Heavy-tailed distributions

We now focus on classes of distributions for which $E(e^{\epsilon X}) = \infty,\ \epsilon > 0$.

Given a non-negative random variable (r.v.) X, its distribution function (d.f.) is denoted by $F(x) = P(X \leq x)$ and its tail by $\overline{F}(x) = 1 - F(x) = P(X > x)$. A d.f. F (or the r.v. X) is said to be *heavy-tailed* if $\overline{F}(x) > 0,\ x \geq 0$, and for all $y \geq 0$,

$$\lim_{x\to\infty} P(X > x+y \mid X > x) = \lim_{x\to\infty} \frac{\overline{F}(x+y)}{\overline{F}(x)} = 1. \tag{1}$$

Letting $a(x) \sim b(x)$ mean that $a(x)/b(x) \to 1$ as $x \to \infty$, we can express (1) as

$$\overline{F}(x+y) \sim \overline{F}(x), \text{ for all } y \geq 0.$$

Intuitively this means that

> if X ever exceeds a large value, then it is likely to exceed any larger value as well; its tail is *heavy* or *fat* or *long*.

We denote the class of heavy-tailed distributions by $\mathcal{L}$ (and use the notation $F \in \mathcal{L}$ or $X \in \mathcal{L}$).

Heavy-tailed distributions differ sharply with the exponential d.f. $F(x) = 1 - e^{-\lambda x}$ which satisfies

$$\frac{\overline{F}(x+y)}{\overline{F}(x)} = e^{-\lambda y}, \ x \geq 0, \ y \geq 0,$$

and hence is not heavy-tailed.

Examples:

1. (*Pareto:*) $\overline{F}(x) = x^{-\alpha}$, $x \geq 1$, with $\alpha > 0$. (Many variations on this exist, such as $\overline{F}(x) = (\frac{c}{c+x})^{\alpha}$, $x \geq 0$, with $c > 0$ and $\alpha > 0$.)

2. (*Lognormal:*) Density

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{\frac{-(\ln(x)-\mu)^2}{2\sigma^2}}, \ x > 0,$$

with $\sigma > 0$ and $\mu \in (-\infty, \infty)$. This is the distribution of the r.v. $X = e^Y$ where Y is normal with mean μ and variance σ^2.

3. (*Heavy-tailed Weibull:*) $\overline{F}(x) = e^{-\lambda x^{\alpha}}$, $x \geq 0$, with $\lambda > 0$ and $0 < \alpha < 1$. Such a r.v. X can be derived from an exponential r.v. $P(Y > x) = e^{-\lambda x}$ via the transformation $X = Y^{1/\alpha}$.

A very important class is (*Regularly varying tails:*) With $\alpha \geq 0$, $\overline{F}$ is said to be regularly varying with index $-\alpha$ if it is a regularly varying function, that is, if

$$\lim_{x \to \infty} \frac{\overline{F}(tx)}{\overline{F}(x)} = t^{-\alpha}, \ t > 0.$$

Such tails can be equivalently represented in the form $\overline{F}(x) = L(x)x^{-\alpha}$, where $L(x)$ is a slowly varying function (that is, regularly varying with $\alpha = 0$; $L(tx)/L(x) \to t$). Examples of regularly varying tails include the Pareto tail, but by using slowly varying factors such as $L(x) = c\ln(x)$ or $c\ln(\ln(x))$, or a function $L(x)$ that converges to a constant, a large variety of tails is included here.

Much of the beginning of heavy-tailed applications to queues began with using regularly varying tails, due to their nice form and relative ease of manipulation. Even today, sometimes it is possible/desirable to first prove a result for regularly varying tails to motivate trying to prove the result more generally. These tails however are not exhaustive (they don't contain the Weibull for example) and it turns out that a larger class of heavy-tailed distributions called *subexponential* distributions has become the standard, and includes all the examples we have given.

Definition 1 *(**Subexponential distributions**)* *The d.f. F (or the r.v. X) is called subexponential if $\overline{F}(x) > 0,\ x \geq 0$, and for all $n \geq 2$,*

$$\lim_{x \to \infty} \frac{\overline{F^{*n}}(x)}{\overline{F}(x)} = n. \tag{2}$$

It can be shown that if the condition holds for some $n \geq 2$, then it holds for all $n \geq 2$.

