Logarithmic Asymptotics for the Supremum of a Stochastic Process

Ken Duffy[4], John T. Lewis and Wayne G. Sullivan.
Dublin Institute of Technology.
20/06/2002

Abstract

Logarithmic asymptotics are proved for the tail of the supremum of a stochastic process, under the assumption that the process satisfies a restricted large deviation principle on regularly varying scales. The formula for the rate of decay of the tail of the supremum, in terms of the underlying rate–function, agrees with that stated by Duffield and O'Connell [4]. The rate–function of the process is not assumed to be convex. A number of queueing examples are presented which include applications to Gaussian processes and Weibull sojourn sources.

1. Introduction. Let $\{W_t\}$ be a stochastic process. Define $Q := \sup_{t\geq 0} W_t$. We investigate the tail–asymptotics of $\mathbb{P}[Q > q]$ as q becomes large. Assume that the process $\{W_t\}$ satisfies a restricted form of the Large Deviation Principle (LDP): for some scaling functions a, v regularly varying with indices $A > 0, V > 0$, respectively, the limit

$$\lim_{t\to\infty} v(t)^{-1} \log \mathbb{P}[a(t)^{-1} W_t > c] = -J(c) \tag{1.1}$$

exists for all $c \geq 0$; under additional technical assumptions on $\{W_t\}$, we prove that

$$\lim_{q\to\infty} h(q)^{-1} \log \mathbb{P}[Q > q] = -\inf_{c>0} c^V J(c^{-A}) \tag{1.2}$$

with $h = v \circ a^{-1}$.

It is conceivable that the tail–asymptotics of $\mathbb{P}[a(t)^{-1}W_t > c]$ for a single t–value could dominate those of $\mathbb{P}[Q > q]$. We show by example that this can happen. Without an additional assumption, (1.1) can hold while (1.2) fails. To exclude this possibility, we introduce a uniform individual decay–rate hypothesis: there exist constants $F > V/A$, $K > 0$, such that for all t and all $c > K$,

$$\frac{1}{v(t)} \log \mathbb{P}[W_t > ca(t)] \leq -c^F.$$

[1]American Mathematical Society 1991 subject classifications. Primary 60F10. Secondary 60K25.
[2]Keywords and phrases: Extrema of Stochastic Process, Large Deviation Theory, Queueing Theory
[3]Work supported by Science Foundation Ireland under the National Development Plan
[4]K.D's work supported in part by a grant from Microsoft Research, Cambridge.

The novel aspects of our treatment of the tail–asymptotics of $\mathbb{P}[Q > q]$ are:

- the use of a weakened form of the LDP in which a non–convex rate–function J is allowed;
- the introduction of the uniform individual decay–rate hypothesis.

Our main results are proved in Section 2. Examples are given in Section 3. The connections with the existing literature are discussed in Section 4. We conclude this section with a brief description of our strategy:

We assume that a restricted LDP of the form (1.1) is satisfied by the pair $(W_t/a(t), v(t))$ but do not make assumptions about how it has been deduced; in particular, we admit non–convex rate–functions. We base our estimates directly on the probabilities, making clear their interpretation, and we provide a simple form for the resulting asymptotic rate of decay. With these assumptions, it is not difficult to prove the lower bound,

$$\liminf_{q\to\infty} \frac{1}{h(q)} \log \mathbb{P}[Q > q] \geq - \inf_{c>0} c^V J\left(\frac{1}{c^A}\right) =: -\delta; \tag{1.3}$$

details are provided in Section 2.

More work is required, however, in proving the corresponding upper bound. This begins by splitting the set $\{Q > q\}$ into the union of three, not necessarily disjoint, subsets:

$$\{Q > q\} = \{\sup_{t:a(t)<q\underline{c}} W_t > q\} \cup \{\sup_{t:q\underline{c}\leq a(t)\leq q\overline{c}} W_t > q\} \cup \{\sup_{t:a(t)>q\overline{c}} W_t > q\}, \tag{1.4}$$

where $0 < \underline{c} < \overline{c} < \infty$. The principle of the largest term ensures that the rate of decay of the probability of $\{Q > q\}$ is less than or equal to the slowest rate of decay of these three sets. It will be shown in Proposition 2.3. that the rate of decay of the probability of the middle term in (1.4), for any $0 < \underline{c} < \overline{c} < \infty$, is bounded above by $-\delta$. It is shown in Proposition 2.2. that there exists a $\overline{c} < \infty$ such that the rate of decay of the probability of the final term in (1.4) is bounded above by $-\delta$.

The restricted LDP hypothesis refers to the limiting behavior of $\log \mathbb{P}[W_t > ca(t)]$. The asymptotics of a single W_t could dominate those of Q. We need to impose an additional condition which excludes this possibility, ensuring that there exists a $\underline{c} > 0$ such that the rate of decay of the probability of the first term in (1.4) is as fast as $-\delta$. We then have that

$$\lim_{q\to\infty} \frac{1}{h(q))} \log \mathbb{P}[Q > q] = - \inf_{c>0} c^V J\left(\frac{1}{c^A}\right). \tag{1.5}$$

2. Main Results. We consider a family of random variables $\{W_t : t \in T\}$ where T is an unbounded subset of $\mathbb{R}_+$. In this work we shall be primarily interested in $T = \mathbb{Z}_+$ but provide an additional hypothesis under which the work extends to $T = \mathbb{R}_+$. We define $Q := \sup_{t\geq 0} W_t$.

