Estimating Tail Probabilities in Queues via Extremal Statistics

Peter W. Glynn* Stanford University

Assaf Zeevi† Stanford University

This version: December 1999

Abstract

We study the estimation of tail probabilities in a queue via a semi-parametric estimator based on the maximum value of the workload, observed over the sampled time interval. Logarithmic consistency and efficiency issues for such estimators are considered, and their performance is contrasted with the (non-parametric) empirical tail estimator. Our results indicate that in order to "successfully" estimate and extrapolate buffer overflow probabilities in regenerative queues, it is in some sense necessary to first introduce a rough model for the behavior of the tails. In the course of developing these results, we establish new almost sure limit theory, in the context of regenerative processes, for the maximal extreme value and related first passage times.

Short Title: Estimating Tail Probabilities in Queues

Keywords: Extreme values, queues, regenerative processes, rare events, estimation, asymptotics.

2000 AMS Subject Classification (Primary): 60G70, 60K25; Secondary: 60K30, 62G05, 62G20

1 Introduction

Consider a finite-buffer queue that is being monitored over time, and suppose that we wish to exert some control over the input process so as to ensure that the proportion of jobs arriving to a full buffer is less than some given value. For example, in a communication network, we may wish to implement some form of admission control so as to ensure that the long-run fraction of dropped packets (or cells) at the buffers feeding the switches is acceptably low. In such settings, we expect that the admission control policy would need to, either explicitly or implicitly, estimate the fraction of time that a buffer is full based on the observed traffic. In particular, it is natural to assume that

*EES&OR, e-mail: glynn@stanford.edu

†Information Systems Lab, e-mail: assaf@isl.stanford.edu

this estimate will be based on the observed buffer occupancy; see [20] and [6] for examples of such admission control policies.

In this paper, our goal is to develop some insight into this estimation problem where "extremal statistics" are used as the statistical vehicle by which to estimate the buffer loss probabilities. By "extremal statistics", we mean here that we shall base our estimator on the behavior of the observed maximum (i.e., the maximal extreme value) of the workload process associated with the system.

We shall simplify this problem in several different ways. First, we shall deal only with a single-station network. Furthermore, we shall replace the finite buffer system with its infinite capacity analogue. In particular, rather than consider the problem of estimating the probability that a finite buffer system is full, we shall instead consider the problem of estimating the probability that the workload of an infinite capacity system is greater then some level b. For large b, it seems reasonable to expect that the estimator we consider here has a qualitatively similar behavior in the finite and infinite capacity settings.

The workload process to the single-server queue is regenerative under quite modest assumptions on the input processes to the queue. For example, the workload is regenerative in the context of renewal input processes in which the inter-arrival and processing times are i.i.d. But regeneration is also a useful theoretical tool in queues with dependent inputs. For example, if the arrival stream to the queue is generated by a Markov-modulated Poisson process with a finite-state irreducible continuous time Markov chain as a modulator we can expect the workload process to be regenerative. Since virtually all of the theory developed here requires only the existence of regeneration structure, we have phrased most of the theory in this paper in terms of general regenerative processes.

We prove that under sharp conditions on the regenerative process, tail probabilities for the marginal distribution of the process may be estimated from the maximal extreme value. We introduce the notion of *logarithmic consistency*, and prove that the maximal extreme value can be used to construct an estimator of the tail probability that is logarithmically consistent when the stationary marginal has an exponential-like tail (Theorem 2.5). We also prove logarithmic consistency for the corresponding tail estimator that can be constructed when the marginal is Pareto-like (Theorem 2.8) or Weibull-like (Theorem 2.10). Part of this study concentrates on contrasting the performance of the extremal based estimator, which will be seen to be essentially semi-parametric, with that of the obvious non-parametric empirical tail estimator. Our results prove that extremal statistics can potentially do a better job of roughly predicting vanishingly small tail probabilities than the empirical tail estimator, at least when the observed time horizon is of "small" to "moderate" size; see Section 2. However, the rate of convergence of the extremal tail estimator is slow. In Section 2 we introduce an extrapolation based empirical tail estimator that has a better theoretical conver-

gence rate than the extremal tail estimator. Research into additional improved tail estimators is ongoing. We note, however, that the extremal estimator has positive characteristics in that its specification does not require any user-defined tuning constants (as opposed to our extrapolation-based estimator) and it is logarithmically consistent under relatively mild assumptions on the observed process.

In the course of our investigation of the extremal estimator, we developed several new results regarding extreme values in the context of regenerative processes (and, hence, for a large class of regenerative queueing systems):

1. Almost sure limit theorems and associated L^p convergence (Theorems 2.5, 2.8, and 2.10) for regenerative extreme values when the tail is exponential-like, Pareto-like, and Weibull-like (Glasserman and Kou [11] prove a similar result in the exponential-like case, but under different hypotheses than ours; Theorem 2.8 and 2.10 are extensions of the limit theory to Pareto-like and Weibull-like tails).

2. A compound Poisson limit theorem for the amount of time that the workload process for a single-server queue spends above level b.

3. An almost sure limit theorem for the passage time required for a regenerative process to exceed level b, when b is large (again, this is an extension of the results in [11] to Pareto and Weibull tails).

The paper is organized as follows. Section 2 introduces the extremal tail estimator in the context of exponential-like, Pareto-like, and Weibull-like tails, and establishes its basic logarithmic consistency properties. In Section 3, we compare the extremal estimator to its most obvious non-parametric competitor, namely the empirical tail estimator. Section 4 discusses the implications of our limit theory for first passage times of regenerative processes. Finally, Section 5 is concerned with some explicit computations when the underlying observed process is reflecting Brownian motion.

2 Consistency Results for Extremal Statistics

We start this section by briefly reviewing some basic terminology and theory associated with single server queues.

Let $\Gamma(t)$ be the cumulative amount of work to arrive to a buffer over the interval $[0, t]$. In communications applications we think of $\Gamma(t)$ as the total amount of packets or cells arriving in

$[0,t]$. Then, $\Gamma = (\Gamma(t) : t \geq 0)$ is a real-valued non-decreasing process with $\Gamma(0) = 0$. Assume in addition that Γ is right-continuous with left limits, and has stationary ergodic increments such that $\mathbb{E}\Gamma(1) < \infty$. Without loss of generality, we may presume that the deterministic server works at a unit rate. Given any right continuous Γ with left limits we can represent the workload $W(t)$ present in the system at time t, if $W(0) = 0$, as

$$W(t) = \Gamma(t) - t - \inf_{0 \leq s \leq t}[\Gamma(s) - s];$$

see, for example, Harrison [18] for this representation of the workload.

If $\mathbb{E}\Gamma(1) < 1$, it is easily shown that

$$W(t) \Rightarrow W(\infty)$$

as $t \to \infty$, where $W(\infty)$ is a proper random variable. In fact, we can construct a probability space supporting both the process Γ and a stationary process $W^* = (W^*(t) : t \geq 0)$ such that

i.) $W^*(t) \stackrel{\mathcal{D}}{=} W(\infty)$ for all $t \geq 0$, where $\stackrel{\mathcal{D}}{=}$ denotes "equality in distribution";

ii.) $W^*(t) = (\Gamma(t) - t) + W^*(0) \bigvee L^*(t)$ where $L^*(t) := -\inf_{0 \leq s \leq t}[\Gamma(s) - s]$, for $t \geq 0$;

see Konstantopoulos, Zazanis, and De Veciana [22] for details. Thus, W^* is a stationary version of the workload process, for the system with input process Γ.

Our goal in this paper is to develop an efficient means of estimating $\alpha(b) = \mathbb{P}(W^*(0) \geq b)$, based on the observed trajectory of the process W^*. In other words, we are concerned in this paper with a special case of the following more general problem:

> Given a real-valued stationary process $X = (X(t) : t \geq 0)$, estimate $\alpha(b) = \mathbb{P}(X(0) \geq b)$ from the observed trajectory $(X(s) : 0 \leq s \leq t)$.

Most of the analysis in this paper will focus on this (more general) estimation problem. In this setting, the most natural estimator of $\alpha(b)$ is the *empirical tail estimator*

$$\hat{\alpha}_1(t; b) = \frac{1}{t} \int_0^t \mathbb{I}_{\{X(s) \geq b\}} ds \quad .$$

Let $\mathcal{M}_1$ be the set of probability measures on the path space of X under which X is stationary and ergodic; for simplicity, in the following discussion we assume that the underlying probability space supporting X is its path space. An immediate consequence of Birkhoff's ergodic theorem is the following (strong) consistency result for $\hat{\alpha}_1(t; b)$.

Proposition 2.1 *If $\mathbb{P} \in \mathcal{M}_1$, then $\hat{\alpha}_1(t;b) \to \alpha(b)$ almost surely (a.s.) as $t \to \infty$ for each $b \geq 0$.*

Remark 2.2 *When Γ has stationary ergodic increments then it follows that the stationary version of the workload process W must necessarily be ergodic as well (cf., for example [22]). It follows from Proposition 2.1 that $\hat{\alpha}_1(t;b)$ is strongly consistent in the queueing context (where the stationary version of the workload process W^* plays the role of X).*

However, alternative estimators for the tail probability become pertinent if we have reason to believe that the tail probability may be suitably modeled. In particular, suppose that the tail is asymptotically exponential in the following sense:

$$\textbf{A1}. \qquad \frac{1}{b} \log \alpha(b) \to -\theta^* \quad as \;\; b \to \infty$$

for $0 < \theta^* < \infty$.

Remark 2.3 *In a queueing context, there is a long history of results that offer sufficient conditions for the validity of* **A1**. *The earliest such result is the classical Cramér-Lundberg approximation for the tail of the steady-state waiting time distribution associated with the single server queue with "light tailed" renewal inputs; see Asmussen [2] for details. More recently, very general results guaranteeing* **A1** *have been developed by Glynn and Whitt [15] and Duffield and O'Connell [7]. These results include queues in which the input process Γ can exhibit complex dependency structure. Under certain conditions, these results extend from a single node to an intree network, which is a useful model of "real-world" high speed ATM networks; the reader is referred to Chang [5] for details.*

The use of extremal statistics is one means of taking advantage of **A1** To illustrate this point, consider a discrete time real-valued stationary sequence $(X_n : n \geq 0)$ for which the tail probability satisfies **A1**. If $M_n = \max\{X_j : 0 \leq j \leq n-1\}$ and $F_n(\cdot)$ is the empirical distribution function corresponding to $(X_j : 0 \leq j \leq n-1)$, then **A1** suggests that

$$\frac{1}{M_n} \log \bar{F}_n(M_n-) \to -\theta^*$$

as $n \to \infty$, where $\bar{F}_n(x) = 1 - F_n(x)$, and M_n is the largest order statistic, such that $\bar{F}_n(M_n-) = 1/n$. Thus, one might expect that under suitable conditions on the X_i's, it ought to be that

$$\frac{M_n}{\log n} \to \frac{1}{\theta^*}$$

as $n \to \infty$, in some suitable sense. As a consequence, we are led to consider the estimator

$$\hat{\alpha}_2(t;b) = \exp\left(-\frac{b \log t}{M(t)}\right) \quad ,$$

where $M(t) = \sup\{X(s) : 0 \leq s \leq t\}$. Because **A1** asserts only that θ^* captures the principal behavior of the logarithm of the probability $\alpha(b)$, we cannot expect $\hat{\alpha}_2(t;b)$ to be strongly consistent for $\alpha(b)$ in general. Instead, we demand that $\hat{\alpha}_2(t;b)$ satisfy a weaker type of consistency.

