Conditioned Asymptotics for Tail Probabilities in Large Multiplexers *

N.G. Duffield

AT&T Laboratories,
Room 2C-323, 600 Mountain Avenue,
Murray Hill, NJ 07974, USA

E-mail: duffield@research.att.com

Abstract

Consider a buffer whose input is a superposition of L independent identical sources, and which is served at rate sL. Recent work has shown that, under very general circumstances, the stationary tail probabilities for the queue of unfinished work Q in the buffer have the asymptotics $\mathbf{P}[Q > Lb] \approx e^{-LI(b)}$ for large L. Here the *shape function*, I, is obtained from a variational expression involving the transient log cumulant generating function of the arrival process.

In this paper, we extend this analysis to cover time-dependent asymptotics for Markov arrival processes subject to conditioning at some instant. In applications we envisage that such conditioning would arise due to knowledge of the queue at a coarse-grained level, for example of the number of current active sources. We show how such partial knowledge can be used to predict future tail probabilities by use of a time dependent, conditioned shape function. We develop some heuristics to describe the time-dependent shape function in terms of a reduced set of quantities associated with the underlying arrivals process and show how to calculate them for renewal arrivals and a class of ON-OFF arrivals. This bypasses the full variational calculation of the shape function for such models.

1 Introduction

In this paper we explore the time evolution of estimates of loss ratios in multiplexers of many superposed Markovian sources. Specifically we are interested in their response to conditionings applied at an instant of time. In practice such conditioning could arise through determination of a reduced set of state variables, for example the number of currently active sources.

The phenomenology of the conditioned evolution of loss ratios can be anticipated from the following observation. It has been established in a large class of queues serving increasingly many sources that at a given time, the most likely way for the queue length Q to exceed a given level b is for arrivals to build up over the previous interval whose duration $\tau(b)$ is asymptotically proportional to b for large b. (See [2, 12] for some general results of this nature for sources obeying mixing conditions; see [15, 16] specifically for large superpositions). Now suppose that the arrival process takes a time t_r to relax back to stationarity from its

*To be published in Performance Evaluation

conditioned state. Then the probability that the queue exceeds a level b at time $u > 0$ will be the same as for the stationary queue provided that b is sufficiently small that $\tau(b) < u - t_r$ holds. For then, only arrivals since the return to stationarity effect the probabilities of interest. However, for larger b, the $P[Q > b]$ will be determined by the arrivals prior to relaxation also. Since $\tau(b)$ is asymptotically linear in b one can summarize these observations as saying that the conditioning at time 0 can be accounted for as a change to the tail distribution $P[Q > b]$ which passes out to infinity as time increases, leaving the stationary distribution in its wake. In this paper we shall substantiate this description for Markovian arrivals processes which satisfy appropriate mixing conditions, and provide some heuristics to describe the evolution of the tail distributions. We illustrate the heuristics with the subclasses of renewal and ON-OFF arrivals.

First, we introduce the framework within the asymptotics of the tail distributions of the buffer occupations. Consider an infinitely buffered queue which processes stationary mixing arrivals. Using the theory of large deviations it can be shown that the tail asymptotics satisfy

$$\lim_{b\to\infty} b^{-1} \log \mathbf{P}[Q > b] = -\delta, \tag{1.1}$$

where the exponential decay rate δ is calculated in terms of the arrivals process and the service rate of the queue. This result has been proved in various degrees of generality in [6, 18, 23].

The consequent *effective bandwidth approximation*

$$\mathbf{P}[Q > b] \approx e^{-\delta b} \tag{1.2}$$

has been proposed as an estimate of the loss ratio in a buffer of size b (See [26, 34] and references therein). However, there is recent work, both numerical [7] and theoretical [5, 9, 31], which shows that this estimate can be inaccurate when the arrivals are composed of superposition of L streams each with a high degree of autocorrelation, increasingly so as L becomes large. (See also [33] for an earlier large deviation treatment of the large L asymptotic for Markov fluid sources). This is illustrated in Figure 1 for a set of simulated loss curves for various L. These are generated at constant load in that the service rate is proportional to L. Each source was a discrete time Markov Modulated process, in which the modulating process was a 2-state Markov chain which generated periodic arrivals of fixed size in the one state and no arrivals in the other. This was proposed as a model of packetized voice traffic in [11]. The simulation curves are taken from [8].

Large deviation theory provides the explanation of this behavior. We present briefly from [5]. A basic result is that for the L-fold superposition, served at a rate proportional to L so that the offered load is independent of L:

$$\lim_{L\to\infty} L^{-1} \log \mathbf{P}[Q > Lb] = -I(b) \tag{1.3}$$

for some *shape function* I depending on the service rate and the arrival process. I is determined as follows. Let time t take values in $T = \mathbf{R}_+$ or $\mathbf{Z}_+$. Consider an L-fold superposition (possibly heterogeneous) of arrival streams. Let A_t^L denote the work arriving from this superposition during the interval $[-t, 0)$. The service rate is sL for some fixed s. Define the excess workload process by $W_t^L = A_t^L - Lst$, i.e., the difference between the work arriving in $[-t, 0)$ and the amount of work which would be processed during a busy period of length t. The transient cumulant generating function (CGF, or log moment generating function) of the workload process is defined by

$$\lambda_t^L(\theta) = (Lt)^{-1} \log \mathbf{E}[e^{\theta W_t^L}]. \tag{1.4}$$

We assume

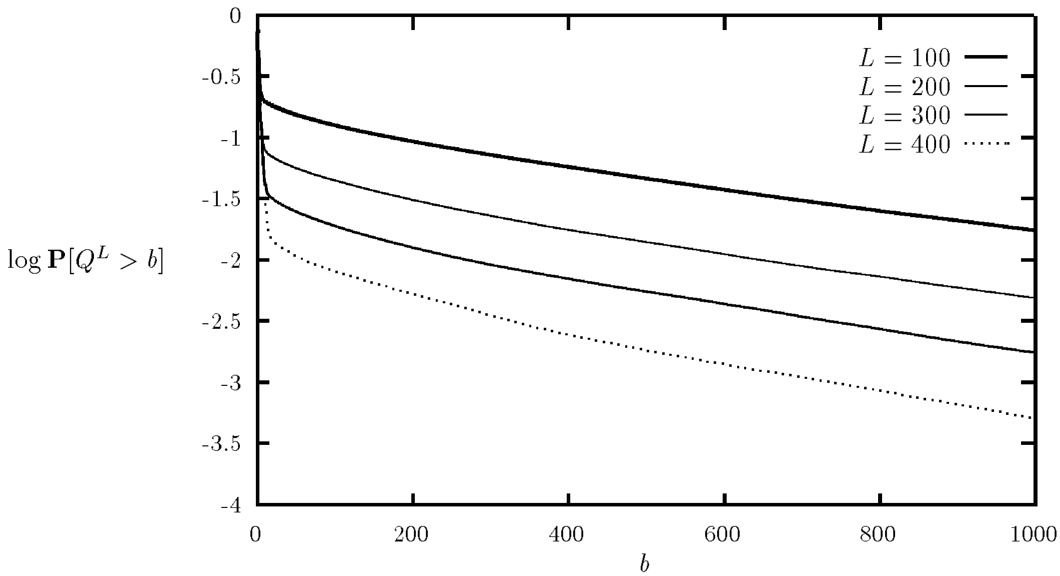

Figure 1: Simulated loss curves for increasing L.

Hypothesis 1 *(i)* Limiting CGF: *The limits* $\lambda_t(\theta) = \lim_{L\to\infty} \lambda_t^L(\theta)$ *and* $\lambda(\theta) = \lim_{t\to\infty} \lambda_t(\theta)$ *exist as extended real numbers for all* $\theta \in \mathbf{R}$*, the first limit uniform for* t *sufficiently large.*

(ii) Stability: *There exists* $\theta > 0$ *such that* $\lambda_t(\theta) < 0$ *for* t *sufficiently large.*

(iii) Smoothness: λ_t *and* λ *are essentially smooth (see [29]).*

Let Q denote the amount of unprocessed work at $t = 0$. Pathwise we have (see [4])

$$Q = \sup_{t\geq 0} W_t^L. \tag{1.5}$$

Denote by f^* the Legendre Transform of a real function f, i.e. $f^*(x) = \sup_\theta(x\theta - f(\theta))$.

Proposition 1 *Under Hypothesis 1 (and with the addition of a regularity condition 1(iv) of [5] when* $T = \mathbf{R}_+$*) then eq. (1.3) holds with*

$$I(b) = \inf_{t>0} (t\lambda_t)^*(b). \tag{1.6}$$

Hypothesis 2 *(i) The equation* $\lambda(\delta) = 0$ *has a positive solution.*

(ii) $\nu := \lim_{t\to\infty} \nu_t$ *exists and is finite, where* $\nu_t = -t\lambda_t(\delta)$*, and* δ *is the positive solution of (i).*

Proposition 2 *Under Hypothesis 2 (and with the addition of a regularity condition 3(ii) of [5] for* $T = \mathbf{Z}_+$*) then* δ *can be expressed as*

$$\delta = \inf_{x>0} x\lambda^*(1/x), \quad \text{and} \tag{1.7}$$

$$\lim_{b\to\infty} (I(b) - \delta b) = \nu. \tag{1.8}$$

It is worth remarking that neither stationarity of the increments of the W_t^L nor homogeneity of the arrivals are required for Propositions 1 and 2.