Recall from Bingham *et. al.* [2] the definition of a regularly varying function:

DEFINITION 1 *A strictly positive measurable function $f : \mathbb{R}_+ \to \mathbb{R}_+$ is regularly varying of index $\rho \neq 0$ if, for all $c > 0$,*

$$\lim_{t\to\infty} \frac{f(ct)}{f(t)} = c^\rho.$$

Scaling hypothesis: The function a is continuous, strictly increasing, regularly varying with index $A > 0$ and $\lim_{t\to\infty} a(t) = \infty$. The measurable function v is regularly varying with index $V > 0$ and $\lim_{t\to\infty} v(t) = \infty$.

Note that, although we have assumed that $a(t)$ is both strictly increasing and continuous, given any function $a(t)$ which is regularly varying with index $A > 0$, it is possible to construct a function $a'(t)$ which is both strictly increasing and continuous, so that

$$\lim_{t\to\infty} \frac{a(ct)}{a(t)} = \lim_{t\to\infty} \frac{a'(ct)}{a'(t)} = \lim_{t\to\infty} \frac{a'(ct)}{a(t)} = c^A,$$

for all $c > 0$.

Also note that in Loynes' original work [10], where the distribution $Q := \sup_{t\geq 0} W_t$ was introduced in the queueing context, $\{W_t : t \in \mathbb{Z}_+\}$ is defined to be a process with stationary increments $\{Z_t\}$; that is, $W_t := \sum_{i=-t}^{-1} Z_i$ and $W_0 := 0$. In this setting, if Z_1 is integrable, the individual ergodic theorem (see page 18 of Halmos [9]) holds, so that

$$\lim_{t\to\infty} \frac{1}{t} \sum_{i=1}^{t} Z_{-i} = \lim_{t\to\infty} \frac{W_t}{t} = \mathbb{E}[Z_1 \,|\mathcal{F}],$$

where $\mathcal{F}$ is the invariant $\sigma-$algebra. Hence, in this situation, it seems likely that $a(t)$ would be set to be t; if $a'(t)$ is any other scale such that $\lim a'(t)/t \in \{0, \infty\}$, then the information about the mean behavior of the arrivals less service is lost on the scale $a'(t)$.

If $a(t)$ is regularly varying with constant A, then Theorem 1.5.12 of [2] proves that $a^{-1}(t)$ is regularly varying with constant $1/A$.

DEFINITION 2 *We define the scaling function* $h : \mathbb{R}_+ \to \mathbb{R}_+$ *by*

$$h(q) := v(a^{-1}(q)). \tag{2.6}$$

From chapter 1 of [2], in particular Theorems 1.4.1 and 1.5.6, we deduce the following:

LEMMA 2.1. *Under the scaling hypothesis the function* h *defined by (2.6) satisfies for each* $c > 0$

$$\lim_{t\to\infty} \frac{h(ct)}{h(t)} = c^{V/A}.$$

As $t \to \infty$ *the ratio* $h(ct)/h(t)$ *converges to* $c^{V/A}$ *uniformly as a function of* c *on compact subsets of* $\mathbb{R}_+$. *Similarly, the ratios* $v(ct)/v(t)$ *and* $a(ct)/a(t)$ *converge uniformly to* c^V *and* c^A *as functions of* c *on compact subsets of* $\mathbb{R}_+$. *For each* $\delta > 0$ *and* $C > 1$ *there exists* $b_{\delta,C}$ *so that* $q, t \geq b_{\delta,C}$ *implies*

$$h(t)/h(q) \leq C \max\{(t/q)^{V/A+\delta}, (t/q)^{V/A-\delta}\}.$$

Analogous inequalities hold for $v(t)/v(q)$ *and* $a(t)/a(q)$.

For each $x \in \mathbb{R}$, define $\lceil x \rceil$ to be the least integer greater than x. The proof of the next Lemma is elementary and we omit it:

LEMMA 2.2. *Under the scaling hypothesis, for all* $c > 0$,

$$\lim_{t\to\infty} \frac{v(\lceil a^{-1}(ct)\rceil)}{v(a^{-1}(t))} = c^{V/A}.$$

LEMMA 2.3. *Under the scaling hypothesis, for each* $\gamma > 0$,

$$\lim_{n\to\infty} \frac{1}{-\gamma v(n)} \log \sum_{k=n}^{\infty} e^{-\gamma v(k)} = 1. \tag{2.7}$$

PROOF We need only show that the left has side in (2.7) is not less than 1. Fix α satisfying $0 < \alpha < V$. By Lemma 2.1., for $C > 1$ there exists b_C so that for $n \geq b_C$ $v(k) \geq v(n)(k/n)^\alpha/C$ when $k > n$. By considering Riemann sums we deduce

$$\sum_{k=n+1}^{\infty} \exp -\gamma v(k) \leq \sum_{k=n+1}^{\infty} \exp -\frac{\gamma v(n)(k/n)^\alpha}{C} \leq n \int_1^\infty \exp -\frac{\gamma v(n)\, x^\alpha}{C}\, dx.$$

For all sufficiently large n, $\gamma v(n)/C > 1$, so

$$\sum_{k=n}^{\infty} \exp -\gamma v(k) \leq \exp -\gamma v(n) + n \left(\int_1^{\infty} \exp -(x^{\alpha} - 1)\, dx \right) \exp -\frac{\gamma v(n)}{C} .$$

Since $(\log n)/v(n) \to 0$, the desired inequality follows by taking $C \downarrow 1$.