Definition 2.4 *We say that $\hat{\alpha}(t;b)$ is logarithmically consistent for $\alpha(b)$ if for every (deterministic) function $b(t)\to\infty$ as $t\to\infty$, we have that*

$$\frac{\log \hat{\alpha}(t;b(t))}{\log \alpha(b(t))} \Rightarrow 1$$

as $t\to\infty$.

Roughly speaking, *logarithmic consistency* is a reasonable demand to place upon an estimator when only the order of magnitude of the probability is required. Note that assumption **A1** only provides a rough measure of the tail decay, i.e., only *logarithmic asymptotics* of the tail are given. Thus, in some sense, *logarithmic consistency* is the natural measure of performance to capture this model specification. In the communications networking context, this seems like a reasonable minimal requirement to expect from an estimator.

We now turn to establishing logarithmic consistency for $\hat{\alpha}_2(t;b)$. Now, given that $\log \hat{\alpha}_2(t;b)/\log \alpha(b) = -(\log t/M(t))(b/\log \alpha(b))$ it is clear that the only issue remaining here is obtaining conditions that make rigorous the limit theorem $M(t)/\log t \to 1/\theta^*$ as $t\to\infty$.

To precisely state the main result here, we let $\mathcal{M}_2$ be the set of probabilities under which $X = (X(t) : t \geq 0)$ is a stationary process and classically regenerative with respect to random times $(T(n) : n \geq 0)$ satisfying $0 \leq T(0) < T(1) < T(2) < \cdots$ and in which $\mathbb{E}\tau_1^p < \infty$ for all $p \geq 1$, where $\tau_n := T(n) - T(n-1)$. By classically regenerative we mean that the cycles $((X(T(j-1)+s) : 0 \leq s < \tau_j), \tau_j)$ are independent for $j \geq 0$, where $T(-1) := 0$, and are identically distributed for $j \geq 1$.

Theorem 2.5 *Suppose that $\mathbb{P} \in \mathcal{M}_2$ and that under the probability measure $\mathbb{P}$:*

i.) X satisfies **A1***;*

ii.) there exists $\epsilon > 0$ such that

$$\liminf_{b\to\infty} \mathbb{P}\left(\int_0^{\tau_1} \mathbb{I}_{\{X(T(0)+s)>b\}} ds \geq \epsilon \middle| \beta_1 > b\right) \geq \epsilon \quad .$$

Then,

$$\frac{M(t)}{\log t} \to \frac{1}{\theta^*} \qquad \mathbb{P}-a.s.$$

as $t\to\infty$, and in L^p for any $p \in [1,\infty)$.

Proof: We break the proof up into several steps. In the first step we show that the cycle-maximum r.v. has the same tail behavior (in logarithmic scale) as the stationary marginal. The second step reduces the problem of studying the behavior of the maximum value $M(t)$ to the maximum of a sequence of cycle-maximum r.v.'s, for which the asserted asymptotics are established. Finally, the last step proves the L^p convergence result. We now turn to the detailed derivation.

1^0. Let $\beta_j = \sup\{X(s) : T(j-1) \leq s < T(j)\}$ be the maximum of X over the j'th regenerative cycle. We first show that under the conditions of the theorem,

$$\frac{1}{b}\log \mathbb{P}(\beta_j > b) \rightarrow -\theta^* \tag{2.1}$$

as $b \rightarrow \infty$. By the ratio formula for regenerative processes (cf. [2, p. 126]),

$$\mathbb{P}(X(t) > b) = \frac{1}{\mathbb{E}\tau_1}\mathbb{E}\int_0^{\tau_1} \mathbb{I}_{\{X(T(0)+s)>b\}} ds \quad . \tag{2.2}$$

By assumption ii.) in the theorem, and Markov's inequality we have also that

$$\frac{1}{\mathbb{E}\tau_1}\mathbb{E}\int_0^{\tau_1} \mathbb{I}_{\{X(T(0)+s)>b\}} ds \geq \frac{1}{\mathbb{E}\tau_1}\epsilon^2 \mathbb{P}(\beta_1 > b) \quad ,$$

for sufficiently large b. Hence,

$$\limsup_{b\rightarrow\infty} \frac{1}{b}\log \mathbb{P}(\beta_1 > b) \leq \frac{1}{\theta^*} \quad . \tag{2.3}$$

On the other hand, for $p > 1$ and q such that $p^{-1} + q^{-1} = 1$,

$$\begin{aligned}\frac{1}{\mathbb{E}\tau_1}\mathbb{E}\int_0^{\tau_1} \mathbb{I}_{\{X(T(0)+s)>b\}} ds &\leq \frac{1}{\mathbb{E}\tau_1}\mathbb{E}\tau_1 \mathbb{I}_{\{\beta_1>b\}} \\ &\leq \frac{1}{\mathbb{E}\tau_1}\left(\mathbb{E}\tau_1^p\right)^{1/p} \mathbb{P}^{1/q}(\beta_1 > b)\end{aligned}$$

by Hölder's inequality. So, (2.2) yields

$$\liminf_{b\rightarrow\infty} \frac{1}{b}\log \mathbb{P}(\beta_1 > b) \geq \frac{1}{\theta^*} \tag{2.4}$$

by letting $p \rightarrow \infty$ and $q \downarrow 1$. Relation (2.1) now follows from (2.3) and (2.4).

2^0. We now reduce the problem to the study of the maximum of the β_i sequence. Let $S_n = \sum_{i=1}^n \tau_i$, set $N(t) = \sup\{n \geq 1 : S_n \leq t\}$, and let $M(t) = \sup\{X(s) : s \in (0, t]\}$. Note that by definition of the stationary workload X with dynamics following ii.) in the theorem, we have

$$\frac{\bigvee_{i=1}^{N(t)} \beta_i}{\log t} \leq \frac{M(t)}{\log t} \leq \frac{\beta_0}{\log t} + \frac{\bigvee_{i=1}^{N(t)+1} \beta_i}{\log t} \tag{2.5}$$

where $\beta_0 := \sup\{X(s) : 0 \leq s < T(0)\}$ is due to the delayed regenerative process, and $T(0)$ has the distribution of the forward recurrence time in the associated stationary renewal process. In

particular, $\beta_0/\log t = o(1)$ a.s. It suffices therefore to prove that the normalized maximum of a random number of copies of β_1 converges as asserted. We will do this in two steps.

i.) Upper bound: We first observe that due to the strong law of large numbers (SLLN) for renewal processes we have that $N(t,\omega)/t \to 1/\mathbb{E}\tau_1$ for $\mathbb{P}$ almost all (a.a.) $\omega \in \Omega$. Now, for any $\delta > 0$ we have

$$\begin{aligned}\sum_{n=1}^{\infty} \mathbb{P}(\beta_n > ((1+\delta)\log n)/\theta^*) &= \sum_{n=1}^{\infty} \mathbb{P}(\exp(\theta^*\beta_1/(1+\delta)) > n) \\ &\leq \mathbb{E}\exp(\theta^*\beta_1/(1+\delta)) \\ &< \infty \quad ,\end{aligned}$$

where the last step follows since (2.1) implies that $\mathbb{E}\exp(\theta'\beta_1) < \infty$ for all $\theta' < \theta^*$. Thus, $\beta_n/\log n \leq (1+\delta)/\theta^*$ a.s., for all but finitely many n. Fix an $\omega \in \Omega$ such that the above holds, and such that $N(t,\omega)/t \to 1/\mathbb{E}\tau$. Then, there exists $K(\omega)$ such that

$$\begin{aligned}\frac{\bigvee_{k=1}^{N(t,\omega)} \beta_k(\omega)}{\log N(t,\omega)} &\leq \frac{\bigvee_{k=1}^{K(\omega)} \beta_k(\omega)}{\log N(t,\omega)} + \bigvee_{k=K(\omega)}^{N(t,\omega)} \frac{\beta_k(\omega)}{\log k} \\ &\leq \frac{\bigvee_{k=1}^{K(\omega)} \beta_k(\omega)}{\log N(t,\omega)} + \frac{(1+\delta)}{\theta^*} \quad .\end{aligned}$$

Since $\delta > 0$ is arbitrary, and since $N(t)/t \to \mathbb{E}\tau_1$, we conclude that for $\mathbb{P}-$a.a. $\omega \in \Omega$

$$\limsup_{t\to\infty} \frac{\bigvee_{k=1}^{N(t,\omega)} \beta_k(\omega)}{\log t} \leq \frac{1}{\theta^*} \tag{2.6}$$

ii.) Lower bound: Fix $\delta > 0$, and observe that

$$\begin{aligned}\mathbb{P}\left(\bigvee_{k=1}^{n} \beta_k \leq \frac{((1-\delta)\log n)}{\theta^*}\right) &= \mathbb{P}^n(\beta_1 \leq ((1-\delta)\log n)/\theta^*) \\ &\leq \exp(-n\mathbb{P}(\beta_1 > ((1-\delta)\log n)/\theta^*)) \\ &\leq \exp(-n^{\delta}) \quad ,\end{aligned}$$

where the last step used again (2.1). Thus, $\vee_{k=1}^{n}\beta_k/\log n \geq (1-\delta)/\theta^*$ for all but finitely many n. Since $\delta > 0$ is arbitrary

$$\liminf_{n\to\infty} \frac{\bigvee_{k=1}^{n} \beta_k}{\log n} \geq \frac{1}{\theta^*} \quad \mathbb{P}-a.s. \tag{2.7}$$

To extend this to the case where n is random, we use again the SLLN for renewal processes. Specifically, fix $\omega \in \Omega$ such that $N(t,\omega)/t \to 1/\mathbb{E}\tau_1$. Now, fix $\delta > 0$ sufficiently small. Then, for $t \geq K_1(\omega)$, say, it follows that $|N(t,w)/t - 1/\mathbb{E}\tau_1| \leq \delta$. This in turn implies that

$$\frac{\bigvee_{k=1}^{N(t,\omega)} \beta_k(\omega)}{\log t} \geq \frac{\bigvee_{k=1}^{\lfloor (1/\mathbb{E}\tau_1 - \delta)t \rfloor} \beta_k(\omega)}{\log t}$$

for $t \geq K_1(\omega)$. But according to (2.7) for $t \geq K_2(\omega)$, say, it holds that $(\bigvee_{k=1}^{\lfloor t \rfloor} \beta_k)/\log\lfloor t \rfloor \geq 1/\theta^*$. Thus, it follows that for $\mathbb{P}$ a.a. ω

$$\liminf_{t\to\infty} \frac{\bigvee_{k=1}^{N(t,\omega)} \beta_k(\omega)}{\log t} \geq 1/\theta^* \tag{2.8}$$

which concludes the derivation of the lower bound. Putting together (2.8) and (2.6) with (2.5) in Step 1^0 concludes the proof that $M(t)/\log t \to 1/\theta^*$ almost surely.