The work of this paper rests upon the observation (which we shall substantiate in classes of models) that whereas δ is insensitive to conditioning of the arrivals process A at some point in time, ν is not. One can understand this latter property at an intuitive level by remarking that the mixing conditions sufficient for the existence of λ (see e.g. [13]) should mean that the effects of conditioning decay with t and are time-averaged out in (1.4) as $t \to \infty$. However, we see from Hypothesis 2(ii) that ν is *not* a time-averaged quantity: so it is plausible that the effects of initial conditions persist at this finer level of the shape function.

Now the motivation for the shape function was that is should provide a correction to the effective bandwidth approximation for queue-tail asymptotics. Consider a buffer serving a large number of stationary sources; take, for example, MMPP sources. Suppose at $t = 0$ we observe a deviation in the empirical distribution of the modulating processes away from its mean. Conditioned upon this event, the large buffer asymptotic of the shape function $I(b) \approx \delta b + \nu$ as $b \to \infty$ will be shifted from its average value, due to the shift in ν. The ramifications of this are that predictions of the cell-loss ratios based on the effective bandwidth approximation (1.2), or its refinements using (1.3) and (1.8), can be inaccurate in the presence an observed large deviation in the aggregate arrival process.

What can be said about the evolution in time of such phenomena? We shall show that for finite b a good approximation to $I(b)$ can be got by taking $I(b) \approx \delta b + \nu_{\tau(b)}$ where $\tau(b)$ is the time-scale over which arrivals must build up in order for the queue to exceed the level b. This enables us to substantiate the behavior announced at the start of this section: at times $u > 0$ and buffer sizes b such that $\tau(b) < u - t_r$ where t_r is the time required to relax to stationarity after conditioning, then $\nu_{\tau(b)}$ is the same as it would be for the unconditioned process. The consequent approximation for loss ratios is

$$\mathbf{P}[Q^L > Lb] \approx e^{-LI(b)} \approx e^{-L\delta b} e^{-L\nu_{\tau(b)}}. \tag{1.9}$$

We envisage that the consequent of loss ratios under conditioning could be used as the basis for admission control under two different scenarios. The first is in circumstances where information on the modulating state (for example, the number of sources currently active) is more readily available that detailed information about the state of the queue. This is reasonable if modulating state information resides is more accessible that detailed state describing the queue and short term arrival patterns. For example, the former could reside high in the protocol stack, the latter being buried within the implementation of the queue. However, with an appropriate model of the detailed behavior of the sources, admission control would then be based on determining, through knowledge of the current modulating state, the bandwidth required in order to guarantee sufficiently small loss ratios in the future. These probabilities would be determined by the conditioned, time-dependent shape function determined from a traffic model. An example of admission and other controls based on such modulating state information is presented in [19].

In the second scenario we assume instead that we have access to detailed state of the arrivals process. Some recent work [10, 17] has demonstrated the possibility of determining the CGF λ and the consequent slope δ for real-traffic. In principle the offset ν can be measured in the same manner and used to determine loss probabilities via (1.3) and (1.8) in a model independent manner without identifying a specific conditioning of the arrivals.

Our work in this paper is divided into three parts. Firstly, we demonstrate the generality of the behavior described above by establishing it in queues with a broad class of arrivals processes, namely Markov Additive Processes (MAP's). The tasks involved in this are as follows. In section 2 we establish the sensitivity of ν to conditioning and show how to combine conditionings on different MAP sources in order to approximate the loss curve for the queue fed by their superposition. In section 3 we demonstrate the time-dependent behavior of the conditioned shape function, and in section 4 we propose two heuristics to describe this. The heuristics have the advantage that they are characterized by a reduced set of quantities rather than the full conditioned shape function. Having established the behavior described above to the level of generality of MAP's, the second part of our work is to provide formula for its application in two simpler subclasses of models, namely renewal arrivals processes, and alternating renewal arrivals processes, or ON-OFF models. This is done in section 5. As a by-product we are able to provide formulas for the *stationary* offset ν for such models; this is a matter of independent interest which seems not to have been worked out yet. With these results to hand, the third part of our work, done in section 6, is to demonstrate the accuracy of the heuristics by comparing them with the full time dependent conditioned shape function with the heuristic approximations in an example.

2 Conditioning in a queue with Markovian arrivals.

When a queue is fed by an L-fold superposition of independent sources served at rate sL then the transient CGF for W_t^L decomposes as a sum

$$\lambda_t^L(\theta) = L^{-1} \sum_{i=1}^{L} \lambda_{i,t}(\theta) \tag{2.1}$$

where $\lambda_{i,t}(\theta) = t^{-1} \log \mathbf{E}[e^{\theta W_{i,t}}]$ is the transient CGF of source i assigned a service rate s. Our first step in analysis of the conditioned asymptotics for large superpositions will be to analyze the conditioned CGF for a single Markovian source and to find the appropriate form for ν.

Consider a single source for which the backward excess workload process W is the additive component of an MAP (X, W). That is to say, let $X = (X_t)_{t\in T}$ be an irreducible aperiodic Markov process on a state space E with σ-field $\mathcal{E}$, adjoined to which is an additive component $W = (W_t)_{t\in T}$ with $W_0 = 0$ such that (X, W) is a Markov process on the state space $E \times \mathbf{R}$. Here X plays the role of a modulating process, while W_t gives the cumulative additions to the excess workload as determined by $(X_s : 0 \le s \le t)$. As a special case, X may itself be the increment process of the workload so that $W_t = \int_0^t ds\, X_s$. For each $t' \ge 0 \in T$ the joint distribution of $X_{t+t'}$ and the increments $W_{t+t'} - W_t$, conditioned on $(X_{t''}, W_{t''})_{0\le t''\le t}$ depend only on X_t. This dependence can be expressed through the kernel

$$P_{t'}(x, G \times B) := \mathbf{P}[X_{t+t'} \in G, W_{t+t'} - W_t \in B \mid X_t = x] \quad ,$$

for $G \in \mathcal{E}$ and B a Borel set of $\mathbf{R}$. We assume the existence of a stationary distribution q for X.

A recurrence condition on the kernel P is required for what follows (see [25], or the summary in Hypothesis 5 in [5]). We assume it to be satisfied. Define the transformed kernel

$$\hat{P}_t(\theta) = \hat{P}_t(x, dy; \theta) = \int P_t(x, dy \times dw) e^{\theta w}. \tag{2.2}$$

Let $e^{\lambda(\theta)}$ denote the maximal eigenvalue of $\hat{P}_1(\theta)$, $r(\theta)$, $\ell(\theta)$ the corresponding (right) eigenfunction and (left) eigenmeasure respectively. Define the conditioned CGF

$$\lambda_t(\theta, x) := t^{-1} \log \mathbf{E}[e^{\theta W_t} \mid X_0 = x] = t^{-1} \log \hat{P}_t(x, E; \theta), \tag{2.3}$$

and its unconditioned version

$$\lambda_t(\theta) = t^{-1} \log \mathbf{E}[e^{\theta W_t}] = t^{-1} \log \int q(dx) \hat{P}_t(x, E; \theta). \tag{2.4}$$

Our first result establishes the convergence of the single source CGF under conditioning and gives the appropriate conditioned form of ν.

Proposition 3 *(i) $\lambda(\cdot)$ is essentially smooth and essentially strictly convex on some domain $\mathcal{D}$.*

(ii) For $\theta \in \mathcal{D}$, $\ell(\theta)$ is absolutely continuous with respect to q; $r(\theta)$ and $d\ell(\theta)/dq$ are uniformly positive and bounded.

(iii) For all $\theta \in \mathcal{D}$

$$e^{t\lambda_t(\theta,x)} = e^{-\nu(\theta,x)} e^{t\lambda(\theta)} \left(1 + \mathbf{O}(\varepsilon^t(\theta))\right) \tag{2.5}$$

for some $\varepsilon(\theta) \in (0,1)$ where λ is the unconditioned CGF and

$$e^{-\nu(\theta,x)} = \frac{r(\theta,x)\ell(\theta,E)}{\int \ell(\theta,dy) r(\theta,y)} > 0. \tag{2.6}$$

(iv) For all $\theta \in \mathcal{D}$ and $x \in E$ $\lim_{t\to\infty} \lambda_t(\theta,x) = \lambda(\theta)$. Moreover, the convergence is uniform in x.

(v) If a strictly positive solution δ to the equation $\lambda(\delta) = 0$ exists

$$-\lim_{t\to\infty} t\lambda_t(\delta, x) = \nu(x) := \nu(\delta, x). \tag{2.7}$$

Proof: (i,ii,iii) follow from Lemmas 3.1 and 3.4 of [25]. Then (iv) and (v) become trivial corollaries. The uniformity of convergence in (iv) follows from the form (2.5) and the boundedness of $r(\theta)$ and $d\ell(\theta)/dq$. ■

We now treat heterogeneous conditioning within otherwise homogeneous superpositions of sources. (The extension to superpositions of multiple classes of arrival processes, each with heterogeneous conditioning is straightforward). Consider at L-fold superposition of MAP sources $(X_{i,t}, W_{i,t})_{i=1,\dots L, t\in\mathbf{R}}$ with identical transition kernels. Set $X_0^L = \{X_{i,0} : i = 1, 2, \dots L\}$ and let μ_L denote the empirical distribution of the $X_{i,0}$, i.e.

$$\mu_L = \frac{1}{L} \sum_i \delta_{X_{i,0}}, \tag{2.8}$$

where δ_x is measure with a single atom at $x \in E$. Then the conditioned CGF for finite L is

$$\lambda_t^L(\theta, X_0^L) = \int \mu_L(dx) \lambda_t(\theta, x). \tag{2.9}$$

Proposition 4 *Assume μ_L converges weakly to some measure μ as $L \to \infty$. Then*

(i) $\lim_{L\to\infty} \lambda_t^L(\theta, X_0^L) = \lambda_t(\theta, \mu) := \int \mu(dx)\lambda_t(\theta,x)$ uniformly for $t \geq 1$ and $\theta \in \mathbf{R}$.