Restricted LDP hypothesis: $(W_t/a(t), v(t))$ satisfies a restricted LDP, in the sense that there exists a function $J : \mathbb{R}_+ \to [0, \infty]$ such that, for each $c \geq 0$,

(2.8)
$$\lim_{t\to\infty} \frac{-1}{v(t)} \log \mathbb{P} \left[\frac{W_t}{a(t)} > c \right] = J(c).$$

If $(W_t/a(t), v(t))$ satisfies a Large Deviation Principle with rate–function $I(x)$ which is continuous where it is finite, then it satisfies the restricted LDP hypothesis with $J(x) := \inf_{y\geq x} I(y)$, for $x \geq 0$.

Stability and Continuity hypothesis: $J(0) > 0$ and there is some $c > 0$ such that $J(c) < \infty$. Moreover $J(x)$ is assumed to be continuous on the interior of the set upon which it is finite, which we denote $\mathcal{J}$.

In the queueing context, $J(0) > 0$ is the usual stability condition. If $J(x) = \infty$ for all $x > 0$, then $\mathbb{P}[Q > q]$ will be asymptotically zero with rate ∞.

Standard monotonicity arguments show the following:

LEMMA 2.4. *Under the restricted LDP, stability and continuity hypotheses, the limit*

$$\lim_{t\to\infty} \frac{-1}{v(t)} \log \mathbb{P} \left[\frac{W_t}{a(t)} > c \right] = J(c)$$

exists uniformly in c on compact intervals contained in $\mathcal{J}$, the interior of the set upon which $J(c)$ is finite.

The restricted LDP hypothesis refers to limiting behavior of $\log \mathbb{P}[W_t > ca(t)]$, not values for specific t. The asymptotics of a single W_t could dominate those of Q. The condition below excludes this possibility.

Uniform individual decay rate hypothesis: There exist constants $F > V/A$, $K > 0$, so that for all t and all $c > K$,

$$\frac{1}{v(t)} \log \mathbb{P}[W_t > ca(t)] \leq -c^F. \tag{2.9}$$

The results which follow apply for the parameter space $T = \mathbb{Z}_+$. If the following hypothesis obtains, they extend to $T = \mathbb{R}_+$.

Extension hypothesis:

$$\lim_{q\to\infty} \sup_{k\in\mathbb{Z}_+} \frac{\log \mathbb{P}[\sup_{k<t\leq k+1} W_t > q]}{h(q)} = \lim_{q\to\infty} \sup_{k\in\mathbb{Z}_+} \frac{\log \mathbb{P}[W_k > q]}{h(q)}.$$

This hypothesis is trivially satisfied when $T = \mathbb{Z}_+$. For $T = \mathbb{R}_+$, additional information about $\{W_t\}$ is needed to assure that the supremum over $k < t \leq k+1$ does not differ significantly from W_{k+1}. Though this hypothesis may be difficult to prove for specific models, in the queueing context it should be quite clear whether there is a significant difference between the maximum over all t and the maximum over $t \in \mathbb{Z}_+$.

The asymptotic lower bound for Q is a direct consequence of the restricted LDP for $(W_t/a(t), v(t))$.

PROPOSITION 2.1. *Under the scaling and restricted LDP hypotheses, we have*

$$\liminf_{q\to\infty} \frac{1}{h(q)} \log \mathbb{P}[Q > q] \geq -\inf_{c>0} c^V J\left(\frac{1}{c^A}\right). \tag{2.10}$$

PROOF Fix $c > 0$, as $Q = \sup_{t\geq 0} W_t$ we know that

$$\frac{1}{h(q)} \log \mathbb{P}[Q > q] \geq \frac{1}{h(q)} \log \mathbb{P}[W_{\lceil a^{-1}(cq)\rceil} > q]. \tag{2.11}$$

Elementary estimates yield

$$\frac{1}{v(\lceil a^{-1}(cq)\rceil)} \log \mathbb{P}\left[W_{\lceil a^{-1}(cq)\rceil} > q\right] = \frac{1}{v(\lceil b\rceil)} \log \mathbb{P}\left[W_{\lceil b\rceil} > \frac{a(b)}{c}\right],$$

with $b = a^{-1}(cq)$. The restricted LDP hypothesis and Lemma 2.2. ensure that

$$\liminf_{q\to\infty} \frac{1}{h(q)} \log \mathbb{P}[Q > q] \geq -c^{V/A} J\left(\frac{1}{c}\right).$$

As this is true for all $c > 0$,

$$\liminf_{q\to\infty} \frac{1}{h(q)} \log \mathbb{P}[Q > q] \geq -\inf_{c>0} c^{V/A} J\left(\frac{1}{c}\right).$$

Substituting $c' = c^{1/A}$ in for c, we get the result (2.10).

The upper bound is treated by splitting the event $\{Q > q\}$ into three parts, each of which is dealt with separately.