3^0. To prove the L^p convergence, it suffices to show that $\{(M(t)/\log t)^p\}_{t\geq 2}$ is a uniformly integrable family of r.v.'s. To that extent, it suffices to prove that for all $p \geq 1$, $\sup_{t\geq 2} \mathbb{E}(M(t)/\log t)^p < \infty$. Recall that $T(0)$ denotes the time of the first renewal in the delayed renewal process. Then,

$$\mathbb{E}\left[\frac{M(t)}{\log t}\right]^p \leq \mathbb{E}\left[\frac{M(T(0))}{\log t}\right]^p + \mathbb{E}\left[\frac{\bigvee_{i=1}^{\lceil t \rceil} \beta_i}{\log t}\right]^p \quad ,$$

where $M(T(0)) = \sup\{X(s) : s \in [0, T(0)]\}$. We first deal with the second term on the right hand side. Define

$$K_1 = \inf\left\{ y > 0 : \text{such that } \frac{\log P(\beta_1 > x)}{x} \leq -\frac{\theta^*}{2}, \ \ \forall x \geq y \right\}$$

and note that $K_1 < \infty$ follows from assumption **A1** and (2.1). Finally, set $K = \max\{K_1, 4/\theta^*\}$. Then,

$$\begin{aligned}
\mathbb{E}\left[\frac{\bigvee_{i=1}^{\lceil t \rceil} \beta_i}{\log t}\right]^p &= \int_0^\infty p y^{p-1} \mathbb{P}\left(\bigvee_{i=1}^{\lceil t \rceil} \beta_i \geq y \log t\right) dy \\
&\leq C_1 + \int_K^\infty p y^{p-1} \lceil t \rceil \mathbb{P}\left(\beta_i \geq y \log t\right) dy \\
&\overset{(a)}{\leq} C_1 + \int_K^\infty p y^{p-1} \exp(\log\lceil t \rceil) \exp(-(\theta^* y \log t)/2) dy \\
&\overset{(b)}{\leq} C_1 + \int_0^\infty p y^{p-1} \exp(-(\theta^* y \log t)/4) dy \\
&\leq C_1 + C_2/(\theta^* \log t)^p \\
&< \infty \quad ,
\end{aligned}$$

where the first inequality follows from the union bound, and (a) and (b) follow from the choice of K, and $C_1 = K^p$.

Turning our attention to the first term on the right hand side, by ergodicity of X we can appeal to a "path" version of the regenerative ratio formula (see, e.g., [12]). In particular, with $M(T(0)) = \sup_{t\in(0,T(0)]} X(t)$ we have that

$$\mathbb{E}^*\left[M(T(0))\right]^p = \frac{1}{\mathbb{E}^0 \tau_1} \mathbb{E}^0 \int_0^{\tau_1} \left[\sup_{t\in(s,\tau_1]} X(t)\right]^p ds \ \ ,$$

where $\mathbb{E}^*\{\cdot\}$ and $\mathbb{E}^0\{\cdot\}$ denote expectation w.r.t the stationary, and zero delayed versions of X respectively. Thus,

$$\mathbb{E}^* \left[\frac{M(T(0))}{\log t} \right]^p \leq \frac{1}{\mathbb{E}^0 \tau_1} \frac{\mathbb{E}^0[\tau_1 \beta_1^p]}{(\log t)^p}$$

and by assumption $\mathbb{E}^0 \tau_1^q$ is finite for all $q \geq 1$, and (2.1) ensures the same is true for β_1. This concludes the proof. ■

Corollary 2.6 *Under the conditions of Theorem 1 $\hat{\alpha}_2(t;b)$ is logarithmically consistent for $\alpha(b)$.*

Remark 2.7 *From Theorem 2.5 it follows straightforwardly that if X is a discrete-time sequence satisfying* **A1***, then $\hat{\alpha}_2(t;b)$ is logarithmically consistent for $\alpha(b)$ over the class of probability measures $\mathcal{M}_2$. The waiting time sequence associated with the single server queue typically satisfies the conditions of Theorem 2.5. For example, if the queue has i.i.d. inter-arrival times and i.i.d processing times (with finite moment generating function in a neighborhood of the origin and with the mean processing time strictly less than the mean of the inter-arrival time), then* **A1** *is generally satisfied, and $\mathbb{E}\tau_1^p < \infty$ for all $p \geq 1$, is guaranteed to occur; see, e.g., Gut [17].*

Hypothesis ii.) asserts that conditional on the process X hitting level b over a cycle, the process X spends a uniformly positive amount of time over that level within a cycle. This type of behavior is typical of queues whose service rate does not depend on the number of customers in the queue. In particular, in great generality one may expect a conditional weak convergence of the following form

$$\mathbb{P}\left(\int_0^{\tau_1} \mathbb{I}_{\{X(T(0)+s)>b\}} ds \in \cdot \big| \beta_1 > b \right) \Rightarrow \mathbb{P}(Z \in \cdot)$$

as $b \to \infty$, with Z a positive random variable. See Section 3 and 5 for related, and more explicit, computations. Related results on weak convergence of extremal statistics in regenerative processes can be found in [24], while almost sure results in the i.i.d. context are derived in [10]; see also [8, §3.5] for a recent survey.

We next turn to consideration of other models for the tail probability $\alpha(b)$. In particular, suppose that $\alpha(b)$ decays, roughly speaking, as a power of b.

$$\textbf{A2}. \qquad \frac{\log \alpha(b)}{\log b} \to -\theta^* \qquad as \;\; b \to \infty$$

for $0 < \theta^* < \infty$. In this setting, an argument similar to the one following **A1** suggests that we may expect

$$\frac{\log M(t)}{\log t} \to \frac{1}{\theta^*} \tag{2.9}$$

as $t \to \infty$. This suggests that in the presence of **A2**, we ought to consider the tail probability estimator

$$\hat{\alpha}_3(t;b) = b^{-(\log t / \log M(t))} \quad .$$

The logarithmic consistency of $\hat{\alpha}_3(t;b)$ as an estimator of $\alpha(b)$ requires precisely that we verify (2.9).

Theorem 2.8 *Suppose that $\mathbb{P} \in \mathcal{M}_2$ and that under the probability measure $\mathbb{P}$:*

i.) X satisfies **A2***;*

ii.) there exists $\epsilon > 0$ such that

$$\liminf_{b\to\infty} \mathbb{P}\left(\int_0^{\tau_1} \mathbb{I}_{\{X(T(0)+s)>b\}} ds \geq \epsilon \middle| \beta_1 > b\right) \geq \epsilon \quad .$$

Then,

$$\frac{\log M(t)}{\log t} \to \frac{1}{\theta^*} \quad \mathbb{P}-a.s.$$

as $t\to\infty$, and in L^p for any $p \in [1,\infty)$.

The proof is similar to that of Theorem 2.5 and is omitted for brevity.

Thus, Theorem 2.8 provides an a.s. extremal result for real-valued regenerative processes with a Pareto-like stationary marginal distribution. It covers, for example, i.i.d. sequences in discrete time having power law tails, and basically establishes the logarithmic consistency over $\mathcal{M}_2$ for discrete time regenerative processes. But, unfortunately, the result is not typically applicable to queueing problems. The difficulty is that a "fat tail" in the marginal stationary distribution of the queue goes hand-in-hand with a "fat tailed" cycle length, τ_1. Consequently, for a queue satisfying **A2**, it generally is false that $\mathbb{E}\tau_1^p < \infty$ for all $p \geq 1$. Thus, when **A2** holds in the queueing setting, $\mathbb{P}$ is typically not a member of $\mathcal{M}_2$. In fact, the conclusions of Theorem 3 do not generally hold in queues. Rather, one may expect that

$$\frac{\log M(t)}{\log t} \to \frac{1}{\theta^*+1} \quad a.s.$$

as $t\to\infty$. The reason for this behavior is that under suitable conditions on the queue (e.g., single-server queue with renewal input and sub-exponential processing times), the tail of $\beta_1 := \sup\{X(s) : s \in (0,\tau_1]\}$ is lighter in logarithmic scale than $\alpha(b)$; see Asmussen [3] for further details. In particular, when **A2** holds for the waiting time sequence marginal, evidently $\log \mathbb{P}(\beta_1 > b)/\log b \to -(\theta^*+1)$ as $b\to\infty$.

Example 2.9 *Consider, the GI/G/1 queue, when the associated random walk has increments X_i such that $\bar{F}(x) := \mathbb{P}(X_1 > x) \sim L(x)x^{-\theta^*}$ with $\theta^* > 1$ and where $L(x)$ is a slowly varying function.*

Then, it is well known that the stationary version of the waiting time sequence $(W_n^* : n \geq 0)$ *has*

$$\mathbb{P}(W_0^* > x) \sim \frac{1}{\mathbb{E}X_1} \int_x^\infty \bar{F}(y)dy$$

see, e.g., [3]. On the other hand, Heath, Samorodnitsky and Resnick [19] have shown recently that

$$\mathbb{P}(\beta_1 > x) \sim \mathbb{E}\tau_1 \bar{F}(x) \quad .$$

A simple calculation shows that the tail of the stationary marginal is one power "heavier" than that of the cycle maximum.

The previous discussion suggests that in the presence of such queueing structure, the estimator $\hat{\alpha_3}(t;b)$ should be modified to

$$\hat{\alpha}_4(t;b) = b^{1-(\log t/\log M(t))} \quad .$$

In the interests of completeness, we conclude this section with an a.s. extremal result in the context of Weibull-type stationary tails.

Theorem 2.10 *Suppose that* $\mathbb{P} \in \mathcal{M}_2$ *and that under the probability measure* $\mathbb{P}$*:*

i.) there exist positive finite constants γ *and* θ^* *such that*

$$\frac{\log \alpha(b)}{b^\gamma} \to -\theta^*$$

as $b \to \infty$*;*

ii.) there exists $\epsilon > 0$ *such that*

$$\liminf_{b\to\infty} \mathbb{P}\left(\int_0^{\tau_1} \mathbb{I}_{\{X(T(0)+s)>b\}} ds \geq \epsilon \middle| \beta_1 > b \right) \geq \epsilon \quad .$$

Then,

$$\frac{\log M(t)}{(\log t)^{1/\gamma}} \to \left(\frac{1}{\theta^*}\right)^{1/\gamma} \quad \mathbb{P}-a.s.$$

as $t \to \infty$*, and in* L^p *for any* $p \in [1, \infty)$*.*

The proof is similar to that of Theorem 2.5; the details are omitted.