(ii) $\lim_{t\to\infty}\lambda_t(\theta,\mu)=\lambda(\theta)$ *where* λ *is the* CGF *of the unconditioned process.*

(iii) Suppose a positive solution δ *to equation* $\lambda(\delta)=0$ *exists. Then*

$$\nu(\mu) := \int \mu(dx)\nu(x) = -\lim_{t\to\infty} t\lambda_t(\delta,\mu). \tag{2.10}$$

Proof: (i) The convergence is uniform since by Prop. 3(ii,iii), $|\lambda_t(\theta,x)-\lambda(\theta)|$ is uniformly bounded for $t\geq 1$. (ii) This follows, from Prop. 3(iv), using the uniformity in x of convergence of $\lambda_t(\theta,x)$ to $\lambda(\theta)$ as $t\to\infty$. (iii) By Prop. 3(ii,iii) $\nu(x)$ is a bounded on E, and $|\nu(x)-t\lambda_t(\delta,x)|$ converges to 0 as $t\to\infty$ and is uniformly bounded for all $x\in E$ and $t\geq 1$. ■

In conjunction with Proposition 3, Proposition 4 tells us when a positive solution δ to the equation $\lambda(\delta)=0$ exists then Hypotheses 1 and 2 are satisfied. The main theorem of the section then follows as a corollary.

Theorem 1 *For the conditioned queueing system described above then the asymptotic (1.3) holds with a shape function* I *for which*

$$\lim_{b\to\infty}(I(b)-\delta b)=\nu(\mu). \tag{2.11}$$

3 Evolution after conditioning

How does the shift in ν from its stationary value evolve after conditioning? We will answer this question at the level of single source shape functions, or equivalently, for homogeneous conditionings. The results extend to heterogeneous conditions by use of Proposition 4; one need only take the expectation of a given conditioned shape function with the distribution ν if initial conditions. We will recapitulate how to do this at the end of section 4. Our ability to predict the future shape function rests on the assumption of *a priori* stationarity of X and of the increments of W. In this case the Markov property for the backward process (X,W) is equivalent to that of the corresponding re-reversed (i.e. forward) process $(\tilde{X},\tilde{W})$. We examine the effects of condition at some point $t=-u$ in the past, and derive the shape function for the queue at $t=0$. In the main result of this section, Theorem 3, we show that for any fixed buffer level b, the offset ν of the shape function converges to its stationary value as we move forward in time. However, the conditioned value of ν persists at increasingly large buffer levels. We conclude this section by obtaining a relation between the conditioned offsets for a given MAP and its time-reversed process.

Let $\tilde{R}$ be the transition kernel for $\tilde{X}$. Then the moment generating function of the work arriving in the t units of time previous to some $u>0$, conditioned on the modulating process X taking the value x at time 0 is, by stationarity of (X,W), equal to

$$\mathbf{E}[e^{\theta W_t}\mid X_u=x]=\begin{cases}\int \tilde{R}_{u-t}(x,dy)\mathbf{E}[e^{\theta W_t}\mid X_t=y] & \text{if } t\leq u\\ \mathbf{E}[e^{\theta(W_t-W_u)}\mid X_u=x]\mathbf{E}[e^{\theta W_u}\mid X_u=x] & \text{if } t>u\end{cases}. \tag{3.1}$$

In the last expression we have used the Markov property of the MAP. Using $\lambda_t,\tilde{\lambda}_t$ to denote the transient CGF's for the backward and forward workloads respectively, and similarly defining ν and $\tilde{\nu}$, we get from (3.1) that

$$t\Lambda_{x,u,t}(\theta):=\log\mathbf{E}[e^{\theta W_t}\mid X_u=x]=\begin{cases}\log\int\tilde{R}_{u-t}(x,dy)e^{t\tilde{\lambda}_t(\theta,y)} & \text{if } t\leq u\\ (t-u)\lambda_{t-u}(\theta,x)+u\tilde{\lambda}_u(\theta,x) & \text{if } t>u\end{cases}. \tag{3.2}$$

Let $I_{x,u}$ be the shape function derived from the $\Lambda_{x,u,t}$, i.e.,

$$I_{x,u}(b) = \inf_{t>0}(t\Lambda_{x,u,t})^*(b), \tag{3.3}$$

and I that from the λ_t. The next proposition shows that in the limit $t \to \infty$, $\Lambda_{x,u,t}$ is completely insensitive to the time u since the conditioning took place.

Proposition 5 *Let $U(\cdot)$ be any positive function on $\mathbf{R}_+$. Then* $\lim_{t\to\infty} \Lambda_{x,U(t),t}(\theta) = \lambda(\theta) = \tilde{\lambda}(\theta)$.

Proof: The equality of $\tilde{\lambda}$ and λ follows from the (trivial) equality of $\tilde{\lambda}_t$ and λ_t. To obtain the first equality, substitute (2.5) into (3.2) and use the fact (from Prop. 3(ii,iii)) that $|\nu(x)|$ is bounded to show that $\Lambda_{x,u(t),t}(\theta, x) = \lambda(\theta) + \mathbf{O}(t^{-1})$ for $t \geq 1$. ■

The last result suggests that we will be able to assign the same asymptotic slope δ (the positive root of $\lambda(\delta) = 0$) to any of the shape functions I_u. However, while the slope will be independent of u, the offset of the asymptote from the origin is not. The next proposition helps determine the location $\hat{\tau}(u)$ of the infimum in the variational expression (1.6) for $I_{x,u}$, and to express the offset $I_u(b) - \delta b$ in terms of it.

Proposition 6 *Let $\tau : \mathbf{R}_+ \to \mathbf{R}_+$ be strictly increasing to infinity. Then there exist functions $\beta, \hat{\tau} : \mathbf{R}_+ \to \mathbf{R}_+$, with $\beta(u)$ eventually increasing to infinity, such that as $u \to \infty$*

$$\beta(u)/\tau(u) \to \lambda'(\delta)\,, \qquad \tau(u)/\hat{\tau}(u) \to 1 \tag{3.4}$$

and for which the following bounds hold for all $u > 0$

$$-\hat{\tau}(u)\Lambda_{x,u,\hat{\tau}(u)}(\delta) \leq I_{x,u}(\beta(u)) - \delta\beta(u) \leq -\tau(u)\Lambda_{x,u,\tau(u)}(\delta) \tag{3.5}$$

Proof: Define $\beta(u) = \tau(u)\Lambda'_{x,u,\tau(u)}(\delta)$. By Lemma 5 and Lemma IV.6.3 of [20],

$$\lim_{u\to\infty} \Lambda'_{x,u,\tau(u)}(\delta) = \lambda'(\delta) \tag{3.6}$$

and hence β is eventually increasing, and the first limit in (3.4) holds.

Upper Bound: $I_{x,u}(\beta(u)) - \delta\beta(u) \leq (\tau(u)\Lambda_{x,u,\tau(u)})^*(\beta(u)) - \delta\beta(u) = -\tau(u)\Lambda_{x,u,\tau(u)}(\delta)$.
Lower Bound: Let the infimum (1.6) for $I_{x,u}(\beta(u))$ be attained at $\hat{\tau}(u)$. Then

$$I_{x,u}(\beta(u)) - \delta\beta(u) = (\hat{\tau}(u)\Lambda_{x,u,\hat{\tau}(u)})^*(\beta(u)) - \delta\beta(u) \geq -\hat{\tau}(u)\Lambda_{x,u,\hat{\tau}(u)}(\delta) \tag{3.7}$$

The results then follows if we can establish that $\tau(u)/\hat{\tau}(u) \to 1$ as $u \to \infty$. This follows from the Theorem 4 in [15] provided that for some $t_0 > 0$

$$\liminf_{b\to 0} \inf_{0<t<t_0} t\lambda^*_{bt}(1/t) \geq \delta. \tag{3.8}$$

Now from (2.5) it follows that for some $K > 0$, $e^{t\lambda_t(\theta)} \leq e^K e^{t\lambda(\theta)}$ for all θ and for all $t > 0$. Consequently

$$\lambda_t(\theta) \leq \lambda(\theta) + K/t \quad \text{and hence} \quad t\lambda^*_{bt}(1/t) \geq t\lambda^*(1/t) - K/b. \tag{3.9}$$

(3.8) then follows for any $t_0 > 0$ by use of (1.7). ■

Proposition 6 is the technical result which will enable us to substantiate (1.9) in an appropriate sense. It indicates that when $\beta(u)$ is equal to some large b, then $\tau(b)$ of (1.9) is roughly $b\lambda'(\delta)$ and so the quantity denoted by $\nu_{\tau(b)}$ in (1.9) which approximates $I_{x,u}(b) - \delta b$ is roughly $-b\lambda'(\delta)\Lambda_{x,u,b\lambda'(\delta)}(b\lambda'(\delta))$. We now formalize this in the main theorems of this section. First we establish the extreme values of the offset as $b \to \infty$ and $u \to \infty$, according to the order in which we take limits.