THEOREM 2.1. *For all $\underline{c} > \overline{c} > 0$, we have*

$$\limsup_{q\to\infty} \frac{1}{h(q)} \log \mathbb{P}[Q > q] = \max\{\Delta_1, \Delta_2, \Delta_3\}.$$

where:

$$\begin{aligned}
\Delta_1 &:= \limsup_{q\to\infty} h(q)^{-1} \log \mathbb{P}[\sup_{t:a(t)<q\underline{c}} W_t > q];\\
\Delta_2 &:= \limsup_{q\to\infty} h(q)^{-1} \log \mathbb{P}[\sup_{t:q\underline{c}\leq a(t)\leq q\overline{c}} W_t > q];\\
\Delta_3 &:= \limsup_{q\to\infty} h(q)^{-1} \log \mathbb{P}[\sup_{t:a(t)>q\overline{c}} W_t > q].
\end{aligned}$$

PROOF Using (1.4), a direct application of the principle of the largest term (see Lemma 1.2.15 of [3]) suffices.

First we treat Δ_3 defined in Lemma 2.1.:

PROPOSITION 2.2. *Under the scaling, restricted LDP, stability and continuity hypotheses, there exists $\infty > \overline{c} > 0$ such that*

$$\Delta_3 := \limsup_{q\to\infty} \frac{1}{h(q)} \log \mathbb{P}[\sup_{t:a(t)>q\overline{c}} W_t > q] < -\inf_{c>0} c^V J\left(\frac{1}{c^A}\right). \tag{2.12}$$

PROOF For $\overline{c} > 0$ we have

$$\mathbb{P}[\sup_{k:a(k)>q\overline{c}} W_k > q] = \mathbb{P}[\bigcup_{k:a(k)>q\overline{c}} W_k > q] \leq \sum_{k:a(k)>q\overline{c}} \mathbb{P}[W_k > q].$$

The restricted LDP and stability hypotheses imply

$$\lim_{t\to\infty} \frac{-1}{v(t)} \log \mathbb{P}\left[\frac{W_t}{a(t)} > 0\right] = J(0) > 0.$$

Select γ, $0 < \gamma < J(0)$. Then for all sufficiently large k

$$\mathbb{P}[W_k > q] \le \mathbb{P}[W_k > 0] \le e^{-\gamma v(k)},$$

and for all sufficiently large q,

$$\sum_{k:a(k)>q\bar{c}} \mathbb{P}[W_k > q] \le \sum_{k:a(k)>q\bar{c}} e^{-\gamma v(k)}.$$

By Lemma 2.3., there exists $\delta > 0$ so that $-\log\sum_{k=n}^{\infty} e^{-\gamma v(k)} > \delta\gamma v(n)$, for all sufficiently large n. Then

$$\limsup_{q\to\infty} \frac{1}{h(q)} \log \sum_{k:a(k)>q\bar{c}} \mathbb{P}[W_k > q] \le \limsup_{q\to\infty} -\delta\gamma\frac{h(q\bar{c})}{h(q)} = -\delta\gamma\bar{c}^{V/A}.$$

Since $c^{V/A} \to \infty$ as $c \to \infty$, we may choose $\bar{c} > 0$ so that

$$-\delta\,\gamma\,\bar{c}^{V/A} < -\inf_{c>0} c^V J\left(\frac{1}{c^A}\right).$$

Now we treat Δ_2 defined in Lemma 2.1.:

PROPOSITION 2.3. *Under the scaling, restricted LDP, stability and continuity hypotheses, for all* $0 < \underline{c} < \bar{c} < \infty$*,*

$$\Delta_2 := \limsup_{q\to\infty} \frac{1}{h(q)} \log \mathbb{P}[\sup_{t:q\underline{c}\le a(t)\le q\bar{c}} W_t > q] \le -\inf_{c>0} c^V J\left(\frac{1}{c^A}\right). \tag{2.13}$$

PROOF We have that

$$\mathbb{P}[\sup_{k:q\underline{c}\le a(k)\le q\bar{c}} W_k > q] \le a^{-1}(q\bar{c}) \max_{k:q\underline{c}\le a(k)\le q\bar{c}} \mathbb{P}\left[W_k > q\right]. \tag{2.14}$$

As $V > 0$

$$\limsup_{q\to\infty} \frac{1}{h(q)} \log a^{-1}(q\bar{c}) = 0.$$

Thus, using equation (2.14) and the fact that log is order–preserving, we have that the left hand side of equation (2.13) is less than or equal to

$$\limsup_{q\to\infty} \max_{k:q\underline{c}\le a(k)\le q\bar{c}} \frac{1}{h(q)} \log \mathbb{P}\left[W_k > q\right]. \tag{2.15}$$

Let

$$J_k(c) := \frac{-\log \mathbb{P}\left[W_k > c\,a(k)\right]}{v(k)}. \tag{2.16}$$

Select $\varepsilon > 0$, $\varepsilon < 1/\bar{c}$. If $J(1/\underline{c}) < \infty$, let $c^* := 1/\underline{c}$. Otherwise select c^* so that $J(c^*) < \infty$, $c^* + \varepsilon \leq 1/\underline{c}$ and $J(c^* + \varepsilon) = +\infty$. Now $\lim_k J_k(c) = J(c)$ for each $c \in [0, c^*]$ by the restricted LDP hypothesis. By lemma 2.4. we have uniform convergence on $[0, c^*]$. Note $J(c) \geq J(0) > 0$ for $c > 0$. Then there exists N_ε so that $n \geq N_\varepsilon$ implies

$$(2.17) \qquad J_n(c) > J(c)(1-\varepsilon) \text{ for } c \in [0, c^*] \text{ and } J_n(c^* + \varepsilon) > \frac{1}{\varepsilon} \text{ if } J(1/\underline{c}) = +\infty.$$

