *for all* m *and all* $n\ge N$;
(b) $ES_1^i = \nu_i\mu$, *and as* $m\to\infty$, $\mathrm{Var}\, S_1^i \to \nu_i\sigma_0^2$, $\limsup E|S_1^i|^3 < \infty$.

Proof (a) The condition $\mu_0=0$ ensures that $P^{(0)}$ is recurrent (Asmussen [1] X.4) and hence there exists N such that $P_j^{(0)}(\lambda(i)>N/2) < \delta_1 < 1$ for all $j \in E$. Since $G \to G(0)$ implies $P_j(\lambda(i)>N/2) \to P^{(0)}(\lambda(i)>N/2)$, we can choose $\delta_2<1$ such that $P_j(\lambda(i)>N/2) < \delta_2$ for all m. A geometric trial argument then shows that $\eta=\delta_2^{1/N}$ satisfies the requirements. In (b), the statements on the mean and variance follow by general results on regenerative processes ([1] V.3). Further the conditions of Th. 1 ensure that $E(|X_1^*|^3 |\ J_0^*=i,\ J_1^*=j)$ is bounded, say by c. Hence by Minkowski's inequality

$$E|S_1^i|^3 = E[E(|S^*_{\lambda(i)}|^3 \,|\lambda(i), J_0^*, \dots, J^*_{\lambda(i)})]$$

$$\leq E[\sum_{n=0}^{\lambda(i)-1} E(|X_n^*|^3 \,|\lambda(i), J_0, \dots, J_{\lambda(i)})^{1/3}]^3 \leq c\, E\lambda(i)^3$$

which remains bounded according to part (a). □

Now let $\{c\} = \{c^{(m)}\}$ be any sequence of real numbers with $c^{(m)} \to \infty$ and define

$$B^{(c)}(t) = (\sigma^2 c)^{-1/2}\{S_{[ct]} - [ct]\mu\}, \quad B^*(t) = |\mu|\, Q_{[t\sigma^2/\mu^2]}/\sigma^2.$$

Lemma 2 $B^{(c)} \to B_0$ *and* $B^* \to B_{-1}^{(R)}$.

Proof The first statement is essentially well known. In fact, for a fixed Markov-modulated random walk the central limit theorem (and the expression for σ^2 stated in the Introduction) is contained in Keilson and Wishart [4], whereas the functional form is given, e.g., in Billingsley [2]. To obtain the present triangular array version one may, e.g. use the standard random walk result to obtain $B^{(i)} \to B_0$ where

$$B^{(i)}(t) = (c\nu_i \mathrm{Var} S_1^i)^{-1/2}(S^i_{[ct\nu_i]} - [ct\nu_i]\mu/\nu_i)$$

$$= (c\nu_i \mathrm{Var} S_1^i)^{-1/2}(S^*_{\omega([ct\nu_i])} - [ct\nu_i]\mu/\nu_i).$$

Using $\omega([ct\nu_i] \cong ct$, one may then approximate $B^{(c)}$ by $B^{(i)}$, the necessary bounds being provided by Lemma 1. The details are fairly standard and omitted.

Letting $c = \sigma^2/\mu^2$ we get

$$\{|\mu| S^*_{[t\sigma^2/\mu^2]}/\sigma^2\}_{t\geq 0} \to \{B_{-1}(t)\}_{t\geq 0}.$$

Therefore $B^* \to B_{-1}^{(R)}$ follows immediately by (2), (5) and the continuous mapping theorem. □