From a queueing standpoint, the most obvious potential application of Theorem 2.10 is to the study of the workload process of a queue fed by fractional Brownian motion (fBM) input. This process has recently been proposed as a suitable model for network traffic in Ethernet and wide area networks (cf. [9] for further discussion). It is known that such queues have Weibull-type tails; see [25] and [7]. However, there is no obvious regenerative structure to which one can appeal in the fBM setting. Moreover, fBM exhibits long-range dependence which precludes "standard" mixing techniques. Other methods are therefore needed to study the extremal behavior of such queues; see Zeevi and Glynn [27] for details.

3 Rates of Convergence

The most important class of queueing models for which extremal estimators appear relevant is the class of processes for which the marginal tail probabilities satisfy **A1**. In this section, we consider rates of convergence for the extremal estimators of Section 2 in the setting of such exponential-type tail models.

We start by stating a general result for regenerative processes that can be expected to cover a broad class of queueing systems. Let $\mathcal{M}_3$ be the class of probabilities under which X is stationary and classically regenerative with regeneration times $0 \leq T(0) < T(1) < T(2) < \cdots$, with $\mathbb{E}\tau_1 < \infty$. In what follows we write $f(t) = \Theta(h(t))$ if $|f(t)/h(t)|$ is bounded above and below by finite positive constants, as $t \to \infty$.

Theorem 3.1 *Suppose that $\mathbb{P} \in \mathcal{M}_3$ and that under the probability measure $\mathbb{P}$:*

i.) $\log \alpha(b) = -\theta^* b + O(1)$ *as* $b \to \infty$ *for* $0 < \theta^* < \infty$*;*

ii.) $\mathbb{E}\left[\int_0^{\tau_1} \mathbb{I}_{\{X(T(0)+s)>b\}} ds | \beta_1 > b\right] = \Theta(1)$.

Then,

$$\frac{M(t)}{\log t} = \frac{1}{\theta^*} + O_p\left(\frac{1}{\log t}\right)$$

as $t \to \infty$.

Proof: The assumption combined with an application of the regenerative ratio formula easily yield

$$\log \mathbb{P}(\beta_1 > b) = -\theta^* b + O(1) \quad . \tag{3.1}$$

Now, recall that standard convergence of types theory for extremes asserts that for $\{X_i\}$ a sequence of i.i.d. r.v.'s with $\mathbb{P}(X_1 > x) \sim \eta \exp(-\theta^* x)$ one has

$$\theta^* \max_{1 \leq i \leq n} X_i - \log n - \log \eta \Rightarrow Z \quad ,$$

where Z has a Gumbel or type I distribution, i.e., $\mathbb{P}(Z \leq x) = \exp(-e^{-x})$, $x \in \mathbb{R}$. The "coarser" tail asymptotic that we assume in i.), resulting in (3.1), does not allow us to conclude a similar weak convergence for the scaled and translated maximum value, however a simple calculation shows that it is sufficient for the asserted rate of convergence in the i.i.d. context. The tail result in (3.1) together with an application of Lemma 1.1 in [3] conclude the proof for the regenerative case. ■

Remark 3.2 *Hypothesis i.) of Theorem 3.1 is known to hold in a wide variety of queueing settings. For example, the Cramér-Lundberg exact asymptotic guarantees i.) in the context of the stationary*

waiting time sequence of the single-server queue fed by renewal input; see [2, p. 269]. A similar exact asymptotic is known for the stationary workload of the same queue. As for ii.), the amount of time spent by a queue above level b, conditional on attaining level b, is typically $\Theta(1)$ *in b, due to the random walk structure implicit in queues. For a related calculation, see the proof of Theorem 7.5.1 of Glynn and Torres [13].*

Evidently, Theorem 3.1 implies that the rate of convergence at which the ratio $\log \hat{\alpha}_2(t;b)/\log \alpha(b)$ tends to one (i.e., the rate at which logarithmic consistency is attained) is very slow, namely $(\log t)^{-1}$ in the observed time horizon $[0,t]$. The slow rate is at least in part a consequence of the fact that we are demanding logarithmic consistency over an extremely broad class of models, namely all models satisfying the hypotheses of Theorem 2.5. An obvious competing estimator, which we now study in further detail, is $\hat{\alpha}_1(t;b)$, namely the empirical tail estimator.

We shall study this estimator in a specific model setting, namely that of a single-server queue with first-come first-serve queue discipline, i.i.d. inter-arrival times $(U_i : i \geq 1)$ and independent i.i.d. processing times $(V_i : i \geq 0)$. Let $\xi_i = V_i - U_{i+1}$ and assume that the ξ_i's are bounded, non-lattice r.v.'s for which there exists $\theta^* > 0$ satisfying the equation $\mathbb{E}\exp(\theta^*\xi_i) = 1$. To avoid trivialities, we impose $\mathbb{P}(\xi_i > 0) > 0$. Under the above conditions it is well known that there exists a stationary version of the waiting time sequence and $\alpha(b) = \mathbb{P}(W_0^* > b)$ satisfies the hypotheses of Theorem 3.1. A key to our analysis is the following compound Poisson limit theorem for the single server queue's waiting time sequence. For a heuristic explanation of this limit theorem see [1], while a more rigorous statement in the context of regenerative processes is given in [24, Theorem 2.6.1]. We note however that the conditions appearing in [24, Theorem 2.6.1] require verification, which in itself constitutes much of the proof below.

Proposition 3.3 *Suppose that* $n_b \uparrow \infty$ *so that* $n_b \exp(-\theta^* b) \to \eta$ *for some* $\eta \in [0,\infty)$ *as* $b \to \infty$. *Under the above conditions on* $(W_j^* : j \geq 0)$, *there exists a non-zero r.v. Z such that*

$$\sum_{j=0}^{n_b-1} \mathbb{I}_{\{W_j^* > b\}} \Rightarrow \mathcal{CP}(\lambda; \tilde{Z})$$

as $b \to \infty$. *Here* $\mathcal{CP}(\lambda;\tilde{Z})$ *is a compound Poisson r.v. that can be represented as* $\sum_{j=1}^{N} \tilde{Z}_j$ *with N a Poisson r.v. with mean* $\lambda = \eta C_\infty / \mathbb{E}\tau_1$ *(C_∞ is an explicit constant, that is identified explicitly) and* $\tilde{Z}_1, \tilde{Z}_2, \ldots$ *is an independent sequence of i.i.d copies of* $\tilde{Z}$. *The distribution of* $\tilde{Z}$, *the compounded r.v., is given in terms of its Laplace-Stieltjes transform in (3.7).*

Proof: Let $T(n) = \inf\{m > T_{n-1} : W_m = 0\}$ for $n \geq 1$, and $\tau_i := T_i - T_{i-1}$. Note that $T(0)$ is the duration of the first (delayed) cycle. Put $\ell(n) = \sup\{k \geq 0 : T(k) \leq n\}$, so that $\ell(n)$ counts

the number of completed regenerative cycles in $[0, n]$. Also, let $S_n = \sum_{i=1}^{n} \xi_i$ with $S_0 = 0$ be the random walk associated with W. In what follows we will use $n \equiv n_b$ to represent the sample size, suppressing the index b for clarity, where no confusion arises. The proof proceeds by first observing that

$$\sum_{i=1}^{\ell(n)-1} Z_i(b) \leq \sum_{j=0}^{n-1} \mathbb{I}_{\{W_j^* > b\}} \leq Z_0(b) + \sum_{i=1}^{\ell(n)+1} Z_i(b)$$

with $\{Z_i(b)\}_{i \geq 1}$ a sequence of independent copies of $Z_1(b) := \sum_{j=0}^{\tau_1 - 1} \mathbb{I}_{\{S_j > b\}}$, since W is identical to the random walk S until time $\tau := \tau_1$. Now, noting that $Z_0(b) \leq T(0)\mathbb{I}(\max\{W_j^* : 0 \leq j \leq T(0)\})$ is follows that $Z_0(b) \to 0$ a.s. as $b \to \infty$. The key therefore is to establish that $\sum_{i=1}^{\ell(n)} Z_j(b) \Rightarrow \mathcal{CP}(\lambda; \tilde{Z}$; a compound Poisson process. Establishing the latter, together with an application of the converging together lemma will conclude the proof. We now proceed to verify this asymptotic, and identify explicitly the rate λ and compounding distribution of $\tilde{Z}$ that together characterize the weak limit. We break the proof up into several steps.

1^0. The first step establishes that instead of the random time $\ell(n)$ one can consider $\ell(n) \approx n/\mathbb{E}\tau_1$. We now make this statement rigorous. Fix $\epsilon > 0$. Then on the event $A = \{\omega : |\ell_n(\omega)/n - 1/\mathbb{E}\tau_1| \leq \epsilon\}$ we have

$$\sum_{i=1}^{\lfloor n/\mathbb{E}\tau_1 - \epsilon n \rfloor} Z_i(b) \leq \sum_{i=1}^{\lfloor n/\mathbb{E}\tau_1 \rfloor} Z_i(b) + \mathcal{E}_n(b) \leq \sum_{i=1}^{\lceil n/\mathbb{E}\tau_1 + \epsilon n \rceil} Z_i(b) \tag{3.2}$$

with $\mathcal{E}_n(b)$ the error term arising from the approximation of ℓ_n on the event A. Fix $x \in \mathbb{R}_+$ so that

$$\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^{\ell(n)} Z_i(b) \leq x\right) = \\
&\quad = \mathbb{P}\left(\sum_{i=1}^{\ell(n)} Z_i(b) \leq x, \left|\frac{\ell(n)}{n} - \frac{1}{\mathbb{E}\tau_1}\right| \leq \epsilon\right) + \mathbb{P}\left(\sum_{i=1}^{\ell(n)} Z_i(b) \leq x, \left|\frac{\ell(n)}{n} - \frac{1}{\mathbb{E}\tau_1}\right| > \epsilon\right) \\
&\quad = \mathbb{P}\left(\sum_{i=1}^{\lfloor n/\mathbb{E}\tau_1 \rfloor} Z_i(b) + \mathcal{E}_n(b) \leq x, \left|\frac{\ell_n}{n} - \frac{1}{\mathbb{E}\tau_1}\right| \leq \epsilon\right) + o(1) \\
&\quad = \mathbb{P}\left(\sum_{i=1}^{\lfloor n/\mathbb{E}\tau_1 \rfloor} Z_i(b) + \mathcal{E}_n(b) \leq x\right) \\
&\qquad - \mathbb{P}\left(\sum_{i=1}^{\lfloor n/\mathbb{E}\tau_1 \rfloor} Z_i(b) + \mathcal{E}_n(b) \leq x, \left|\frac{\ell_n}{n} - \frac{1}{\mathbb{E}\tau_1}\right| > \epsilon\right) + o(1) \\
&\quad = \mathbb{P}\left(\sum_{i=1}^{\lfloor n/\mathbb{E}\tau_1 \rfloor} Z_i(b) + \mathcal{E}_n(b) \leq x\right) + o(1)
\end{aligned}$$

where $a_n = o(1)$ if $a_n \to 0$ as $n \to \infty$; these terms equaling $o(1)$ follows from the SLLN for renewal

processes. Thus, the problem is reduced to considering a deterministic number of summands, and a (random) error term.