Note that $J_n(c) \geq J_n(c^*)$ for $c \in [c^*, c^* + \varepsilon]$ and $J_n(c) \geq J_n(c^* + \varepsilon)$ for $c \in [c^* + \varepsilon, 1/\underline{c}]$. For $q > a(N_\varepsilon \underline{c})$ define c_k for each k, $a^{-1}(q/\underline{c}) \leq k \leq a^{-1}(q/\bar{c})$, by $c_k := q/a(k)$ so that we have $v(k) = h(q/c_k)$ and from (2.16) and (2.17),

$$(2.18) \qquad -\log \mathbb{P}[W_k > q] > \begin{cases} h(q/c_k)J(c_k)(1-\varepsilon) & \text{if } 0 < c_k \leq c^*; \\ h(q/c_k)J(c^*)(1-\varepsilon) & \text{if } c^* < c_k < c^* + \varepsilon; \\ h(q/c_k)/\varepsilon & \text{if } c^* + \varepsilon \leq c_k \leq 1/\underline{c}. \end{cases}$$

The value of k at which the max in (2.15) occurs corresponds to the minimal term in (2.18). For sufficiently small ε the minimum does not occur at the $h(q/c_k)/\varepsilon$ term. Then this minimal term divided by $h(q)$ is not less than

$$\inf_{c \in [1/\bar{c}, 1/\underline{c}]} \frac{h(q/c)}{h(q)} J(c-\varepsilon)(1-\varepsilon) \geq \inf_{c>0} \frac{h(q/c)}{h(q)} J(c-\varepsilon)(1-\varepsilon).$$

Taking $q \to \infty$ and then the limit $\varepsilon \to 0$, and substituting $c' = c^{-A}$, yields (2.13).

Finally, we treat Δ_1 defined in Lemma 2.1.:

PROPOSITION 2.4. *Let the sequence $\{W_t\}$ satisfy the scaling, restricted LDP, stability and continuity hypotheses, and the uniform individual decay rate hypothesis, then there exists $\underline{c} > 0$ such that*

$$(2.19) \qquad \Delta_1 := \limsup_{q \to \infty} \frac{1}{h(q)} \log \mathbb{P}[\sup_{k: a(k) < q\underline{c}} W_k > q] < \inf_{c>0} c^V J\left(\frac{1}{c^A}\right).$$

PROOF Note that

$$\mathbb{P}[\sup_{k: a(k) < q\underline{c}} W_k > q] \leq q\underline{c} \max_{k: a(k) < q\underline{c}} \mathbb{P}[W_k > q],$$

and, as $V/A > 0$,

$$\limsup_{q \to \infty} \frac{1}{h(q)} \log q\underline{c} = 0.$$

Therefore

$$(2.20) \quad \limsup_{q \to \infty} \frac{1}{h(q)} \log \mathbb{P}[\sup_{k: a(k) < q\underline{c}} W_k > q] = \limsup_{q \to \infty} \max_{k: a(k) < q\underline{c}} \frac{1}{h(q)} \log \mathbb{P}[W_k > q].$$

Define c_k for each integer k, $0 < k \le a^{-1}(q\underline{c})$, by $c_k := q/a(k)$, so that we have $v(k) = h(q/c_k)$ and from (2.9) for each $c_k > K$,

$$\frac{1}{h(q)} \log \mathbb{P}[W_k > c_k a(k)] \le -\left(\frac{h(q/c_k)}{h(q)}\right) c_k^F = -\frac{h(a(k))}{h(q)} \left(\frac{q}{a(k)}\right)^F . \tag{2.21}$$

Take $C > 1$ and δ, $0 < \delta < F - V/A$. Letting $t = a(k)$, by Lemma 2.1. there exists $b_{\delta,C}$ so that for $q > t > b_{\delta,C}$,

$$\frac{h(t)}{h(q)} \ge \frac{1}{C} \left(\frac{t}{q}\right)^{V/A+\delta},$$

and, by (2.21), for $c_k := q/a(k) > \max\{K, 1\}$ and $a(k) > b_{\delta,C}$,

$$\frac{1}{h(q)} \log \mathbb{P}[W_k > c_k a(k)] \le -\frac{1}{C} c_k^{-V/A-\delta} c_k^F .$$

Choose $1/\underline{c} > \max\{K, 1\}$ so that

$$-\frac{1}{C}\left(\frac{1}{\underline{c}}\right)^{(F-V/A-\delta)} < -\inf_{c>0} c^V J\left(\frac{1}{c^A}\right).$$

Then for each $q > b_{\delta,C}$ and any k which satisfies $b_{\delta,C} < a(k) \le q\underline{c}$ we have

$$\frac{1}{h(q)} \log \mathbb{P}[W_k > c_k\, a(k)] < -\inf_{c>0} c^V J\left(\frac{1}{c^A}\right). \tag{2.22}$$

Inequality (2.21) implies that for each fixed k, $\log \mathbb{P}[W_k > c_k a(k)]/h(q) \to -\infty$ as $q \to \infty$. Since there are only finitely many k with $a(k) \le b_{\delta,C}$, (2.20) and (2.22) together imply (2.19).