Theorem 2 *Assume that δ exists as the positive solution of $\lambda(\delta) = 0$ and the hypotheses of Proposition 2 are satisfied. Then*

(i) $\lim_{b\to\infty} \lim_{u\to\infty} (I_{x,u}(b) - \delta b) = \nu = \tilde{\nu}$.

(ii) $\lim_{u\to\infty} \lim_{b\to\infty} (I_{x,u}(b) - \delta b) = \nu(x) + \tilde{\nu}(x)$.

Proof: (i) Since by assumption (1.8) holds, it suffices to show that $\lim_{u\to\infty} I_{x,u}(b) = I(b)$ for all x and b. Since X is assumed to irreducible and aperiodic, $\tilde{R}_{u-t}(x, dy) \to q(dy)$ weakly as $u \to \infty$, and hence

$$\lim_{u\to\infty} \Lambda_{x,u,t}(\theta, x) = \tilde{\lambda}_t(\theta) = \lambda_t(\theta) \tag{3.10}$$

independent of x, pointwise for t and θ. Thus by Lemma 1 of [15],

$$\lim_{u\to\infty} (t\Lambda_{x,u,t})^*(b) = (t\lambda_t)^*(b) \tag{3.11}$$

pointwise for t, and for b in the effective domain of λ_t^*. Then we get a lower bound:

$$\begin{aligned}
\liminf_{u\to\infty} I_{x,u}(b) &= \liminf_{u\to\infty} \inf_{t>0} (t\Lambda_{x,u,t})^*(b) && (3.12)\\
&\geq \inf_{t>0} \lim_{u\to\infty} \inf (t\Lambda_{x,u,t})^*(b) && (3.13)\\
&= \inf_{t>0} (t\lambda_t)^*(b) = I(b). && (3.14)
\end{aligned}$$

For the corresponding upper bound, observe that for any $t' > 0$, $I_{x,u}(b) \leq (t'\Lambda_{x,u,t'})^*(b)$ so that

$$\limsup_{u\to\infty} I_{x,u}(b) \leq (t'\lambda'_t)^*(b). \tag{3.15}$$

Since t' is arbitrary

$$\limsup_{u\to\infty} I_{x,u}(b) \leq \inf_{t\geq 0} (t\lambda_t)^*(b) = I(b). \tag{3.16}$$

(ii) From (1.8) and (3.2) we have

$$\lim_{b\to\infty} (I_{x,u}(b) - \delta b) = \lim_{t\to\infty} t\Lambda_{x,u,t}(\delta) = u\tilde{\lambda}_u(\delta, x) + \lim_{t\to\infty} t\lambda_t(\delta, x) \tag{3.17}$$

from which the result follows upon taking $u \to \infty$.

∎

We can make a more precise identification of the location of the change in the offset $I_{x,u}(b) - \delta b$ between the two limiting values in Theorem 2. Roughly speaking in the offset moves outwards as u increases and is located around $b = u\lambda'(\delta)$.

Theorem 3 *For each $k \in (0,1)$ or $(1,\infty)$ there exist k_u with $\lim_{u\to\infty} k_u = k$ such that*

$$\lim_{u\to\infty} I_{x,u}(k_u u\lambda'(\delta)) - \delta k_u u\lambda'(\delta) = \begin{cases} \nu & \text{if} \quad 0 < k < 1 \\ \nu(x) + \tilde{\nu}(x) & \text{if} \quad k > 1 \end{cases} . \tag{3.18}$$

Proof: Let $\tau(u) = ku$ with $k > 1$. Combining (3.2) and Proposition 6 we have

$$(u - \hat{\tau}(u))\lambda_{\hat{\tau}(u)-u}(\theta, x) + u\tilde{\lambda}_u(\theta x) \le I_{x,u}(ku\Lambda'_{x,u,ku}(\delta)) - \delta ku\Lambda'_{x,u,ku}(\delta) \le u(1-k)\lambda_{u(k-1)}(\theta, x) + u\tilde{\lambda}_u(\theta, x). \tag{3.19}$$

As $u \to \infty$, the upper bound converges to $\nu(x) + \tilde{\nu}(x)$. Now we have seen in Proposition 6 that $\hat{\tau}(u)/\tau(u) \to 1$ and $k_u := k\Lambda'_{x,u,\tau(u)}(\delta)/\lambda'(\delta) \to k$ as $u \to \infty$. From this it is not difficult to show that the lower bound converges also to $\nu(x) + \tilde{\nu}(x)$ and we are done.

When $k < 1$ the proof is similar, except now the upper bound is $-\log \int \tilde{R}_{u(1-k)}(x, dy)e^{ku\tilde{\lambda}_{ku}(\theta,y)}$, with a similar lower bound having $\hat{\tau}(u)$ is place of ku. As $u \to \infty$, then by the weak convergence of $\tilde{R}_{u(1-k)}(x, dy)$ to the stationary distribution $q(dy)$ and (2.5), this converges to $-\log \int q(dy)e^{-\nu(y)}$. This is equal to ν by the following Proposition 7(iv). ∎.

Theorem 3 says that the step in the offset becomes sharp at $b = u\lambda'(\delta)$, as measured in a time scale proportional to u. This phenomenon conforms to the sharpness of relaxation times found in classes of Markov processes when their dimensionality becomes large; see e.g. [1] for some results and references.

What can be said about the accuracy of approximations based on Theorems 1, 2 and 3 as applied to models with finite L? It is not difficult to show that the convergence in Theorem 1 is uniform for b in bounded sets. This means that predictions of the shape of I will follow the shape of the loss curve for finite L arbitrarily closely (out to any b) for sufficiently large L. Moreover, a finer analysis of the convergence for superpositions of independent sources shows that the the error $I(b) - L^{-1}\log \mathbf{P}[Q > Lb]$ is $O(L^{-1}\log L)$ as $L \to \infty$ [28].

The relation between $\nu(x)$ and $\tilde{\nu}(x)$. We have seen that (trivially) $\nu = \tilde{\nu}$. We now examine the relation between the conditioned offsets $\nu(x)$ and $\tilde{\nu}(x)$. As before let (X, W) the backward MAP, with transformed kernel $\hat{P}_t(\theta)$ and (maximal) right eigenvector $r(\theta)$ and left eigenmeasure $\ell(\theta)$. The same symbols, with a tilde, are used to denote the corresponding quantities for the forward process. We abbreviate $r(\theta)$ and $\ell(\theta)$ by simply r and ℓ.

Proposition 7 *(i) $\tilde{r} = d\ell/dq$ and hence $r = d\tilde{\ell}/dq$*

(ii) For a reversible MAP, $r = d\ell/dq$ and hence $\nu(x) = \tilde{\nu}(x)$.

(iii) $\mathbf{E}[e^{-\nu(X_0)-\tilde{\nu}(X_0)}] = e^{-\nu}$.

(iv) $\mathbf{E}[e^{-\nu(X_0)}] = e^{-\nu}$.

Proof: (i) Let $A, B \in \mathcal{E}$. Then

$$\mathbf{E}[e^{\theta W_t}; X_t \in A, X_0 \in B] = \int dq(x)I_B(x)\hat{P}_t(x, A; \theta) = \int dq(x)I_A(x)\tilde{\hat{P}}_t(x, B; \theta). \tag{3.20}$$

Thus

$$\tilde{\hat{P}}_t(y, B; \theta) = \frac{d\int dq(x)I_B(x)\hat{P}_t(x, \cdot; \theta)}{dq}(y), \tag{3.21}$$

and so

$$\int \tilde{\hat{P}}_t(y, dz; \theta)\frac{d\ell}{dq}(z) = \frac{d\int d\ell(x)\hat{P}_t(x, \cdot; \theta)}{dq}(y) = e^{t\lambda(\theta)}\frac{d\ell}{dq}(y). \tag{3.22}$$

The second relation follows by symmetry.

(ii) $r = d\ell/dq$ is a trivial consequence of (i) and $\hat{P} = \tilde{\hat{P}}$ in the reversible case. Use $\mathbf{1}$ to indicate the identity function, and (ℓ, r) for $\int \ell(dx)r(x)$. Since in the general

$$e^{-\nu(x)} = \frac{r(x)(\ell, \mathbf{1})}{(\ell, r)}, \tag{3.23}$$

$$e^{-\tilde{\nu}(x)} = \frac{\tilde{r}(x)(\tilde{\ell}, \mathbf{1})}{(\tilde{\ell}, \tilde{r})} = \frac{\frac{d\ell}{dq}(x)(q, r)}{(\ell, r)}, \tag{3.24}$$

then specializing to the reversible case with $r = d\ell/dq$ the result obtains.

(iii) From (3.23)

$$\mathbf{E}[e^{-\nu(X_0)-\tilde{\nu}(X_0)}] = \frac{(q, r\frac{d\ell}{dq})(q, r)(\ell, \mathbf{1})}{(\ell, r)^2} = \frac{(q, r)(\ell, \mathbf{1})}{(\ell, r)} = e^{-\nu}. \tag{3.25}$$

(iv)

$$\mathbf{E}[e^{-\nu(X_0)}] = \mathbf{E}[\lim_{t\to\infty} e^{t\lambda_t(\delta, X_0)}] = \lim_{t\to\infty} \mathbf{E}[e^{t\lambda_t(\delta, X_0)}] = \lim_{t\to\infty} e^{t\lambda_t(\delta)} = e^{-\nu} \tag{3.26}$$

Here, the second step follows by dominated convergence, since by Prop. 3(ii,iii), $\mathbf{E}[e^{t\lambda_t(\delta,x)}]$ is bounded uniformly for $x \in E$ and $t > 1$. ■

4 Heuristics for the time dependent behavior

Theorem 2 shows that for fixed large u there is a shift in $I_u(b) - \delta b$ between its value for smaller b and that for larger b. Theorem 3 suggests that in describing the shape of $I_u(b)$ we use the following:

Rough Heuristic.

$$I_{x,u}(b) \approx \begin{cases} \delta b + \nu & b < u\lambda'(\delta) \\ \delta b + \nu(x) + \tilde{\nu}(x) & b > u\lambda'(\delta) \end{cases} \tag{4.1}$$

The heuristic is specified by four parameters: the offsets $\nu, \nu(x), \tilde{\nu}(x)$ and the speed $\lambda'(\delta)$. This gives considerable simplification of description as compared with the full conditioned shape function. The parameters can be found using the formulas in Theorem 3 based on the eigenfunction and eigenmeasure of the transformed kernel $\hat{P}_t(\delta)$. Examples of explicit calculations of these can be found in [14].