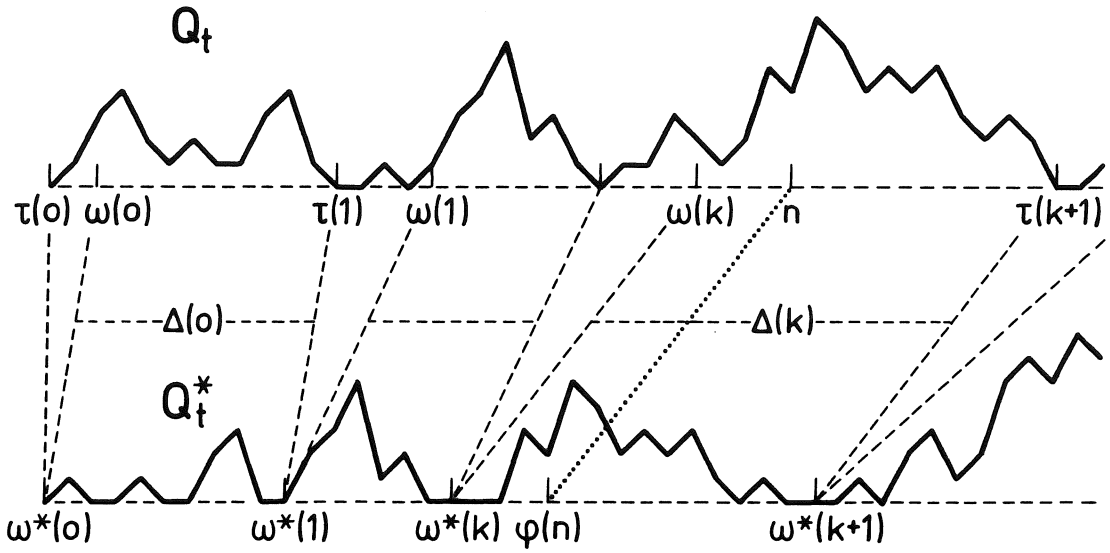


Figure 1

We next recursively define random times $\tau(k)$, $\omega(k)$, $\Delta(k)$, $\omega^*(k)$ by $\tau(0)=\omega^*(0)=0$,

$$\omega(k) = \inf\{n>\tau(k): Q_n>0,\ J_n = J^*_{\omega^*(k)}\},\quad \tau(k+1) = \inf\{n>\tau(k): Q_n = 0\},$$

$\Delta(k) = \tau(k+1)-\tau(k)$, $\omega^*(k+1) = \omega^*(k) = \Delta(k)$, cf. Fig. 1. We may assume that

$$J_{\omega(k)+r} = J^*_{\omega^*(k)+r},\quad Q_{\omega(k)+r+1} - Q_{\omega(k)+r} = X^*_{\omega^*(k)+r},\quad r < \Delta(k)-1.$$

It is basic to observe that $Q_n = 0$ if and only if n is a descending ladder point for $\{S_n\}$. Letting $n = \omega^*(k)$, it follows by induction that $Q_{\omega^*(k)} = 0$ for all k. Therefore, since $q\rightarrow(q+x)^+$ is a contraction,

$$|Q_{\omega(k)+r} - Q^*_{\omega^*(k)+r}| \le |Q_{\omega(k)} - Q^*_{\omega^*(k)}| = Q_{\omega(k)},\quad r < \Delta(k) \tag{6}$$

Now define

$$N_t = \sup\{k: \omega^*(k) < t\},\quad M_k = \sup\{Q_n: \tau(k) \le n \le \omega(k)\},$$

$$\varphi(n) = \begin{cases} \omega^*(k) & \tau(k) \le n \le \omega(k) \\ \omega^*(k) + n - w(k) & \omega(k) \le n \le \tau(k+1) \end{cases},$$

$$B''(t) = |\mu| \, Q^{*}_{[\varphi(t\sigma^2/\mu^2)]}/\sigma^2 = B^*(\varphi(t\sigma^2/\mu^2)\mu 2/\sigma 2).$$

Then

$$\sup_{0\le s\le t} |B''(s) - B'(s)| < |\mu| \max_{k\le N_t} M_k/\sigma^2 \tag{7}$$

(if $\varphi(s\sigma^2/\mu^2)$ is in $[\omega(k),\tau(k+1))$ this follows from (6), on $[\tau(k),\omega(k)]$ it is obvious from the definition of M_k). Hence the proof of $B' \to B^{(R)}_{-1}$ will be complete if we can show that subject to the heavy traffic limit $m\to\infty$ it holds for any fixed t that

$$EN_{t\sigma^2/\mu^2} = O(|\mu|^{-1}) \tag{8}$$

$$|\mu| \max_{k\le c/|\mu|} M_k \xrightarrow{P} 0 \tag{9}$$

$$\mu^2 \sup_{0\le s\le t\sigma^2/\mu^2} |\varphi(s)-s| \xrightarrow{P} 0 \tag{10}$$

Indeed, (10) and $B^* \to B^{(R)}_{-1}$ imply $B'' \to B^{(R)}_{-1}$, and (7)-(9) then yields $B' \to B^{(R)}_{-1}$.

Proof of (8). Let $\kappa_-(i)$ be the first descending ladder epoch of $\{S^i_n\}$ and N^i_t the number of descending ladder epochs before time t. Then Lemma 1(b) is well-known to imply $E\kappa_-(i) = O(|\mu|^{-1})$, $E\kappa_-(i)^2 = O(|\mu|^{-3})$. Hence by Lorden's inequality for the renewal function ([1] VI.4),

$$EN_{t\sigma^2/\mu^2} \le \sum_{i\in E} EN^i_{t\sigma^2/\mu^2}$$

$$\le \sum_{i\in E} \left\{ \frac{t\sigma^2/\mu^2}{E\kappa_-(i)} + \frac{E\kappa_-(i)^2}{(E\kappa_-(i))^2} \right\} = O(|\mu|-1).$$