From Proposition 2.1., Theorem 2.1. and Propositions 2.2., 2.3., 2.4., we deduce the following:

THEOREM 2.2. *If the sequence $\{W_t\}$ satisfies the scaling, restricted LDP, stability and continuity hypotheses, and the uniform individual decay rate hypothesis, then*

$$\lim_{q\to\infty} \frac{1}{h(q)} \log \mathbb{P}[Q > q] = -\inf_{c>0} c^V J\left(\frac{1}{c^A}\right). \tag{2.23}$$

2.1. The Scaled Cumulant Generating Function. The Cumulant Generating Function (sCGF) of W_t, scaled by $(a(t), v(t))$, is defined by

$$\lambda_t(\theta) := \frac{1}{v(t)} \log \mathbb{E}\left[\exp\left(\theta v(t) W_t / a(t)\right)\right]. \tag{2.24}$$

The analysis in [8] and [4] is based on the sCGF. The hypotheses of this paper have simple expressions in terms of the sCGF, when it exists. The conditions we specify here for the sCGF case are intended for easy applicability, rather than maximum generality. Under these assumptions, the large deviation rate–function is convex. The sCGF technique is not applicable to models which have non–convex rate–functions.

LDP Hypothesis, sCGF case: For all $\theta \in \mathbb{R}$ and all t the scaled cumulant generating function given by (2.24) exists as a finite real value. For each $\theta \in \mathbb{R}$ the following finite limit exists and defined $\lambda(\theta)$:

$$\lambda(\theta) := \lim_{t\to\infty} \lambda_t(\theta).$$

Furthermore $\lambda(\theta)$ is assumed to be continuously differentiable. These assumptions imply the following (see [3]):

PROPOSITION 2.5. *Under the above assumptions the pair $\big(W_t/a(t), v(t)\big)$ satisfies a large deviation principle with rate–function $I(x)$ given by the Legendre–Fenchel transform of $\lambda(\theta)$,*

$$I(x) := \sup_{\theta}\{\theta x - \lambda(\theta)\}. \tag{2.25}$$

This implies that $I(x)$ is a convex function and continuous on the interior of the set where it is finite.

Stability Hypothesis, sCGF case: There exists $\overline{\theta} > 0$ so that $\lambda(\overline{\theta}) < 0$.

The above implies that $J(x) = I(x)$ for $x \geq 0$ and $J(0) \geq -\lambda(\overline{\theta})$.

The uniform individual decay rate hypothesis can be readily expressed in terms of the sCGF.

PROPOSITION 2.6. *If there exist constants F', M such that $F' > \max\{V/A, 1\}$ and*

$$\lambda_t(\theta) \leq M\theta^{F'/(F'-1)}$$

for all $\theta > 0$ and all t, then for each F, $1 < F < F'$, there exists K_F so that for all $c > K_F$ and all t,

$$\frac{1}{v(t)} \log \mathbb{P}[W_t > c\, a(t)] \leq -c^F$$

and the uniform individual decay rate hypothesis is satisfied.

PROOF An elementary consequence of (2.24) is Chernoff's inequality,

$$\log \mathbb{P}[W_t > c\,a(t)] \le -v(t)\left(c\,\theta - \lambda_t(\theta)\right).$$

It then follows that

$$\log \mathbb{P}[W_t > c\,a(t)] \le -v(t)\left(c\,\theta - M\theta^{F'/(F'-1)}\right).$$

Choosing $\theta = (c(F'-1)/(MF'))^{F'-1}$ we have

$$\log \mathbb{P}[W_t > c\,a(t)] \le -v(t)c^{F'}\left(M^{1-F'}F'^{-F'}(F'-1)^{F'-1}\right). \tag{2.26}$$

Since M and F' are constants, for each F, $F' > F > \max\{V/A, 1\}$, there exists K_F such that, for all $c > K_F$, the right hand side of (2.26) will be less than $-v(t)c^F$.

3. Examples.

3.1. Application to Gaussian Processes. Perhaps the simplest examples to which the theory can be applied are those in which

$$W_t := X_t - \mu t,$$

where $\mu > 0$ is constant and $\{X_t\}$ are mean zero Gaussian random variables with covariance function

$$\Gamma(s,t) := \mathbb{E}[X_s X_t], \qquad \sigma_t^2 := \Gamma(t,t).$$

Define

$$a(t) := t \qquad \text{and} \qquad v(t) \equiv h(t) := t^2/\sigma_t^2.$$

$\lambda_t(\theta)$ is given by

$$\lambda_t(\theta) = \frac{1}{2}\theta^2 - \mu\theta,$$

for all $t \ge 0$. Using Equation (2.25), we get

$$I(x) = \frac{1}{2}(x+\mu)^2. \tag{3.27}$$

$I(x)$ is non–decreasing for $x \ge 0$, hence $J(c) = I(c)$ for all $c \ge 0$. Moreover $J(c)$ is continuous where it is finite. It suffices to assume that, for all $c > 0$,

$$\lim_{t\to\infty} \frac{\sigma_{ct}^2}{\sigma_t^2} = \lim_{t\to\infty} \frac{\Gamma(ct,ct)}{\Gamma(t,t)} = c^{2H}, \qquad 0 < H < 1.$$