2^0. This step derives the weak limit of $\sum_{i=1}^{\lfloor n/\mathbb{E}\tau_1 \rfloor} Z_i(b)$. Fix $s > 0$. Then,

$$\begin{aligned}
\varphi_b(s) &:= \mathbb{E}\exp\left(-s\sum_{i=1}^{\lfloor n/\mathbb{E}\tau_1\rfloor} Z_i(b)\right) \\
&= \left(\mathbb{E}[e^{-sZ(b)};\beta_1 > b] + 1 - \mathbb{P}(\beta_1 > b)]\right)^{\lfloor n/\mathbb{E}\tau_1\rfloor} \\
&= \left(1 + \mathbb{P}(\beta_1 > b)\left[\frac{\mathbb{E}[e^{-sZ(b)};\beta_1 > b]}{\mathbb{P}(\beta_1 > b)} - 1\right]\right)^{\lfloor n/\mathbb{E}\tau_1\rfloor} \qquad (3.3)
\end{aligned}$$

where $\beta_1 = \max(S_0, S_1, \ldots, S_{\tau_1 - 1})$. The main effort is to show that for each fixed $s > 0$

$$\mathbb{E}\left[e^{-sZ(b)};\beta_1 > b\right] \sim h(s)\exp(-\theta^* b) \quad . \qquad (3.4)$$

The reader may now notice that with (3.4), and (3.3) it is clear that $\varphi_b(s) \to \varphi(s)$ as $b \to \infty$, under the premise that $n_b \exp(-\theta^* b) \to \eta$. The limit $\varphi(s)$ will be identified in the sequel as the Laplace-Stieltjes transform of a compound Poisson distribution function.

The key step en route is to establish (3.4). We now define a change-of-measure on the paths of the random walk $(S_n : n \geq 0)$. To be precise, for $b > 0$, let $T(b) = \inf\{n \geq 0 : S_n \geq b\}$, and define the measure $\tilde{\mathbb{P}}(\cdot)$ via the identity

$$\tilde{\mathbb{E}}[\zeta; T(b) < \infty] := \mathbb{E}[\zeta \exp(\theta^* S_{T(b)}); T(b) < \infty]$$

for all non-negative r.v.'s ζ. It follows that

$$E[\zeta; T(b) < \infty] := \tilde{\mathbb{E}}[\zeta \exp(-\theta^* S_{T(b)}); T(b) < \infty] \quad .$$

Consequently,

$$\begin{aligned}
&\mathbb{E}\left[e^{-sZ(b)};\beta_1 > b\right] = \\
&\quad = \tilde{\mathbb{E}}\left[\exp\left(-s\sum_{j=0}^{\tau_1 - 1}\mathbb{I}_{\{S_j \geq b\}}\right)\exp(-\theta^* S_{T(b)}); T(b) < \tau_1, T(b) < \infty\right] \\
&\quad = \tilde{\mathbb{E}}\left[\exp\left(-s\sum_{j=T(b)}^{\tau_1 - 1}\mathbb{I}_{\{S_j \geq b\}}\right)\exp(-\theta^* S_{T(b)}); T(b) < \tau_1 < \infty\right] \\
&\quad = e^{-\theta^* b}\tilde{\mathbb{E}}\left[g(S_{T(b)} - b; b)e^{-\theta^*(S_{T(b)} - b)}; T(b) < \tau_1\right] \qquad (3.5)
\end{aligned}$$

with $g(x; b) := \mathbb{E}[\exp\{-s\sum_{j=0}^{T(-b)}\mathbb{I}_{\{S_j \geq 0\}}\}|S_0 = x]$, and $T(-b) = \inf\{n \geq 0 : S_n \leq -b\}$ for $-b \leq 0$. It is a well know fact that under the "twisted" distribution $\tilde{\mathbb{P}}(\xi_1 \in dx) = \exp(-\theta^* x)\mathbb{P}(\xi_1 \in dx)$ the

random walk will have positive "drift" for $0 \leq n \leq T(b)$, and thus $T(b) < \infty$ $\tilde{\mathbb{P}}$- a.s. Subsequently to crossing level b the random walk has (the original) negative drift again, under $\mathbb{P}$, thus $\tau_1 < \infty$ $\tilde{\mathbb{P}}$- a.s. Now, bounded convergence guarantees that

$$g(x;b) \downarrow g(x) := \mathbb{E}\left[\exp\left(-s\sum_{i=0}^{\infty}\mathbb{I}_{\{S_j \geq 0\}}\right)\big|S_0 = x\right]$$

as $b \to \infty$. In addition, $\mathbb{I}(T(b) < \tau_1) \downarrow \mathbb{I}(\tau_1 = \infty)$, and under the non-lattice hypothesis on ξ_i's the "overshoot" $\Psi(b) = S_{T(b)} - b \Rightarrow \Psi(\infty)$ as $b \to \infty$ (see, e.g., [2, p. 168]). Now, note that $\{\Psi(b)\}$ is a uniformly bounded family of r.v.'s since by assumption $\xi_i \leq K$ for some $K \in \mathbb{R}^+$. Also, since $g(x;b), g(x) \leq 1$ for all x, b, it follows that $g(x;b) \to g(x)$ uniformly on $[0, K]$, as $b \to \infty$. Combining these statements we observe that $|g(\Psi(b);b) - g(\Psi(b))| \to 0$ a.s. as $b \to \infty$. Thus, we can substitute $g(\Psi(b)$ for $g(\Psi(b);b)$ in (3.5). In addition, observe that

$$g(\Psi(b))\exp(-\theta^*\Psi(b)) \Rightarrow g(\Psi(\infty))\exp(-\theta^*\Psi(\infty))$$

since $g(\cdot)$ is non-decreasing and thus has at most a countable number of discontinuities, and since $\Psi(\infty)$ is a continuous r.v. (see, [2, p. 168]). Bounded convergence theorem and Stam's Lemma (cf. [2, p. 271]) are used to conclude that

$$\tilde{\mathbb{E}}\left[g(\Psi(b))\exp(-\theta^*\Psi(b)); T(b) < \tau_1\right] \to \tilde{\mathbb{E}}\left[g(\Psi(\infty))\exp(-\theta^*\Psi(\infty))\right]\tilde{\mathbb{P}}(\tau_1 = \infty)$$

as $b \to \infty$. Going back to (3.5) we conclude that

$$\mathbb{E}\left[e^{-sZ(b)}; \beta_1 > b\right] \sim \exp(-\theta^* b)h(s)$$

with

$$h(s) = \tilde{\mathbb{E}}\left[\mathbb{E}[\exp(-sZ(\infty))|S_0 = \Psi(\infty)]\exp(-\theta^*\Psi(\infty))\right]\tilde{\mathbb{P}}(\tau_1 = \infty) \tag{3.6}$$

and $Z(\infty) = \sum_{j=0}^{\infty}\mathbb{I}_{\{S_j \geq 0\}}$.

Now, for the GI/G/1 queue it is well known (cf. Iglehart [21]) that

$$\mathbb{P}(\beta_1 > b) \sim C_\infty \exp(-\theta^* b)$$

with $C_\infty := c\tilde{\mathbb{P}}(\tau_1 = \infty)$ and $c = \tilde{\mathbb{E}}\exp(-\theta^*\Psi(\infty))$. Then, by choice of $n_b \sim \eta\exp(\theta^* b)$ we have following (3.3) that

$$\varphi_b(s) \to \exp\left(-\frac{\eta C_\infty}{\mathbb{E}\tau_1}\left(1 - \frac{h(s)}{C_\infty}\right)\right)$$

as $b \to \infty$. Moreover, a closer inspection reveals that the limit is continuous from the right at $s = 0$, thus the limit is a Laplace transform of a bona fide distribution function, which by inspection

is compound Poisson with rate $\lambda = \eta C_\infty / \mathbb{E}\tau_1$ and compounding distribution that has Laplace transform

$$\begin{aligned} L(s) &= \frac{h(s)}{C_\infty} \\ &= \tilde{\mathbb{E}}\left[\mathbb{E}[\exp(-sZ(\infty))|S_0 = \Psi(\infty)]\right] \frac{\exp(-\theta^*\Psi(\infty))}{\tilde{\mathbb{E}}\exp(-\theta^*\Psi(\infty))} \end{aligned} \tag{3.7}$$

following the expression derived for $h(s)$ in (3.4).

3^0. We have just shown that

$$\sum_{i=1}^{\lfloor n/\mathbb{E}\tau_1 \rfloor} Z_i(b) \Rightarrow \mathcal{CP}(\lambda; \tilde{Z})$$

that is, the weak limit is compound Poisson with compounding random variable $\tilde{Z}$ characterized via its Laplace-Stieltjes transform $L(s)$. The same reasoning applies to the lower and upper bounds in (3.2). Combining steps 1^0 and 2^0, sending $\epsilon \downarrow 0$ and using the continuity of the distribution function $\mathcal{CP}(\cdot; \tilde{Z})$ gives

$$\sum_{j=0}^{n_b - 1} \mathbb{I}_{\{W_j^* > b\}} \Rightarrow \mathcal{CP}(\lambda; \tilde{Z})$$

which proves the result. ■

With Proposition 3.3 in hand, we can establish the following result.

Theorem 3.4 *Under the conditions of Proposition 3.3 we have the following:*

i.) if $n_b \exp(-\theta^* b) = o(1)$ *as* $b \to \infty$ *then*

$$n_b \hat{\alpha}_1(n_b; b) \Rightarrow 0$$

as $b \to \infty$*;*

ii.) if $n_b \exp(-\theta^* b) \to \eta \in (0, \infty)$ *as* $b \to \infty$ *then*

$$\frac{\log \hat{\alpha}_1(n_b; b)}{\log \alpha(b)} = 1 + O_p\left(\frac{1}{b}\right)$$

as $b \to \infty$*;*

iii.) if $n_b \exp(-\theta^* b) \to \infty$ *as* $b \to \infty$ *then there exists a constant* $r \in (0, \infty)$ *such that*

$$(n_b \exp(-\theta^* b))^{1/2} \left(\frac{\hat{\alpha}_1(n_b; b)}{\alpha(b)} - 1\right) \Rightarrow r^{1/2}\mathcal{N}(0, 1)$$

as $b \to \infty$*.*

Remark 3.5 Note that in case iii.) if $n_b \approx \exp((\theta^* + \epsilon)b)$ then

$$\frac{\log \hat{\alpha}_1(n_b; b)}{\log \alpha(b)} = 1 + O_p\left(e^{-\epsilon b/2}\right) \quad .$$

Thus, the critical growth rate, $\exp(\theta^* b)$, of the observation window n_b defines the breakdown of (log) consistency on the one hand, but on the other asserts that in the regime where the estimator is log-consistent, the rate of convergence is very rapid.