Here, F^{*n} denotes the n-fold convolution of F, $F^{*2}(x) = \int_0^x F(x-y)dF(y)$ and so on, with corresponding tail $\overline{F^{*n}}(x) = 1 - F^{*n}(x)$.

In terms of r.v.s., (2) can thus be re-stated as

$$P(X_1 + \cdots + X_n > x) \sim nP(X > x),$$

and can equivalently and most importantly be stated as

$$P(X_1 + \cdots + X_n > x) \sim P(\max\{X_1, \ldots, X_n\} > x), \tag{3}$$

for all $n \geq 2$ where $X_1, \ldots, X_n$ are i.i.d. distributed as F. In words (3) means that

the sum is likely to get large because one of the r.v.s. gets large.

It is this interpretation that justifies using subexponential distributions in stochastic modeling of stochastic networks.

Two important properties of $\mathcal{S}$ are contained in

Proposition 1 *(a) If $F \in \mathcal{S}$ and G is any d.f. such that*

$$\overline{G}(x) \sim c\overline{F}(x), \text{ with constant } c > 0, \quad (4)$$

then $G \in \mathcal{S}$ and $F * G \in \mathcal{S}$ and $\overline{F * G}(x) \sim (1 + c)\overline{F}(x)$.

(b) If $F \in \mathcal{S}$ and G is any d.f. such that

$$\overline{G}(x)/\overline{F}(x) \to 0, \quad (5)$$

then $F * G \in \mathcal{S}$ and $\overline{F * G}(x) \sim \overline{F}(x)$.

Two d.f.s F and G (or r.v.s. X and Y) satisfying (4) are said to be *tail equivalent*, whereas if they satisfy (5) we say that F has a heavier (or fatter) tail than G (equivalently, we say that G has a lighter tail than F). Note that (5) holds, in particular, for any subexponential F and any light-tailed G.

For technical reasons we sometimes restrict the class $\mathcal{S}$ even further to the class $\mathcal{S}^* \subset \mathcal{S}$, introduced by Klüppelberg and defined by

Definition 2 *(**The class** $\mathcal{S}^*$) Let F be a d.f. on $[0,\infty)$ such that $\overline{F}(x) > 0,\ x \geq 0$. We say that $F \in \mathcal{S}^*$ if F has finite first moment $1/\mu$ and*

$$\lim_{x\to\infty} \mu \int_0^x \frac{\overline{F}(x-y)}{\overline{F}(x)} \overline{F}(y)dy = 2. \qquad (6)$$

$\mathcal{S}^*$ includes (when the mean is finite) all the examples we mentioned for $\mathcal{S}$; so for all practicle purposes, we can (and sometimes do) assume that we are dealing with distributions in $\mathcal{S}^*$. An important property of $\mathcal{S}^*$ is that if $F \in \mathcal{S}^*$, then not only is F subexponential, but so is F_e (the equilibriuim distribution (integrated tail df) of F).

Single-server delay asymptotics

Interarrival times $\{T_n\}$ are i.i.d. distributed as $A(x) = P(T \leq x)$ with finite non-zero mean $E(T) = 1/\lambda$, and independently service times $\{S_n\}$ are i.i.d. distributed as $F(x) = P(S \leq x)$ with finite non-zero mean $E(S) = 1/\mu$. $\rho \stackrel{\text{def}}{=} \lambda/\mu < 1$ (stability). The delay of the n^{th} customer (in queue, not including service) is denoted by D_n and satisfies the recursion

$$D_{n+1} = (D_n + S_n - T_n)_+, \ n \geq 0. \qquad (7)$$

D denotes steady-state delay: $P(D \leq x) = \lim_{n\to\infty} P(D_n \leq x)$.

D has the same distribution as the maximum of the negative drift random walk $R_n,\ n \geq 0$, where

$$R_n = \sum_{j=1}^{n} (S_j - T_j),\ n \geq 1, \quad R_0 = 0; \tag{8}$$

$$D \stackrel{\mathcal{D}}{=} \max_{n \geq 0} R_n. \tag{9}$$

($X \stackrel{\mathcal{D}}{=} Y$ denotes that X and Y have the same distribution.)