Then $v(t)$ is regularly varying with

$$\lim_{t\to\infty} \frac{v(ct)}{v(t)} = c^V, \qquad V = 2 - 2H.$$

Proposition 2.6. holds with $M = 1$ and $F' = 2$, hence the uniform individual decay rate hypothesis is satisfied and

$$\inf_{c>0} c^V J\left(\frac{1}{c}\right) = \frac{2}{V^V}\left(\frac{\mu}{2-V}\right)^{2-V}. \tag{3.28}$$

EXAMPLE 1 *Let* $\{X_t : t \in \mathbb{Z}_+\}$ *be an independent sequence with* $\sigma_t^2 = t$ *and* $\mu = 1$, *then* W_t *has mean* $-t$, *variance* t, *and the appropriate scale is* $h(q) := q$. *The hypotheses of Theorem 2.2. are satisfied, hence*

$$\lim_{q\to\infty} \frac{1}{q} \log \mathbb{P}[Q > q] = -2. \tag{3.29}$$

Replace W_1 *with an exponential distribution which starts at* -2α, *so that:*

$$\mathbb{P}[W_1 > a] = \exp -((a + 2\alpha)/\alpha),$$

for $a > -2\alpha$. *This distribution has mean* $-\alpha$ *and* $\lim q^{-1} \log \mathbb{P}[W_1 > q] = -\alpha^{-1}$. $(W_t/t, t)$ *still satisfies an LDP with rate–function given by Equation (3.27), but if* $\alpha > 1/2$ *the uniform individual decay rate hypothesis is not satisfied, Equation (3.29) does not hold and* W_1 *determines the asymptotics of* Q.

EXAMPLE 2 *Fractional Brownian motion has*

$$2\Gamma(s,t) = s^{2H} + t^{2H} - |s-t|^{2H}, \qquad \sigma_t^2 = t^{2H},$$

where $0 < H < 1$. *Here* $a(t) := t$, $v(t) := h(t) := t^{2(1-H)}$ *and*

$$\lim_{q\to\infty} \frac{1}{h(q)} \log \mathbb{P}[Q > q] = \frac{2}{V^V}\left(\frac{\mu}{2-V}\right)^{2-V}.$$

Similarly, the Ornstein–Uhlenbeck position process has

$$\sigma_t^2 = 2\left(\frac{\mu}{\nu}\right)^2 (t + e^{-t} - 1),$$

where μ and ν are positive constants. Clearly σ_t^2 is regularly varying with $2H = 1$. See [4] for details.

3.2. Application to Heavy Tailed Processes. Define the heavy tailed distribution Y by

$$\mathbb{P}[Y \geq x] := d(x)e^{-v(x)},$$

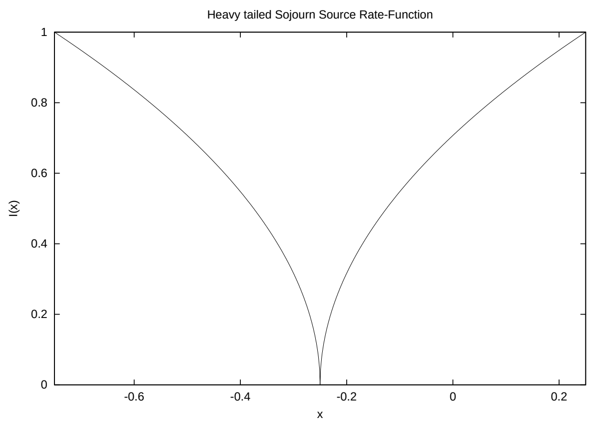

Figure 1: $I(x)$ vs. x for heavy tailed sojourn times.

where $d(x)$ is slowly varying (see Bingham *et al.* [2]), and $v(x)$ is regularly varying with constant $0 < V < 1$. Define $\{Y_t\}$ to be a stationary sequence of two state random variables taking the values 0 and 1 whose sojourn times spent in the 0 and 1 states are distributed by an i.i.d. sequence with distribution Y. Note that the mean of Y_t is $1/2$ as its sojourn times spent in the 'on' and 'off' states have finite, and equal, expectation. Define

$$Z_t := Y_t - \mu,$$

where $\mu \in (1/2, 1)$ so that the stability hypothesis is satisfied. Define the workload process by

$$W_t := \int_0^t Z_s \, ds.$$

It is shown in [5] that $(W_t/t, v(t))$ satisfies a large deviation principle with rate–function, $I(x)$, given by

$$I(x) = \begin{cases} (1 - 2(x+\mu))^V & \text{if } x \in [-\mu, 1/2 - \mu], \\ (2(x+\mu) - 1)^V & \text{if } x \in [1/2 - \mu, 1 - \mu], \\ +\infty & \text{otherwise.} \end{cases}$$