By adding one further parameter we gain some more detail of the time dependent shape function $I_{x,u}(b)$ in the region of the step at $b = u\lambda'(\delta)$. We can use the upper bound in Theorem 6 to give an approximation for $I_{x,u}$. We parameterize b as $b = u\lambda'(\delta) + c$ and set $\tau(u) = u + c/\lambda'(\delta)$. We now combine with (3.2). Taking first the the case $c < 0$, i.e. $t < u$, then when u and hence t are large, $e^{t\tilde{\lambda}_t(\theta,y)}$ will have settled down to $e^{-\nu(y)}$. Likewise for $c > 0$, i.e. $t > u$, $u\tilde{\lambda}_u(\theta, x)$ will have settled down to $e^{-\nu}$. Thus we have the following:

Approximation.

$$I_u(u\lambda'(\delta) + c) - (u\lambda'(\delta) + c)\delta \approx \begin{cases} -\log\int \tilde{R}_{-c/\lambda'(\delta)}(x, dy)e^{-\tilde{\nu}(y)} & \text{if } c \le 0 \\ \tilde{\nu}(x) - (c/\lambda'(\delta))\lambda_{c/\lambda'(\delta)}(\delta) & \text{if } c \ge 0 \end{cases} \tag{4.2}$$

In this approximation, as c increases (or decreases) from 0 we see that the offset $I_u(u\lambda'(\delta)+c)-\delta(u\lambda'(\delta)+c)$ relaxes to the extreme values as determined in Theorem 2. From (4.2) we see that the rate of convergence for $c<0$ is determined by that of $c \mapsto \tilde{R}_{-c/\lambda'(\delta)}$. Thus for $c<0$ relevant relaxation distance is $B_- := \lambda'(\delta)/v_-$ where v_- is the is the (negative of the) supremum of the spectrum of the generator of $\tilde{R}$ after the maximal eigenvalue 0 has been removed.

Similarly, for $c>0$ the relevant relaxation distance is $B_+ := \lambda'(\delta)/v_+$ where v_+ is the (negative logarithm of the) second highest eigenvalue of $\hat{P}_1(\delta)$. Together, these yield the following:

Full Heuristic. *For large u,*

$$I_u(b) \approx \begin{cases} \delta b + \nu & b < u\lambda'(\delta) - MB_- \\ \delta b + \nu(x) + \tilde{\nu}(x) & b > u\lambda'(\delta) + MB_+ \end{cases}, \tag{4.3}$$

for some multiplier $M \geq 1$, and joins continuously between these extremes for $b \in [u\lambda'(\delta) - MB_-, u\lambda'(\delta) + MB_+]$

The multiplier M is somewhat arbitrary: in an exponential decay we might take $M=3$ in order to get within about 5% of the limiting offsets ν and $\nu(x)+\tilde{\nu}(x)$ (measured as a proportion of the difference between them) at $b = u\lambda'(\delta) \pm MB_\pm$. We can describe the full heuristic as saying that $I_{x,u}$ has a 'step' which has a width roughly equal to $\lambda'(\delta)$ times the relaxation distance $M(B_- + B_+)$ and which propagates along the shape function, linearly with u at rate $\lambda'(\delta)$.

Estimation of tail probabilities for heterogeneous conditionings. The location and width of the step are independent of the conditioning x. Consequently, given a limiting distribution μ of conditionings at time $t=0$, such as appears in Theorem 1, we can adapt any of the theorems of section 3 or the heuristics and approximations of the present section by integration of the conditioned offsets $\nu(x)$ or $\tilde{\nu}(x)$ against μ.

To estimate tail probabilities at large finite L use the limit (1.3) to furnish the appropriate approximation for a conditioning distribution μ_L. Thus, for example, the rough heuristic becomes:

$$\mathbf{P}[Q^L > Lb] \approx e^{-L\int d\mu_L(x) I_{u,b}(b)} \approx \begin{cases} e^{-L(\delta b+\nu)} & b < u\lambda'(\delta) \\ e^{-L(\delta b+\nu(\mu_L)+\tilde{\nu}(\mu_L))} & b > u\lambda'(\delta) \end{cases} \tag{4.4}$$

5 Application to renewal and On-Off arrivals.

A special case of MAP's is furnished by renewal processes. Let $(N_t)_{t\in\mathbf{R}_+}$ be a (possibly delayed) renewal process with renewal epochs $(T_n)_{n\in\mathbf{Z}_+}$. Then (X_t, N_t) is a Markov Additive Process when $X_t = t - T_{N_t}$: the time since the last arrival before time t. If we want to calculate the complete conditioned shape function via the variational principle (1.6) then we can calculate $\lambda_t(\theta, X_0)$ within the MAP framework. On the other hand, the *rough* heuristic (4.1) is characterized by a reduced set of quantities: $\delta, \nu, \nu(x), \lambda$. In the remainder of this section we show how, if one is content to use the rough heuristic rather than the full heuristic or the shape function, the other quantities can be calculated fairly directly for renewal and alternating renewal (ON-OFF) processes without recasting them as MAP's.

Renewal Processes. For the renewal process (N_t) with renewal epochs (T_n) let F be the common distribution of the interarrival times $\{T_{n+1}-T_n,\, n \geq 1\}$ and G be the distribution of T_1. We assume that F has finite mean m, and no atom at 0. Recall (see e.g. [30]) that (N_t) is stationary when $G = G_{\rm st}$ where $G_{\rm st}$ has Laplace transform (LT)

$$\hat{G}_{\rm st}(r) = (1 - \hat{F}(r))/(mr), \tag{5.1}$$

where $\hat{F}$ is the LT of F, etc.

Consider a backward arrival process $A_t = aN_t$, some $a > 0$, served at rate s. For the purposes of calculating the parameters of the rough heuristic only, the following renewal theoretic calculation suffices. Let $e^{t\lambda_t(\theta;G)}$ denote the expectation of $e^{\theta(A_t - st)}$ when T_1 has distribution G. For stability we require $ms > 1$.

Theorem 4 *(i) The Laplace Transform with respect to t of $e^{t\lambda_t(\theta;G)}$ is $\hat{Z}(a\theta, s\theta)$ where*

$$\hat{Z}(\omega,\phi) = \frac{1-\hat{G}(\phi)}{\phi} + \frac{e^{\omega}\hat{G}(\phi)(1-\hat{F}(\phi))}{\phi(1-e^{\omega}\hat{F}(\phi))}\,, \qquad \phi > \hat{F}^{-1}(e^{-\omega}). \tag{5.2}$$

(ii) $\lambda(\theta) = \lim_{t\to\infty}\lambda_t(\theta;G)$ exists independent of G and is equal to $\hat{F}^{-1}(e^{-a\theta}) - s\theta$.

(iii) δ is the non-zero root of $\hat{F}(s\delta)e^{a\delta} = 1$ and $\lambda'(\delta) = -1 - ae^{-a\delta}/\hat{F}'(s\delta)$.

(iv) $e^{-\nu(G)} := \lim_{t\to\infty} e^{t\lambda_t(\delta;G)} = \dfrac{(e^{-a\delta}-1)\hat{G}(s\delta)}{s\delta\hat{F}'(s\delta)}$.

(v) $e^{-\nu} = e^{-\nu(G_{\rm st})} = \dfrac{-1}{m\hat{F}'(s\delta)}\left(\dfrac{1-e^{-a\delta}}{s\delta}\right)^2$.

Consistent with the notation of section 2, $\nu(x)$ will denote $\nu(G)$ in the case G is the measure with a single atom at x.

Laplace transform methods. In identifying δ and ν for renewal and ON-OFF models, we shall find it convenient to adopt the following strategy. Consider for simplicity the homogeneous case with unit service rate per source. Define $Z(\theta,t) = \mathbf{E}[e^{\theta A_t}]$, and define the Laplace transform of $Z(\theta)$ with respect to t by $\hat{Z}(\theta,r) = \int_0^\infty Z(\theta,t)e^{-rt}\,dt$. Note that

$$e^{t\lambda_t(\theta)} = e^{-\theta t}Z(\theta,t). \tag{5.3}$$

Suppose now that some $k(\theta) > 0$ and $\rho(\theta)$ we can demonstrate that

$$\varepsilon\int_0^\infty e^{t\lambda_t(\theta)}e^{-(\varepsilon+\rho(\theta))t}\,dt = \varepsilon\hat{Z}(\theta,\theta+\rho(\theta)+\varepsilon) \to k(\theta), \qquad (\varepsilon\to 0^+). \tag{5.4}$$

Then it follows from Karamata's Tauberian Theorem (see Theorem 1.7.6 in [3]) that

$$\lim_{t\to\infty} e^{t(\lambda_t(\theta)-\rho(\theta))} = k(\theta). \tag{5.5}$$

Thus $\lambda(\theta) = \lim_{t\to\infty}\lambda_t(\theta)$ exists and is equal to $\rho(\theta)$. We can identify $\lambda(\theta)$ by the requirement that $\hat{Z}(\theta,\theta+\lambda(\theta)+\varepsilon)$ diverges as $\varepsilon\to 0^+$, i.e.

$$\lambda(\theta) = \inf\{\rho : \hat{Z}(\theta,\theta+\rho) < \infty\}. \tag{5.6}$$

We identify δ by the requirement that $\lambda(\delta) = 0$. Since $\lambda'(\delta) > 0$ this gives

$$\delta = \sup\{\theta : \hat{Z}(\theta,\theta) < \infty\}. \tag{5.7}$$

Finally choosing $\theta = \delta$ in (5.5) we see that $e^{-\nu} = \lim_{t\to\infty} e^{t\lambda_t(\delta)}$ exists and is equal to $k(\delta)$.