Proof: Parts i.) and ii.) of Theorem 3.4 follow immediately from Proposition 3.3. For part iii.) we appeal to Theorem 7.5.2. of Glynn and Torres [13]. ■

Theorem 3.4 asserts that $\alpha_1(n_b; b)$ is not logarithmically consistent for $\alpha(b)$ when $n_b = o(\exp(\theta^* b))$ as $b \to \infty$. In other words, if one is estimating the order of magnitude of a buffer overflow probability from an observed time horizon $[0, n_b]$ where n_b is not enormous (i.e., $n_b = o(\exp(\theta^* b))$), the extremal estimator is superior to the empirical tail estimator $\hat{\alpha}_1(n_b; b)$. On the other hand, if one has available an enormous amount of data (i.e., $n_b \gg \exp(\theta^* b)$), then the empirical estimator's superior convergence rate (see Theorems 3.1 and 3.4) suggests the use of $\hat{\alpha}_1(n_b; b)$ in preference to the extremal estimator $\hat{\alpha}_2(t; b)$.

At an intuitive level, the reason that the extremal estimator is better suited for "small sample sizes" has to do with the fact that it takes explicit advantage of the assumed tail behavior of the marginal distribution of X. In fact, one may consider this estimator semi-parametric in that respect, as opposed to the non-parametric counterpart. Moreover, the presence of a model for the tail probability allows us to extrapolate the loss probabilities out to buffer sizes b for which no losses have yet been observed.

The discussion above suggests an obvious extrapolation-based alternative to the extremal estimator, based not on the observed maximum but on the the observed empirical tail probability. In particular, note that for $\gamma > 0$, **A1** suggests that $\alpha(b) \approx \alpha(\gamma)^{b/\gamma}$ as $b \to \infty$. Hence, an alternative tail probability estimator under the model **A1** for the marginal tail is just

$$\hat{\alpha}_5(t; b) = \hat{\alpha}_1(t; \gamma(t))^{b/\gamma(t)} \quad ,$$

where $(\gamma(t) : t \geq 0)$ is non-decreasing deterministic function of the observed time horizon t.

Theorem 3.6 *Under the conditions of Theorem 3.1, $\hat{\alpha}_3(t; b)$ is logarithmically consistent for $\alpha(b)$ if $\gamma(t)$ is selected so that $\gamma(t) \uparrow \infty$ and $t \exp(-\theta^* \gamma(t)) \to \infty$.*

Proof: Suppose $b(t) \to \infty$ as $t \to \infty$. Then,

$$\begin{aligned}\frac{\log \hat{\alpha}_3(t; b(t))}{\log \alpha(b(t))} &= \frac{b(t)}{\gamma(t)} \frac{\hat{\alpha}_1(t; \gamma(t))}{-\theta^* b(t)(1 + o(1))} \\ &= \frac{\hat{\alpha}_1(t; \gamma(t))}{\log \alpha(\gamma(t))(1 + o(1))} \\ &\Rightarrow 1\end{aligned}$$

as $t \to \infty$, where we used part ii.) of Theorem 3.4 for the last step. ■

Remark 3.7 One obvious variation on the above theme is to fix a finite collection of positive real numbers $\{b_i\}$ and correspondingly form $\alpha_i = -(b_i)^{-1} \log(\hat{\alpha}_1(t; b))$. One can then form an estimator of the buffer overflow for buffer level b by judicious choice of the b_i's, and by taking an (possibly weighted) average of the α_i's. The behavior of this estimator will be qualitatively identical to that of the *extrapolation estimator* studied in Theorem 3.6. In particular, the convergence rate will be no better than that of the extrapolation estimator. However, it is possible that the above estimator may affect multiplicative constants in the convergence rate.

Because $\hat{\alpha}_1(t; \gamma(t))$ obeys a central limit theorem (CLT) with an associated convergence rate that is typically faster than that of $\hat{\alpha}_2(t; b)$, one expects that the extrapolation-based empirical tail estimator $\hat{\alpha}_3(t; b)$ will generally be superior to the extremal estimator $\hat{\alpha}_2(t; b)$ that we have proposed. It should be noted, however, that the extremal estimator $\alpha_2(t; b)$ is very easily implemented. In particular, it does not require the specification of "tuning parameters" like $\gamma(t)$.

We close this section with a brief description of a related, but different, estimation problem. Suppose that we are interested not in estimating steady state tail probabilities but in *predicting* the extremal behavior of the process X over a long time interval. In particular, we wish to estimate $\mathbb{P}(M(c) \leq \nu)$ for large c, based on observing X over $[0, t]$. When the process is regenerative, note that $\mathbb{P}(M(c) \leq \nu) = \mathbb{P}(\tilde{M}(c/\mathbb{E}\tau_1) \leq \nu) + o(1)$ as $c \to \infty$ uniformly in ν, where $\tilde{M}(t) := \max\{\beta_i : 1 \leq i \leq \lfloor t \rfloor\}$; see [3, Lemma 1]. Suppose we observe X over n cycles. Then, we may form the empirical distribution of the β_i's, namely

$$G_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{\beta_i \leq x\}} \quad .$$

Given that $\mathbb{P}(\tilde{M}(t) \leq \nu) = \mathbb{P}(\beta_1 \leq \nu)^{\lfloor t \rfloor}$, a natural estimator for $\mathbb{P}(M(c) \leq \nu)$ is therefore

$$G_n(\nu)^{\lfloor c/\bar{\tau}_n \rfloor}$$

where $\bar{\tau}_n := n^{-1} \sum_{i=1}^{n} \tau_i$.

The following result gives conditions under which the above estimator predicts the extremal behavior of X over $[0, c]$.

Proposition 3.8 *Suppose that X is a classically regenerative process for which $\mathbb{E}\tau_1^2 < \infty$, and let ν_n be given sequence of positive real numbers. If $c_n \to \infty$ in such a way that $\limsup_{n\to\infty} c_n \mathbb{P}(\beta_1 > \nu_n) < \infty$ and $c_n/n = o(1)$ as $n\to\infty$, then*

$$\frac{G_n(\nu_n)^{\lfloor c_n/\bar{\tau}_n \rfloor}}{\mathbb{P}(M(c_n) \leq \nu_n)} \Rightarrow 1$$

as $n\to\infty$.

Proof: Since $c_n \to \infty$, $\mathbb{P}(M(c_n) \leq \nu_n) - \mathbb{P}(\beta_1 > \nu_n)^{\lfloor c_n \mathbb{E}\tau_1 \rfloor} \to 0$ as $n\to\infty$. But

$$\begin{aligned}
\mathbb{P}(\beta_1 \leq \nu_n)^{\lfloor c_n/\mathbb{E}\tau_1 \rfloor} &= \exp\left(\lfloor c_n/\mathbb{E}\tau_1 \rfloor \log(1 - \mathbb{P}(\beta_1 > \nu_n))\right) \\
&= \exp\left(-\frac{c_n}{\mathbb{E}\tau_1}\mathbb{P}(\beta_1 > \nu_n) + o(1)\right) \quad ,
\end{aligned}$$

where we have used the fact that $\limsup_{n\to\infty} c_n \mathbb{P}(\beta_1 > \nu_n) < \infty$ in the last step.

Set $\bar{G}_n(x) = 1 - G_n(x)$, and $\bar{G}(x) = \mathbb{P}(\beta_1 > x)$, $G(x) = 1 - \bar{G}(x)$. Fix $\epsilon > 0$ and note that Chebychev's inequality implies that there exists $x = x(\epsilon)$ such that

$$\mathbb{P}\left(\left|\bar{G}_n(\nu_n) - \bar{G}(\nu_n)\right| > x\sqrt{\frac{\bar{G}(\nu_n)G(\nu_n)}{n}}\right) < \epsilon \quad .$$

On the event

$$A_n = \left\{\left|\bar{G}_n(\nu_n) - \bar{G}(\nu_n)\right| \leq x\sqrt{\frac{\bar{G}(\nu_n)G(\nu_n)}{n}}\right\} \quad ,$$

we have that

$$\begin{aligned}
\frac{G_n(\nu_n)^{\lfloor c_n/\bar{\tau}_n \rfloor}}{\mathbb{P}(M(c_n) \leq \nu_n)} &= \exp\left(\lfloor \frac{c_n}{\bar{\tau}_n} \rfloor \log(1 - \bar{G}_n(\nu_n)) + \frac{c_n}{\mathbb{E}\tau_1}\bar{G}(\nu_n) + o(1)\right) \\
&= \exp\left(-\lfloor \frac{c_n}{\bar{\tau}_n} \rfloor \left(\bar{G}_n(\nu_n) + O\left(\bar{G}(\nu_n) + x\sqrt{n^{-1}\bar{G}(\nu_n)}\right)^2\right) + \right. \\
&\quad \left. = \frac{c_n}{\mathbb{E}\tau_1}\bar{G}(\nu_n) + o(1)\right) \\
&= \exp\left(-\frac{c_n}{\mathbb{E}\tau_1}(1 + O_p(n^{-1/2}))\left(\left(\bar{G}(\nu_n)O\left(\sqrt{n^{-1}\bar{G}(\nu_n)}\right)\right)\right.\right. \\
&\quad \left.\left. - \frac{c_n}{\bar{\tau}_n}O(\bar{G}(\nu_n) + x\sqrt{n^{-1}\bar{G}(\nu_n)})^2\right) + \frac{c_n}{\mathbb{E}\tau_1}\bar{G}(\nu_n) + o(1)\right) \\
&= \exp\left(o_p(1) + o(1)\right) \quad ,
\end{aligned}$$

where the last step used the fact that $c_n\sqrt{\bar{G}(\nu_n)/n} = \sqrt{c_n\bar{G}(\nu_n)}\sqrt{c_n/n} = o(1)$. Consequently, for $\delta > 0$,

$$\mathbb{P}\left(\left|\frac{G_n(\nu_n)^{\lfloor c/\bar{\tau}_n \rfloor}}{\mathbb{P}(M(c_n) \leq \nu_n)} - 1\right| > \delta; A_n\right) \to 0$$

as $n \to \infty$. Letting $\epsilon \downarrow 0$ then establishes the result. ■

The most important consequence of Proposition 3.8 is that this non-parametric estimator does not give good relative accuracy for the tail probability of the extreme value $M(c)$ unless c is small relative to the observed time horizon n. In other words, to obtain good relative accuracy for tail probabilities of extreme values with c large, one must assume some structure on the tail of $M(c)$. The need to model tail structure is therefore a consistent theme of this paper. Berger and Whitt [4] adopted a similar philosophy in their effort to use extreme value limit theory to model the extreme value tails that they were attempting to predict.