Proof of (9), (10). Let $\delta(k) = \omega(k) - \tau(k)$. We first note that since $\delta(k)$ is the exit time from a finite set of states, it follows exactly as in the proof of Lemma 1(a) that the tail of $\delta(k)$ is geometrically small uniformly in m. Hence

$$\mu^2 \sup_{0\le s\le t\sigma^2/\mu^2} |\varphi(s)-s| = \mu^2 E|\varphi(t\sigma^2/\mu^2)- t\sigma^2/\mu^2| \le$$

$$\mu^2 E \sum_{k=1}^{N_{t\sigma^2/\mu^2}} \delta(k) = O(\mu^2 E N_{t\sigma^2/\mu^2}) = O(|\mu|),$$

proving (10). For (9), it is enough to show that $EM_k \le c$ for all k. But

$$EM_0 \le E \sum_{n=0}^{\delta(0)} Q_n \le \sum_{n=0}^{\infty} [P(\delta(0)>n) EQ_n^2]^{1/2}$$

and we only have to obtain some rough bound on EQ_n^2. For example, the conditions of Th.1 imply $E(Q_{n+1}-Q_n)^k = O(1)$, k=1,2, which yields $EQ_n = O(n)$, $EQ_n = O(n^2)$. Hence $EM_0 \le c$ and by the Markov property, $EM_k \le c$ for all k. □

It only remains to prove (3), (4). It is well-known that the r.h.s. of (3) is the limit $P(B_{-1}^{(R)}(t) > x)$ of $P(|\mu| Q_{[t\sigma^2/\mu^2]}/\sigma^2 > x)$ and all that needs to be shown is asymptotic independence of $J_{[t\sigma^2/\mu^2]}$ which can be obtained along the lines of a standard lemma due to Stam ([1] XII.5). Also the proof of (4) follows the one-dimensional case (e.g. [1] VII.6) closely, a main step being an application of Kolmogorov's inequality to $\{S_n^i\}$. We omit the details.

Note that for the matrix-geometric case the l.h.s. of (4) can be computed numerically. Thus for this case, the time-dependent version (3) is the more interesting from the point of practical approximations. Nevertheless, (4) reflects a general tendency of the model in heavy traffic.

References

[1] S. Asmussen, *Applied Probability and Queues*, Wiley, Chichester, 1986 (in print).

[2] P. Billingsley, *Convergence of Probability Measures* , Wiley, New York, 1968.

[3] D.L. Iglehart and W. Whitt, Multiple channels in heavy traffic I-II, *Adv. Appl. Probab.* **15**, 150-177, 355-364 (1970).

[4] J. Keilson and D.M.G. Wishart, A central limit theorem for processes defined on a finite Markov chain, *Proc. Camb. Philos. Soc.* **60**, 547-567 (1964), *ibid.* **63**, 187-193 (1967).

[5] M.F. Neuts, *Matrix-geometric Solutions in Stochastic Models* , Johns Hopkins University Press, Baltimore, 1981.

[6] W. Whitt, Heavy traffic limit theorems for queues: a survey, *Lecture Notes in Economics and Mathematical Systems* **98**. Springer, Berlin, 1974, 307-350.

PREPRINTS 1985

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, 2100 COPENHAGEN Ø, DENMARK.

No. 1 Cohn, Harry: Almost Sure Convergence for Stochastically Monotone Temporally Homogeneous Markov Processes and Applications.

No. 2 Rootzén, Holger: Maxima and Exceedances of Stationary Markov Chains.

No. 3 Asmussen, Søren und Johansen, Helle: Über eine Stetigkeitsfrage betreffend das Bedienungssystem GI/GI/s.

No. 4 Asmussen, Søren and Thorisson, Hermann: A Markov Chain Approach to Periodic Queues.

No. 5 Hald, Anders: Galileo as Statistician.

No. 6 Johansen, Søren: The Mathematical Structure of Error Correction Models.

No. 7 Johansen, Søren and Johnstone, Iain: Some Uses of Spherical Geometry in Simultaneous Inference and Data Analysis.

PREPRINTS 1986

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, 2100 COPENHAGEN Ø, DENMARK.

No. 1 Jespersen, N.C.B.: On the Structure of Simple Transformation Models.

No. 2 Dalgaard, Peter and Johansen, Søren: The Asymptotic Properties of the Cornish-Bowden-Eisenthal Median Estimator.

No. 3 Jespersen, N.C.B.: Dichotomizing a Continuous Covariate in the Cox Regression Model.

No. 4 Asmussen, Søren: On Ruin Problems and Queues of Markov-Modulated M/G/1 Type.

No. 5 Asmussen, Søren: The Heavy Traffic Limit of a Class of Markovian Queueing Models.