Proof of Theorem 4: (i) Let $Z(\theta, t) = \mathbf{E}[e^{\theta N_t}]$ and denote by $Y(\theta, t)$ this expectation for the corresponding non-delayed renewal process (i.e. with $G = F$)). By a standard renewal argument we have

$$Y(\theta; t) = 1 - F(t) + e^{\theta}(F * Y(\theta))(t) \tag{5.8}$$

$$Z(\theta; t) = 1 - G(t) + e^{\theta}(G * Y(\theta))(t). \tag{5.9}$$

Taking Laplace transforms with respect to t we obtain (5.2).

(ii) From (5.2) and (5.6) and taking into account now the arrival size a and service rate s, then $\lambda(\theta)$ is defined by the requirement that the denominator of $\hat{Z}(a\theta, s\theta + \lambda(\theta))$ be infinite. Thus it is the the solution ρ of the equation $e^{a\theta}\hat{F}(s\theta + \rho) = 1$, i.e., $\lambda(\theta) = \hat{F}^{-1}(e^{-a\theta}) - s\theta$.

(iii) Follows from (ii) by the requirement that $\lambda(\delta) = 0$, so that δ is the positive root (unique by convexity of $\log \hat{F}$) of

$$e^{a\delta}\hat{F}(s\delta) = 1. \tag{5.10}$$

(We remark that it is possible to obtain this equation for δ from (5.2) directly within the framework of transient renewal theory: see XI.6 of [22]). The formula for $\lambda'(\delta)$ is obtained by differentiation of the result in (ii).

(iv) Follows by evaluation of $k(\delta)$:

$$e^{-\nu} = \lim_{\varepsilon\to 0+} \varepsilon \hat{Z}(a\delta, s\delta + \varepsilon) \tag{5.11}$$

$$= \frac{(e^{a\delta} - 1)\hat{G}(s\delta)}{s\delta} \lim_{\varepsilon\to 0+} \frac{\varepsilon}{1 - e^{a\delta}\hat{F}(s\delta + \varepsilon)} \tag{5.12}$$

$$= \frac{(e^{-a\delta} - 1)\hat{G}(s\delta)}{s\delta\hat{F}'(s\delta)}. \tag{5.13}$$

(v) then follows from (iv) by inserting (5.1) into (5.13). ■

Alternating renewal (ON-OFF) processes. The above methods can be extended to treat alternating renewal (ON-OFF) processes in the case the the on-duration has finite moment generating function in some neighborhood of the origin. For simplicity we treat the case that all on and off periods are mutually independent. The on periods have distribution F, the off periods distribution H. We let m_F and m_H denote the mean of F and H respectively. Initially, the process is ON with probability p_0, in which case the time till the start of the next OFF period has distribution F_0; it is off with probability $1 - p_0$, in which case the time till the start of the next ON period has distribution H_0. In the stationary case

$$p_0 = p_{\mathrm{st}} := \frac{m_F}{m_F + m_H}, \quad \hat{F}_0(r) = \hat{F}_{\mathrm{st}}(r) := \frac{1 - \hat{F}(r)}{rm_F}, \quad \text{and} \quad \hat{H}_0(r) = \hat{H}_{\mathrm{st}}(r) := \frac{1 - \hat{H}(r)}{rm_H}. \tag{5.14}$$

In the ON periods, fluid arrives at rate a. Finally, as usual, we consider the superposition of L such sources, served at rate sL.

Let T_t denote the amount of time a source is ON during the interval $[-t, 0)$, and set $Z(\theta; t) = \mathbf{E}[e^{\theta T_t}]$, with $Y(\theta, t)$ denoting the corresponding expectation conditioned on an ON period starting at time 0. Then

$$Z(\theta; t) = p\left(e^{\theta t}(1 - F_0(t)) + (F_0(\theta; \cdot) * (1 - H))(t) + (F_0(\theta; \cdot) * H * Y(\theta; \cdot))(t)\right) \tag{5.15}$$

$$+(1-p)\,(1-H_0(t)+(H_0*Y(\theta;\cdot))(t)) \tag{5.16}$$

$$Y(\theta;t) \;=\; e^{\theta t}(1-F(t))+(F(\theta;\cdot)*(1-H))(t)+(F(\theta;\cdot)*H*Y(\theta;\cdot))(t) \tag{5.17}$$

where $dF(\theta;t)=e^{\theta t}dF(t)$ and likewise for F_0. Thus, taking the LT with respect to t, we obtain

$$\hat{Z}(\theta,r) \;=\; \frac{p}{r-\theta}(1-\hat{H}_0(r-\theta))+\frac{p}{r}\hat{F}_0(r-\theta)(1-\hat{H}(r))+\frac{1-p}{r}(1-\hat{H}_0(r)) \tag{5.18}$$

$$+\left(p\hat{F}_0(r-\theta)\hat{H}(r)+(1-p)\hat{H}_0(r)\right)\hat{Y}(\theta,r) \tag{5.19}$$

where

$$\hat{Y}(\theta,r)=\frac{(r-\theta)^{-1}(1-\hat{F}(r-\theta))+r^{-1}\hat{F}(r-\theta)(1-\hat{H}(r)}{1-\hat{F}(r-\theta)\hat{H}(r))}. \tag{5.20}$$

We are now in a position to prove:

Theorem 5 *Suppose the stability condition $p_{\mathrm{st}}a<s$ is satisfied. Then*

(i) $\lambda(\theta)$ is the unique root ρ of the equation

$$\hat{F}((s-a)\theta+\rho)\hat{H}(s\theta+\rho)=1. \tag{5.21}$$

(ii) δ is the unique positive root of the equation $\hat{F}((s-a)\delta)\hat{H}(s\delta)=1$. $\lambda'(\delta)$ is given through

$$(s-a+\lambda'(\delta))\frac{\hat{F}'((s-a)\delta)}{\hat{F}((s-a)\delta)}+(s+\lambda'(\delta))\frac{\hat{H}'(s\delta)}{\hat{H}(s\delta)}. \tag{5.22}$$

(iii) Conditioned on p_0, H_0 and F_0, the offset $\nu(p_0,H_0,F_0)$ is given by

$$e^{-\nu(p_0,H_0,F_0)}=-\frac{p_0\hat{F}_0((s-a)\delta)+(1-p_0)\hat{H}_0(s\delta)}{\hat{F}'((s-a)\delta)\hat{H}(s\delta)+\hat{F}((s-a)\delta)\hat{H}'(s\delta)}\left(\frac{1-\hat{F}((s-a)\delta))}{(s-a)\delta}+\frac{\hat{F}((s-a)\delta)}{s\delta}\right). \tag{5.23}$$

(iv) In the stationary case the offset ν is given by

$$e^{-\nu}=\frac{1}{(m_F+m_H)(\kappa^2\hat{H}'(s\delta)+\hat{F}'((s-a)\delta))}\left(\frac{(\kappa-1)a}{\delta(a-s)}\right)^2, \tag{5.24}$$

where $\kappa=\hat{F}((s-a)\delta)$.

Proof: (i) As in Theorem 4, $\lambda(\theta)$ must make $\hat{Z}(a\theta,s\theta+\lambda(\theta))$ infinite. So from (5.20), $\lambda(\theta)$ is the root ρ of (5.21), when this exists. (If it does not then we take set $\lambda(\theta)=\infty$). One sees that $\partial\rho/\partial\theta$ is negative, so the root is unique when it exists.

(ii) $e^{t\lambda_t(\theta)}=\mathbf{E}[e^{a\theta T_t}]e^{-s\theta t}=\hat{Z}(a\theta,t)e^{-s\theta t}$, so as in the case of renewal processes we find $\delta=\sup\{\theta:\hat{Z}(a\theta,s\theta)<\infty\}$ to be the unique positive root of the equation

$$\hat{F}((s-a)\delta)\hat{H}(s\delta)=1. \tag{5.25}$$

A familiar convexity argument shows that the unique positive solution exists if the derivative of the left hand side w.r.t δ at $\delta=0$ is positive. This condition reduces to $p_{\mathrm{st}}a<s$. Note that the root is independent of the

initial distribution specified by F_0, H_0 and p_0. Setting $\rho = \lambda(\theta)$ (5.21) and differentiating w.r.t. θ at $\theta = \delta$ yields (5.22).