4 First Passage Time Limit Theory for Regenerative Processes

In the previous section, we have studied issues related to the behavior of the maximum value taken on by a real-valued regenerative process over an expanding time horizon. In this section, we briefly describe the implications of these results for passage times. In particular, we study the first passage time

$$T(b) = \inf\{t \geq 0 : X(t) > b\}$$

to a level b, as $b \to \infty$. Because $\{M(t) < b\} = \{T(b) > t\}$, it is relatively straightforward to derive such results based on the limit theory we have already established for $M(t)$.

Our main result is the following.

Theorem 4.1 *Suppose that $\mathbb{P} \in \mathcal{M}_2$ and that there exists $\epsilon > 0$ such that*

$$\liminf_{b \to \infty} \mathbb{P}\left(\int_0^{\tau_1} \mathbb{I}_{\{X(T(0)+s) > b\}} ds \geq \epsilon \middle| \beta_1 > b \right) \geq \epsilon \quad .$$

Then,

i.) if $\log \alpha(b)/b \to -\theta^*$ for $0 < \theta^* < \infty$, then $\log T(b)/b \to \theta^*$ $\mathbb{P}$-a.s. as $b \to \infty$.

ii.) if $\log \alpha(b)/\log b \to -\theta^*$ for $0 < \theta^* < \infty$, then $\log T(b)/\log b \to \theta^*$ $\mathbb{P}$-a.s. as $b \to \infty$.

iii.) if $\log \alpha(b)/b^\gamma \to -\theta^*$ for $0 < \gamma, \theta^* < \infty$, then $\log T(b)/b^\gamma \to \theta^*$ $\mathbb{P}$-a.s. as $b \to \infty$.

Proof: Note that Theorem 2.5 proves that for $\epsilon > 0$, there exists $x_1 = x_1(\epsilon)$ such that for $t \geq x_1$, $M(t) < (1/\theta^* + \epsilon) \log t$. But, $T(b) \to \infty$ a.s., so for $T(b) > x_1$, $b \leq M(T(b)) < (1/\theta^* + \epsilon) \log T(b)$, and hence $(1/\theta^* + \epsilon)^{-1} < \log T(b)/b$ a.s. for b sufficiently large. On the other hand, there exists $x_2 = x_2(\epsilon)$ such that $M(t) > (1/\theta^* - \epsilon) \log t$ a.s. for $t \geq x_2$. Then, for $T(b) > x_2$, $b \geq M(T(b)^-) > (1/\theta^* - \epsilon) \log T(b)$ a.s. and consequently, $(1/\theta^* - \epsilon)^{-1} > \log T(b)/b$ for b sufficiently large. Since this

holds for arbitrary $\epsilon > 0$, i.) is proved. The proofs of ii.) and iii.) are similar, and take advantage of Theorems 2.8 and 2.10, in place of Theorem 2.5. ■

Remark 4.2 *While a large class of stationary stochastic processes are classically regenerative, this assumption in some sense limits the applicability of the theory developed in this paper to general state space Markov chains and processes. In particular, it is known that the class of Harris recurrent Markov chains and processes are one-dependent regenerative, in the sense that there exists a sequence of random times* $(T(n) : n \geq 0)$ *such that* $0 \leq T(0) < T(1) < T(2) < \cdots$ *for which the cycles* $\{(X(T(j-1)+s) : 0 \leq s < T(j) - T(j-1)), T(j) - T(j-1)\}$ *are identically distributed and one-dependent for* $j \geq 1$*; see [26] for details. However, note that*

$$M(T(2n)) = \max \left\{ \bigvee_{j=0}^{n} \beta_{2j}, \bigvee_{j=0}^{n} \beta_{2j+1} \right\} \quad ,$$

where $\beta_k = \max\{X(s) : T(k-1) \leq s < T(k)\}$*. The theory developed in this paper can be directly applied on an individual basis to the* $\{\beta_{2j}\}$ *and* $\{\beta_{2j+1}\}$ *sequences, and their maxima, since each of them is respectively a sequence of i.i.d. r.v.'s. Given the normalizations that are present in our limit theory, a simple sample path argument extends the conclusions of Theorem 2.5 through 2.10 and Theorem 4.1 to the one-dependent setting. Consequently, the theory applies to a broad class of Harris chains and processes.*

Glasserman and Kou [11, Theorem 1.1] prove a result similar to part i.) of Theorem 4.1. However, their theorem imposes the hypothesis directly on the tail of the r.v. β, the maximum of X over a regenerative cycle. On the other hand, our results place the assumptions directly on the tail of the $X(t)$ itself. It should be noted that the theory presented in Glynn and Whitt [15] and Duffield and O'Connell [7] establishes **A1** for the tail of X and not for the r.v. β, making our limit results easier to apply to the examples presented in those papers. The key condition that guarantees that the tails of β and $X(t)$ are equivalent in the sense that $\log \mathbb{P}(X(t) > b)/\log \mathbb{P}(\beta_1 > b) \rightarrow 1$ as $b \rightarrow \infty$ is the fact that the class $\mathcal{M}_2$ requires that $\mathbb{E}\tau_1^p < \infty$ for all $p \geq 1$. It turns out that our assumption that all the cycle moments are finite is, in some sense, a sharp result. Our next examples show that if $\mathbb{E}\tau_1^p = \infty$ for some $p \geq 1$, then there exists a stationary classically regenerative process X for which $\log \mathbb{P}(X(t) > b)/\log \mathbb{P}(\beta_1 > b)$ does not converge to one as $b \rightarrow \infty$. This, in turn, provides a counterexample to the conclusions of Theorems 2.5 through 2.10 as well as Theorem 4.1.

Example 4.3 *We restrict attention to discrete time; an obvious modification will give rise to a continuous time regenerative process with paths that are right-continuous with left limits. Suppose there exists a positive real number* $q \geq 1$ *for which* $\mathbb{E}\tau_1^q = \infty$*. Set* $T(0) = 0$*. Draw* $M_1 \sim Exp(1)$*, and conditional on* M_1 *set* $\tau_1 = \exp(M_1/q)$*, i.e., a point mass at* $\exp(y/q)$ *conditional on* $M_1 = y$*.*

Let $T(1) = T(0) + \tau_1$. Set $X_0 = 0$, and put $(X_n : 1 \leq n < \tau)$ equal to M_1 and set $X_{T(1)} = 0$. Repeat this construction to generate the remaining cycles, with M_k taken each time as an independent copy of an $Exp(1)$ random variable, and $T(k) = T(k-1) + \tau_k$, and $\tau_k = \exp(M_k/q)$. Clearly the resulting process is regenerative, with regeneration set equal to $\{0\}$. Moreover, $\mathbb{E}\tau^p < \infty$ for all $p < q$ and diverges for the qth power. Now, since this process is classically regenerative aperiodic, with $\mathbb{E}\tau_1 < \infty$ then it follows that a stationary version of X, say $X^* = (X_n^* : n \geq 0)$ exists, with $X_n^* \stackrel{\mathcal{D}}{=} X_\infty$, where the distribution of X_∞ is given by the regenerative ratio formula (cf. Asumssen, [2] for details). Specializing this argument, the tails of X_∞ are found to be

$$\begin{aligned}
\mathbb{P}(X_\infty \geq x) &:= \frac{1}{\mathbb{E}\tau} \int_{y=x}^{\infty} \mathbb{E}\left[\sum_{i=0}^{\tau-1} \mathbb{I}_{\{X_i \geq x\}} | M = y\right] P_M(dy) \\
&= \frac{1}{\mathbb{E}\tau} \int_{y=x}^{\infty} \mathbb{E}\left[\tau | M = y\right] e^{-y} dy \\
&= \frac{1}{\mathbb{E}\tau} \int_{y=x}^{\infty} e^{y/q} e^{-y} dy \\
&= \frac{1}{(1-1/q)\mathbb{E}\tau} e^{-(1-1/q)x}
\end{aligned}$$

On the other hand, it is evident that

$$\mathbb{P}(\beta_1 > x) = e^{-x}$$

with $\beta_1 = \max\{X_0, X_1, \ldots, X_{\tau_1 - 1}\}$. Clearly,

$$\lim_{b \to \infty} \frac{\log \mathbb{P}(X_\infty \geq x)}{\log \mathbb{P}(\beta_1 \geq x)} \to 1 - \frac{1}{q} \quad .$$

A more striking "mismatch" of the tails can also occur. Suppose that we only have $\mathbb{E}\tau < \infty$. Fix $p > 2$. Then we can repeat the above construction of X but conditional on M we set $\tau = \exp(M)/M^p$. Then, it is clear that $\mathbb{P}(X_\infty > x) = c/x^{p-1}$, i.e., X has a heavy tailed distribution for its stationary marginal, while β_1 has "light" (exponential) tails. Note, that for the process X we have $\mathbb{E}\tau < \infty$ but $\mathbb{E}\tau^{1+\delta} = \infty$ for all $\delta > 0$.

It should further be noted that in the absence of the hypothesis

$$\mathbb{P}\left(\int_0^{\tau_1} \mathbb{I}_{\{X(T(0)+s) > b\}} ds \geq \epsilon \big| \beta_1 > b\right) \geq \epsilon \quad ,$$

it is easy to construct examples in which $\log \mathbb{P}(X(t) > b) / \log \mathbb{P}(\beta_1 > b)$ does not converge to one, showing that a condition like this is, in some sense, also necessary.

5 An Illustrative Example: Reflecting Brownian Motion

Reflecting Brownian motion (RBM) is an approximation to the single-server queue that has a long history. In particular, RBM provides an approximation to the dynamics of such a queue that

depends only on the mean and the variance characteristics of the input processes to the queue, and it is guaranteed to be asymptotically accurate as the server utilization converges to one; see, for example, Glynn [16]. As such, the behavior of RBM can be expected to be representative of many queues.

Stationary RBM is the process W^* obtained when Γ is a Brownian motion with drift less than one. Denoting stationary RBM by $X = (X(t) : t \geq 0)$, the distribution of X is completely determined by the drift $-\mu < 0$ of the "free process" $\Gamma(t) - t$ and its Brownian variance parameter σ^2. Specifically, the stationary marginal is exponential; $\alpha(b) = \mathbb{P}(X(t) > b) = \exp(-\theta^* b)$ with $\theta^* := 2\mu/\sigma^2$.