For example, with $\mu := 3/4$, $d(x) := 1$ and $V := 1/2$, see figure 1 for a graph of $I(x)$. Note that $I(x)$ is non–decreasing for $x \geq 0$, hence $J(c) = I(c)$ for all $c \geq 0$. Moreover, $J(c)$ is continuous where it is finite. We have

$$c^V J\left(\frac{1}{c}\right) = \begin{cases} +\infty & \text{if } c < \frac{1}{1-\mu}, \\ (2 + c\,(2\mu - 1))^V & \text{if } c \geq \frac{1}{1-\mu}, \end{cases}$$

and

$$\inf_{c>0} c^V J\left(\frac{1}{c}\right) = \left(2 + \frac{2\mu - 1}{1-\mu}\right)^V . \tag{3.30}$$

As Z_t is bounded above by $1-\mu$, the uniform individual decay rate hypothesis is satisfied. Hence, Theorem 2.2. and Equation (3.30) give:

$$\lim_{q\to\infty} \frac{1}{h(q)} \log \mathbb{P}[Q > q] = \left(2 + \frac{2\mu - 1}{1-\mu}\right)^V .$$

4. Connections with existing literature. The basic aim of tail asymptotics is to compute

$$\lim_{q\to\infty} \frac{1}{h(q)} \log \mathbb{P}[Q > q] = -\delta,$$

where $h(q)$ is an appropriately chosen scaling function. Glynn and Whitt [8] treat the case where $h(q) = q$, which is appropriate when a large deviation principle (LDP) is satisfied by the pair $(W_t/t, t)$. Their result does not cover the case of Fractional Brownian Motion (FBM), for example.

The introduction by Duffield and O'Connell [4] of a more general class of scaling functions, wide enough to treat FBM, aroused a great deal of interest and is widely cited. However, there is a gap in their arguments (their equation (36) does not hold in general); our Example 1 shows that it cannot be filled without introducing a further hypothesis. For instance, if one assumes that $\lambda_t(\theta)$, the sCGF defined at (2.24), is independent of t, then their equation (36) holds; this covers the case of FBM. It should be noted that in the case of FBM the approach of Massoulie and Simonian [12] yields the tail asymptotics immediately and, moreover, gives precise (non–asymptotic) probability estimates. Glynn and Whitt [8] exclude the possibility that the asymptotics of Q are dominated by those of a single W_t by assuming that $\mathbb{E}[\exp(\delta W_t)] < \infty$ for all $t \geq 0$, where $\delta := \inf_{c>0} cJ(c^{-1})$. The use of the Gärtner–Ellis Theorem [7, 6] in the approaches adopted by [8, 4] excludes the possibility of the rate–function being non–convex.

None of the work cited above covers the case where the appropriate scale for the large deviations of W_t/t is $\log(t)$. This occurs when W_t has independent increments having a common power–tail distribution. The asymptotics of Q in this case are discussed by Parulekar and Makowski [14], Mikosch and Nagaev [13], Asmussen and Collamore [1], Liu *et. al.* [11] and references therein.

References

[1] S. Asmussen and J. F. Collamore, *Exact asymptotics for a large deviations problem for the GI/G/1 queue*, Markov Processes and Related Fields **5** (1999), no. 4, 451–476.

[2] N. H. Bingham, C. M. Goldie, and J. L. Teugels, *Regular variation*, Cambridge University Press, 1987.

[3] A. Dembo and O. Zeitouni, *Large deviation techniques and applications*, Springer, 1998.

[4] N. G. Duffield and N. O'Connell, *Large deviations and overflow probabilities for the general single server queue, with applications*, Mathematical Proceedings of the Cambridge Philosophical Society **118** (1995), no. 2, 363–374.

[5] K. Duffy, *Logarithmic asymptotics in Queueing Theory and Risk Theory*, Ph.D. Thesis, Trinity College Dublin (2000).

[6] R. Ellis, *Large Deviations for a general class of random vectors*, Annals of Probability, **12** (1984), 1–12.

[7] J. Gärtner, *On large deviations from the invariant measure*, Theory of Probability and its Applications, **22** (1977), 24–39.

[8] P. Glynn and W. Whitt, *Logarithmic asymptotics for steady-state tail probabilities in a single-server queue*, Journal of Applied Probability **31A** (1994), 413–430.

[9] P. R. Halmos, *Ergodic theory*, Chelesa Publishing Company, 1956.

[10] R. M. Loynes, *The stability of a queue with non-independent inter-arrival and service times*, Proceedings of the Cambridge Philosphical Society **58** (1962), 497–520.

[11] Z. Liu, P. Nain, D. Towsley and Z. L. Zhang, *Asymptotic behavior of a multiplexer fed with a long-range dependent process*, Journal of Applied Probability **36** (1999), no. 1, 105–118

[12] L. Massoulie and A. Simonian, *Large buffer asymptotics for the queue fed with FBM input*, Journal of Applied Probability **36** (1999), no. 3, 894–906

[13] T. Mikosch and A. V. Nagaev, *Large deviations of heavy-tailed sums with applications in insurance*, Extremes **1** (1998), no. 1, 81–110.

[14] M. Parulekar and A. M. Makowski, *M/G/Infinity input processes: A versatile class of models for network traffic*, IEEE Infocom (1997), vol. 2, 419–426.

COMMUNICATIONS NETWORK RESEARCH INSTITUTE,
DUBLIN INSTITUTE OF TECHNOLOGY,
RATHMINES, DUBLIN 6,
IRELAND.
E-MAIL: {ken.duffy, john.lewis, wayne.sullivan}@cnri.dit.ie