(iii,iv) Similarly, by the above Laplace transform methods $e^{-\nu(p_0,\hat{F}_0,\hat{G}_0)} = \lim_{\varepsilon\to 0^+} \varepsilon \hat{Z}(a\delta, s\delta + \varepsilon)$, which yields (5.23). Substituting the stationary distributions in (5.14) yields (5.24) after some algebra. ■

6 Comparison of heuristics, approximation, and shape function.

We illustrate the effects of conditioning on the shape function in a renewal model: the interrupted Poisson process. Although the model is quite simply formulated [27], existing results for distributions appear limited to obtaining the Laplace transform of the stationary distribution of the remaining work; this using matrix methods in $L+1$ dimensions for L-fold superpositions [24]. The value of δ can be obtained by using matrix methods as in, for example, [21, 32]. Fortunately, we may fairly readily go further and calculate the full time dependent shape function in order to test the heuristics against it. First define the interrupted Poisson process as a Markov Modulated Poisson Process $(X_t, N_t)_{t\in\mathbf{R}_+}$. Here X_t is a Markov process $\{0,1\}$, with transitions $0 \to 1$ occurring at rate κ, the reverse transition at rate μ. In state 1, the increments of N_t are Poissonian at rate r; in state 0 N_t remains unchanged. We set the arrival process to be $A_t = aN_t$: each arrival is of fixed size a. It is not difficult to see that N_t is a renewal process; this is established in [27] and the interarrival time distribution derived. From this we could determine the parameters entering into the rough heuristic (4.1) using Theorem 4. However, to make a comparison of this with the full shape function, conditioned on X_0, requires that we calculate the full transient conditioned CGFwithin the Markovian formulation.

We now derive (2.5) for this model. Let F_i denote the distribution of the time till next renewal given the modulated process is in state i. Then an elementary argument shows that

$$F_0 = E_\kappa * F_1 \tag{6.1}$$
$$F_1 = \frac{r}{r+\mu}E_{r+\mu} + \frac{\mu}{r+\mu}E_{r+\mu} * F_0, \tag{6.2}$$

where E_ρ denote the exponential distribution with mean ρ^{-1} and $*$ denotes convolution of distributions. Upon taking the LT and eliminating F_0 we obtain

$$\hat{F}(\phi) = \hat{F}_1(\phi) = \frac{r(\kappa+\phi)}{\phi^2 + \phi(\mu+\kappa+r) + r\kappa}. \tag{6.3}$$

Let $\lambda_t(\theta, \{i\})$ denote the workload CGF conditioned on being in state i. The corresponding initial distribution G is F_i. Upon substitution of F and the appropriate G into (5.2) and inversion of the LT, one obtains (after some algebra):

$$e^{-st\theta}e^{t\lambda_t(\theta,\{0\})} = \frac{c_+(\theta)}{c_+(\theta)-c_-(\theta)}e^{c_-(\theta)t} - \frac{c_-(\theta)}{c_+(\theta)-c_-(\theta)}e^{c_+(\theta)t} \tag{6.4}$$
$$e^{-st\theta}e^{t\lambda_t(\theta,\{1\})} = \frac{c_+(\theta)-r(e^{a\theta}-1)}{c_+(\theta)-c_-(\theta)}e^{c_-(\theta)t} - \frac{c_-(\theta)-r(e^{a\theta}-1)}{c_+(\theta)-c_-(\theta)}e^{c_+(\theta)t}, \tag{6.5}$$

where

$$2c_\pm(\theta) = r(e^{a\theta}-1) - (\kappa+\mu) \pm \sqrt{(\kappa+\mu+r(e^{a\theta}-1))^2 - 4\mu r(e^{a\theta}-1)}. \tag{6.6}$$

Finally, for the stationary renewal process we have

$$e^{t\lambda_t(\theta)} = \frac{\mu}{\mu+\kappa} e^{t\lambda_t(\theta,\{0\})} + \frac{\kappa}{\mu+\kappa} e^{t\lambda_t(\theta,\{1\})}, \tag{6.7}$$

i.e. a sum of the $e^{t\lambda_t(\theta,\{i\})}$ weighted by the stationary probabilities of each of the modulating states. The reader should note that the derivation of (6.4) to (6.7) is just a derivation of the spectral decomposition of the transformed kernel $\hat{P}_t(\theta)$ which leads to (2.5): factorization of the denominator in (6.3) find the eigenvalues, and the consequent partial fraction decomposition find the eigenvectors.

We can use the variational expression (1.6) to determine the corresponding shape functions $I(b,\{i\})$ numerically. We have chosen a service rate $s = 1$. We can determine the parameters of the rough heuristic as follows. δ is found by numerical location of the root of the equation got by combining (5.10) and (6.3). (Alternatively, the requirement that $c_+(\delta) = s\delta$ yields the same equation). Comparing (6.4) to (6.7) with (2.5) then $e^{-\nu}, e^{-\nu(\{0\})}$ and $e^{-\nu(\{1\})}$ can be read of as the coefficients of $e^{c_+(\delta)t}$ in the appropriate CGF. A renewal process is reversible, so by Proposition 7(ii) $\nu(x) = \tilde{\nu}(x)$.

We have performed some calculations for the case $\kappa = \mu = 1/10$ and $r = s = 1$, and have displayed the results as follows. In Figure 2 we show, at time zero, the (negative of the) shape function conditioned by the ON state, the OFF state, and also the unconditioned version. The first two of these represent the asymptotic loss curve in the extremes of superpositions with all state ON or all states OFF: the loss curves for mixed superpositions will lie between these. Observe (a) the asymptotic parallelism of the curves—they share the same asymptotic slope δ; and (b) the variation in the offset: $\nu(\{1\}) < 0 < \nu < \nu(\{0\})$.

In Figure 3 we have displayed the difference $I_{\{1\},u}(b) - \delta b$ for the shape function run forward to time $u = 100$, conditioned on all the sources being initially ON. We also display the approximation (4.2) together with the rough heuristic (4.1). The latter is indicated by the horizontal lines for the upper limit ν and the lower limit $\nu(\{1\}) + \tilde{\nu}(\{1\}) = 2\nu(\{1\})$. The parameters for the rough heuristic are $\nu = 0.352$, $2\nu(\{1\}) = -0.296$, $\delta = 0.372$, $\lambda'(\delta) = 0.389$. The step is predicted to be located around $u\lambda'(\delta) \approx 38.9$.

For the full heuristic, we must invoke the MAP formulation of the model. The generator of the modulating process is the matrix

$$\begin{pmatrix} -\kappa & \kappa \\ -\mu & -\mu \end{pmatrix} = \frac{1}{10} \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \tag{6.8}$$

so that $v_- = 1/10$. Thus with the multiplier $M = 3$ we expect the half-step width on the left hand side of roughly $M\lambda'(\delta)/v_- \approx 12$. From (6.5) we have $v_+ = \delta - c_-(\delta) = 0.494$, yielding the half-step width on the right hand side of $M\lambda'(\delta)/v_+ \approx 2.4$. To summarize, in the full heuristic, the step extends, roughly, from $b = 27$ to $b = 41$. This is close agreement with both the full shape function and the approximation from which the full heuristic was derived. The difference of the calculated and approximating curves, which lies mostly within the step, was investigated further. It turns out that in this region, $t \mapsto (t\Lambda_{\{1\},u,t})^*(b)$ has two local minima. The larger of these is essentially that which is used in the approximation, for it is located very close to $\tau(u)$. Outside the step, the two minima coalesce.

Finally, in Figure 4 we illustrate the propagation of the step in $-I_{\{0\},u}(b)$ as u increases. This is done using the approximation (4.2) for the same model but in this instance with all sources initially OFF.

References

[1] V. Anantharam, Threshold phenomena in the transient behavior of Markovian models of communication networks and databases. Queueing Systems, 5 (1989) 77–98.

[2] V. Anantharam. How large delays build up in a GI/G/1 queue. Queueing Systems, 5 (1988) 345–368.

[3] N.H. Bingham, C.M. Goldie and J.L. Teugels, Regular Variation, Encyclopedia of Mathematics and its Applications, Vol 27. Cambridge University Press, 1989.

[4] A.A. Borovkov, Asymptotic Methods in Queueing Theory, Wiley, Chichester, 1984.

[5] D.D. Botvich and N.G. Duffield, Large deviations, the shape of the loss curve, and economies of scale in large multiplexers. Queueing Systems, 20 (1995) 293-320.

[6] C.S. Chang, Stability, queue length and delay of deterministic and stochastic queueing networks, IEEE Trans. on Automatic Control, 39 (1994) 913–931.

[7] G.L. Choudhury, D.M. Lucantoni and W. Whitt, Squeezing the most out of ATM, IEEE Transactions on Communications, 43 (1995) 1566–1579.

[8] T.J. Corcoran, Prediction of ATM multiplexer performance by simulation and analysis of a model of packetized voice traffic, M.Sc. Thesis, Dublin City University, 1994.

[9] C. Courcoubetis and R. Weber, Buffer overflow asymptotics for a switch handling many traffic sources, J. Appl. Prob., 33 (1996) 886-903.

[10] S. Crosby, I. Leslie, J.T. Lewis, N. O'Connell, R. Russell and F. Toomey, Bypassing modelling: an investigation of entropy as a traffic descriptor in the Fairisle ATM network. *Proceedings of the 12th UK Teletraffic Symposium,* Old Windsor, 1995.

[11] J.N. Daigle and J.D. Langford, Models for analysis of packet voice communication systems, IEEE Journal on Selected Areas in Communications, 4 (1986) 847–855.

[12] A. Dembo and T. Zajic, Large deviations: from empirical mean and measure to partial sums process, Stoch. Proc. Appl. 57 (1995) 191–224.