Because of the analytical tractability of RBM, we can easily compute the explicit distribution of cycle maximum β_1; to avoid cycles of zero length, a regenerative cycle is defined as the path traced out by X that starts from zero, hits level 1, and subsequently returns to 0, thereby completing the cycle. A simple computation verifies that the tail of the distribution of the cycle-maximum is asymptotically

$$\mathbb{P}(\beta_1 > b) \sim (e^(\theta^*) - 1) \exp(-\theta^* b) \quad . \tag{5.1}$$

Extreme value theory establishes that

$$\max_{0 \leq s \leq t} X(s) - \frac{1}{\theta^*} \log t \Rightarrow \frac{1}{\theta^*} \left[Z + \log(e^{\theta^*} - 1) \right] \tag{5.2}$$

where Z is a Gumbel r.v. with distribution $\mathbb{P}(Z \leq x) = \exp(-e^{-x})$ for $x \in \mathbb{R}$ (see [4] for details). Our Theorem 2.5 proves that in fact

$$\frac{M(t)}{\log t} \to \frac{1}{\theta^*} \quad a.s. \tag{5.3}$$

as $t \to \infty$, where $M(t) := \sup\{X(s) : s \in (0, t]\}$. Note that here the explicit distribution of the cycle maximum is available, and is equal up to a constant with the tail of the stationary distribution. The limit theorem (5.2) proves that in this example, the rate of convergence in (5.3) is roughly of order $(\log t)^{-1}$. In other words, the rate of convergence is as predicted in Section 3.

The asymptotics of the empirical tail estimator for RBM can also be computed, and the performance of this estimator can then be contrasted with the extremal based estimator. Specifically, if $t_b \exp(-\theta^* b) \to \infty$ as $b \to \infty$, then

$$\sqrt{t_b \exp(-\theta^* b)} \left(\frac{\hat{\alpha}_1(t_b; b)}{\alpha(b)} - 1 \right) \Rightarrow \sqrt{2}(\sigma/\mu)\mathcal{N}(0, 1)$$

as $b \to \infty$; see [13, Theorem 7.3.2] for this calculation. The above limit theorem provides the exact value of the constant r appearing in part ii.) of Theorem 3.1. Consequently, if $t_b \sim \exp((\theta^* + \epsilon)b)$ for some $\epsilon > 0$, then

$$\frac{\log \hat{\alpha}_1(n_b; b)}{\log \alpha(b)} = 1 + O_p\left(e^{-\epsilon b/2}\right)$$

analogously to the result obtained for the GI/G/1 queue following Theorem 3.4.

The case of critical, and sub-critical growth of the time window $[0, t_b]$ is covered by the analogue of Proposition 3.3, which asserts that the empirical tail estimator is not logarithmically consistent.

Proposition 5.1 *Let X be a stationary version of RBM with infinitesimal drift $-\mu < 0$ and infinitesimal variance $\sigma^2 > 0$, and set $\theta^* := 2\mu/\sigma^2$. Let t_b be a sequence of positive real numbers such that $t_b \exp(-\theta^* b) \to \eta \in [0, \infty)$. Then,*

$$\int_0^{t_b} \mathbb{I}_{\{X(s)>b\}} ds \Rightarrow \mathcal{CP}(\lambda; Z)$$

as $b \to \infty$, where the weak limit can be expressed as $\sum_{j=1}^{N} Z_j$; N is Poisson with parameter $\lambda = \eta\theta^*\mu$ and $\{Z_j\}$ is an independent sequence of i.i.d copies of Z whose distribution is given implicitly in terms of its Laplace transform $\mathbb{E}\exp(-sZ) = 2(\sqrt{1 + 2s\sigma^2/\mu^2} + 1)^{-1}$.

Proof sketch: Using the regenerative structure of RBM, we can follow step 1^0 of the proof of Proposition 3.3 to reduce the problem to the study of the asymptotics of $\sum_{i=1}^{\lfloor t/\mathbb{E}\tau_1 \rfloor} Z_i(b)$, where $(Z_i : i \geq 1)$ is a sequence of i.i.d. copies of $Z_0(b) \stackrel{\mathcal{D}}{=} \int_0^{\tau_1} \mathbb{I}_{\{W(s)>b\}} ds$. Note that $X = W$ for $t \in [0, \tau_1]$ where $W(t) = -\mu t + \sigma B(t)$ is the corresponding "free process", i.e., the underlying negative drift Brownian motion. The key then is to establish an analogue of (3.4). Fix $s > 0$ and let $T(x) = \inf\{t \geq 0 : W(t) = x\}$ for $x \in \mathbb{R}$. Then,

$$\begin{aligned}
&\mathbb{E}[\exp(-sZ_1(b)); T(b) < \tau_1] = \\
&\quad = \mathbb{E}\left[\exp\left(-s\int_{T(b)}^{\tau_1} \mathbb{I}_{\{W(s)>b\}} ds\right); T(b) < \tau_1\right] \\
&\quad = \mathbb{E}\left(\mathbb{I}_{\{T(b)<\tau_1\}} \mathbb{E}\left[\exp\left(-s\int_{T(b)}^{\tau_1} \mathbb{I}_{\{W(s)>b\}} ds\right) | W(0) = b\right]\right) \\
&\quad = \mathbb{E}\left(\mathbb{I}_{\{T(b)<\tau_1\}} \mathbb{E}\left[\exp\left(-s\int_{0}^{T(-b)} \mathbb{I}_{\{W(s)>0\}} ds\right)\right]\right) \\
&\quad = \mathbb{P}(T(b) < \tau_1)\mathbb{E}\left[\exp\left(-s\int_{0}^{T(-b)} \mathbb{I}_{\{W(s)>0\}} ds\right)\right]
\end{aligned}$$

where we have used the strong Markov property and time-homogeneity. Since $T(-b) \uparrow \infty$ a.s. as $b \to \infty$, we can use bounded convergence to conclude that

$$\begin{aligned}
h(s) &= \mathbb{E}\left[\exp\left(-s\int_0^{\infty} \mathbb{I}_{\{W(s)>0\}} ds\right)\right] \\
&= 2(\sqrt{1 + 2s\sigma^2/\mu^2} + 1)^{-1}
\end{aligned} \tag{5.4}$$

where the second equality follows from [1, p. 72], as $L(s)$ is clearly seen to be the Laplace-Stieltjes transform of the total time spent positive by negative drift BM. As in Step 2^0 (3.3) of the proof of

Proposition 3.3, we can use (5.4) above to deduce that

$$\phi_b(s) \to \exp\left(-\lambda(1-h(s))\right)$$

as $b \to \infty$. Moreover, it is not difficult to identify the rate of the Poisson r.v. explicitly $\lambda = \eta(\exp(\theta^*)-1)/\mathbb{E}\tau_1$. Basic martingale arguments establish that $\mathbb{E}\tau_1 = (\exp(\theta^*)-1)/(\theta^*\mu)$ and thus we have finally $\lambda = \eta\theta^*\mu$. Having identified the distribution of the compounding r.v. and the rate of the Poisson r.v., we have the complete characterization of the weak limit $\mathcal{CP}(\lambda; Z)$. ■

Acknowledgements. The authors would like to thank an anonymous referee for helpful comments on the manuscript.

References

[1] Aldous, D. *Probability Approximations via the Poisson Clumping Heuristic.* Applied Mathematical Sciences, 77. Springer-Verlag, New York-Berlin, 1989.

[2] Asmussen, S. *Applied Probability and Queues.* Wiley, NY, 1987.

[3] Asmussen, S. *Extreme value theory for queues via cycle maxima.* Extremes, **1** (1998), 137-168.

[4] Berger, A.W. and Whitt, W. *Maximum values in queueing processes.* Prob. Engrg. and Info. Sci., **9** (1995), 375–409.

[5] Chang, C.-S. *Sample path large deviations and intree networks.* Queueing Systems Theory Appl., **20** (1995), 7–36.

[6] Courcoubetis, C., Kesidis, G., Ridder, A., Walrand, J. and Weber, R. *Admission control and routing in ATM networks using inference from measured buffer occupancy.* IEEE Trans. on Comm., **43** (1995), 1778–1784.

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

[8] Embrechts, P., and Klüppelberg, t. and Mikosch, C. *Modeling Extremal Events.* Springer-Verlag, Berlin, 1997.

[9] Erramilli, A., Narayan, O. and Willinger, W. Fractal queueing models. In *Frontiers in queueing*, Probab. Stochastics Ser., pages 245–269. CRC, Boca Raton, FL, 1997.

[10] Galambos, J. *The Asymptotic Theory of Extreme Order Statistics.* Wiley, NY, 1978.

[11] Glasserman, P. and Kou, S. *Limits of first passage times to rare sets in regenerative processes.* Ann. Appl. Probab., **5** (1995), 424–445.

[12] Glynn, P. W. and Sigman, K. *Uniform Cesàro limit theorems for synchronous processes with applications to queues.* Stochastic Process. Appl. **40** (1992), 29–43.

[13] Glynn, P.W. and Torres, M. *Nonparametric estimation of tail probabilities for the single-server queue.* In Stochastic Networks, 109–138, Lecture Notes in Statist., 117, Springer, New York, 1996.

[14] Glynn, P. W., Torres, M. *Parametric estimation of tail probabilities for the single-server queue.* In Frontiers in Queueing, 449–462, Probab. Stochastics Ser., CRC, Boca Raton, FL, 1997.

[15] Glynn, P.W. and Whitt, W. *Logarithmic asymptotics for steady-state tail probabilities in a single-server queue.* J. Appl. Probab., **31A** (1994), 131–156.

[16] Glynn, P.W. *Diffusion approximations.* In Stochastic Models: Handbooks of OR & MS, volume 2. Elsevier Science Publishers, 1990.

[17] Gut, A. *Stopped random walks. Limit theorems and applications.* Applied Probability. A Series of the Applied Probability Trust, 5. Springer-Verlag, New York-Berlin, 1988.

[18] M.J. Harrison. *Brownian Motion and Stochastic Flow Systems.* Wiley, NY, 1985.

[19] Heath, D., Resnick, S. and Samorodnitsky, G. *Patterns of buffer overflow in a class of queues with long memory in the input stream.* Ann. Appl. Prob., **7** (1997), 1021–1057.

[20] Hsu, I. and Walrand, J. *Dynamic bandwidth allocation for ATM switches.* J. Appl. Probab., **33** (1996), 758–771.

[21] Iglehart, D. L. *Extreme values in the GI/G/1 queue.* Ann. Math. Statist., **43** (1972), 627–635.

[22] Konstantopoulos, T., Zazanis, M. and De Veciana, G. *Conservation laws and reflection mappings with an application to multiclass mean value analysis for stochastic fluid queues.* Stochastic Processes and their Applications, **65** (1996), 139–146.

[23] Leadbetter, M.R., Lindgren, G., and Rootzén, H. *Extremes and Related Properties of Random Sequences and Processes.* Springer-Verlag, New York, 1983.

[24] Leadbetter, M.R. and Rootzén, H. *Extremal theory for stochastic processes.* Ann. Probab., **16** (1988), 431–478.

[25] Norros, I. *A storage model with self-similar input.* Queueing Systems, **16** (1994), 387–396.

[26] Sigman, K. and Wolff, R. *A review of regenerative processes.* SIAM Rev. **35** (1993), 269–288.

[27] Zeevi, A.J. and Glynn, P.W. *On the maximum workload in a queue fed by fractional Brownian motion.* accepted for publication in the Ann. Appl. Probab., 1999.