[13] A. Dembo and O. Zeitouni, Large Deviation Techniques and Applications, Jones and Bartlett, London, 1993.

[14] N.G. Duffield, Exponential bounds for queues with Markovian arrivals, Queueing Systems, 17 (1994) 413–430.

[15] N.G. Duffield, Economies of scale in queues with sources having power-law large deviation scalings. J. Appl. Prob., 33 (1996) 840-857.

[16] N.G. Duffield, Economies of scale for long-range dependent traffic in short buffers, Telecommunications Systems. (1996) to appear.

[17] N.G. Duffield, J.T. Lewis, N. O'Connell, R. Russell and F. Toomey, Entropy of ATM traffic streams: a tool for estimating QoS parameters, IEEE Journal on Selected Areas in Communications, 13 (1995) 981–990.

[18] N.G. Duffield and N. O'Connell, Large deviations and overflow probabilities for the general single-server queue, with applications. Math. Proc. Cam. Phil. Soc., 118 (1995) 363-374.

[19] N.G. Duffield and W. Whitt. Recovery from Rare Congestion Events in a Large Multi-Server System, Preprint, AT&T Laboratories, 1996.

[20] R.S. Ellis, Entropy, Large Deviations, and Statistical Mechanics, Springer, New York, 1985.

[21] A.I. Elwalid and D. Mitra, Effective bandwidth of general Markovian traffic sources and admission control of high speed networks. IEEE/ACM Transactions on Networking, 1:329–344, 1993.

[22] W. Feller, An introduction to probability theory and its applications, Volume 2, Wiley, New York, 1971.

[23] P.W. Glynn and W. Whitt, Logarithmic asymptotics for steady-state tail probabilities in a single-server queue, J. Appl. Prob. 31A (1993) 131–159

[24] I. Ide, Superposition of interrupted Poisson processes and its application to packetized voice multiplexers. Proceedings of ITC 12, Torino, June 1–8, 1988.

[25] I. Iscoe, P. Ney and E. Nummelin, Large deviations of uniformly recurrent Markov additive processes, Adv. in Appl. Math., 6 (1985) 373–412

[26] F.P. Kelly, Effective bandwidths at multi-type queues, Queueing Systems, 9 (1991) 5–16

[27] A. Kuczura, The interrupted Poisson process as an overflow process. Bell. Sys. Tech. J., 52 (1973) 437–448

[28] M. Montgomery and G. de Veciana, On the relevance of time-scales in performance oriented traffic characterizations. Proceedings IEEE INFOCOM'96, pp513–520.

[29] R. T. Rockafellar, Convex Analysis, Princeton University Press, Princeton, 1970.

[30] S.M. Ross, Stochastic Processes, Wiley, New York, 1983.

[31] A. Simonian and J. Guibert, Large deviations approximation for fluid queues fed by a large number of on-off sources, Proceedings of ITC 14, Antibes, 1994, pp 1013–1022.

[32] K. Sohraby, On the theory of general ON-OFF sources with applications in high speed networks, Proceedings IEEE INFOCOM'93.

[33] A. Weiss, A new technique for analysing large traffic systems, J. Appl. Prob., 18, (1986) 506–532.

[34] W. Whitt, Tail probabilities with statistical multiplexing and effective bandwidths in multi-class queues, Telecommunications Systems, 2 (1993) 71–107.

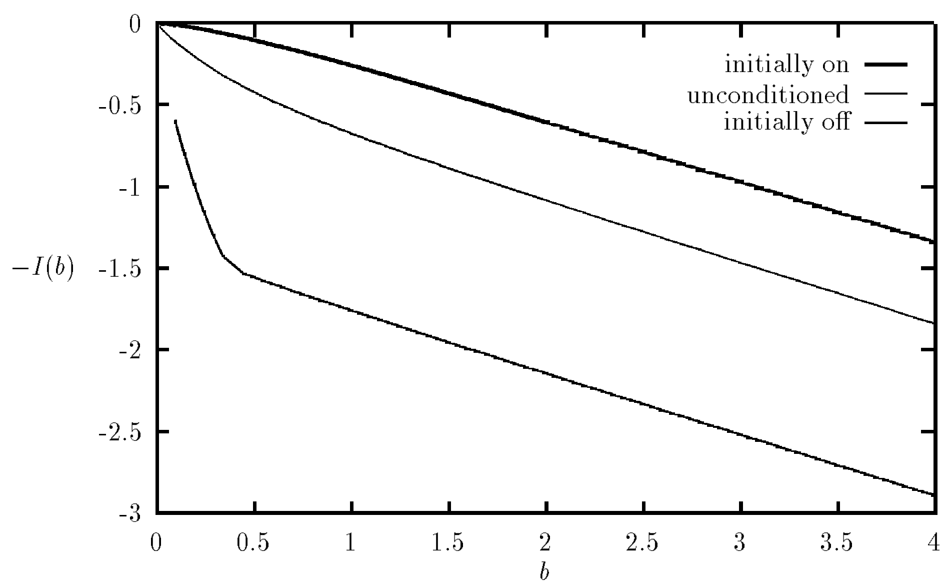

Figure 2: Conditioned and unconditioned shape functions.

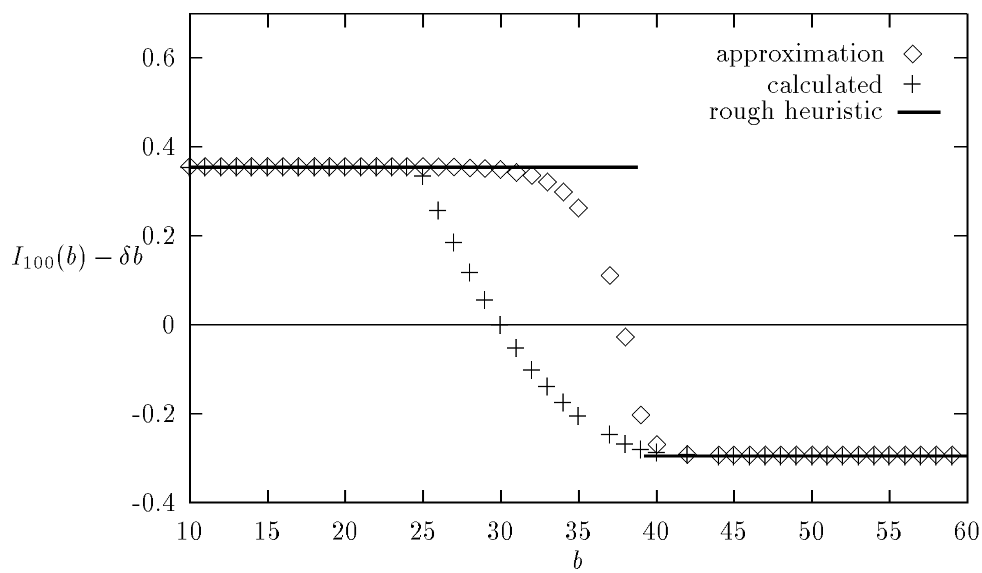

Figure 3: $I(b) - \delta b$ at $u = 100$, calculated, approximated, and by rough heuristic.

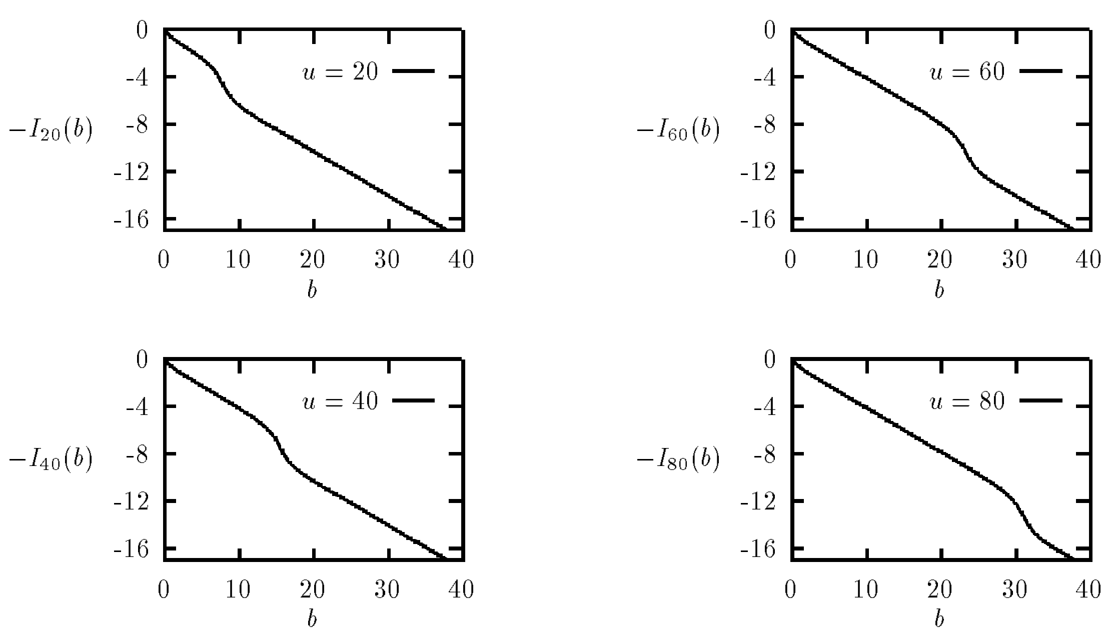

Figure 4: Evolution of shape function at times $u = 20, 40, 60, 80$ after conditioning: approximation