For any non-negative random variable X with distribution F and finite mean $1/\mu$, the *equilibrium distribution* F_e (or integrated tail distribution) is defined by

$$F_e(x) = \mu \int_0^x \overline{F}(y)dy\,, \quad x > 0\ . \tag{10}$$

It arises naturally as the distribution of the stationary forward recurrence time in point processes, the stationary remaining service time distribution in queues, and is intimately related to the *inspection paradox*. By differentiating (10) we see that F_e always has a density $f_e(x) \stackrel{\text{def}}{=} \mu\overline{F}(x)$ on $(0,\infty)$.

We let X_e denote a r.v. distributed as F_e, which in the context of service times will be denoted by S_e.

Lemma 1 (a) *If $F \in \mathcal{L}$, then $F_e \in \mathcal{L}$ and $\overline{F}_e$ is heavier than $\overline{F}$;*

$$\lim_{x\to\infty} \frac{\overline{F}(x)}{\overline{F}_e(x)} = 0. \tag{11}$$

(b) *If $F_e \in \mathcal{L}$, then $\overline{F}_e$ is heavier than $\overline{F}$; (11) holds.*

Summarizing: If either F or F_e is heavy-tailed, then F_e has a fatter tail than F. In particular, if $F_e \in \mathcal{S}$ then F_e has a fatter tail than F and we conclude that $P(X_e + X > x) \sim P(X_e > x)$ when X and X_e are independent (recall Proposition 1 **(b)**).

The following shows how fundamental $\mathcal{S}$ is in the context of queues (A. Pakes (1975)):

Theorem 1 *For the stable FIFO GI/GI/1 queue*

(a) If S_e is subexponential, then

$$P(D > x) \sim \frac{\rho}{1-\rho} P(S_e > x). \qquad (12)$$

(In particular, D is subexponential since it is tail equivalent to S_e.)

(b) If the arrival process is Poisson, then D is subexponential if and only if S_e is subexponential if and only if (12) holds.

The proof of (b) is rather easy, following from the classic *Pollaczek-Khinchine* formula for M/G/1 which expresses D as an independent geometric sum of i.i.d. r.v.s. (ladder heights of the underlying random walk) distributed as S_e:

$$D = \sum_{j=1}^{N} Y_j,$$

where Y_j are iid $\overset{\text{d}}{\sim} S_e$ and $P(N = n) = (1 - \rho)\rho^n$, $n \geq 0$. Noting that $E(N) = \frac{\rho}{1-\rho}$, the proof merely requires justifying that the defining subexponential property (2) can be extended to random summands N, with $E(N)$ replacing N (which indeed is true when N has finite MGF).

The proof of (a) is much harder, requiring further deeper results on ladder heights of random walks and more subtle properties of subexponential distributions.

The key point is that D still has a geometric sum representation as

$$D = \sum_{j=1}^{N} Y_j,$$

but now Y_j are iid with a complex general ladder height distribution (not F_e) and $P(N = n) = (1-\theta)\theta^n$ where θ (the probability of at least one ladder height) is not explicitely known (e.g., it is not ρ anymore). The proof then consists of proving that
$P(Y > x) \sim cP(S_e > x)$, where $cE(N) = \frac{\rho}{1-\rho}$.

Sojourn time $W = D + S$ is easily handled now: when $F_e \in \mathcal{S}$ we now know that $D \in \mathcal{S}$ is tail equivalent to S_e and that S has a lighter tail than S_e; hence $P(W > x) \sim P(D > x)$. Similarly, workload V can be handled since $P(V > x) = \rho P(D + S_e > x)$ holds for any GI/GI/1 queue. Thus when $F_e \in \mathcal{S}$, $P(V > x) \sim \rho(\frac{\rho}{1-\rho} + 1)P(S_e > x) = \frac{\rho}{1-\rho}P(S_e > x)$.

Summarizing: when $F_e \in \mathcal{S}$, all three quantities V, W, D have the same tail asymptotic; $\frac{\rho}{1-\rho}P(S_e > x),\ x \to \infty$.

Tandem queues and beyond

The single-server asymptotics for delay extend in a variety of ways to single-class networks, the general rule being that the sojourn time tail is equivalent to the heaviest service time tail. For example, consider a stable tandem queue with 2 nodes, service time dfs F_1 and F_2 both in $\mathcal{S}^*$ but with F_1 lighter than F_2. Then sojourn time W has the following tail asymptotic: $P(W > x) \sim \frac{\rho_2}{1-\rho_2} P(S_e(2) > x)$. ($\rho_2 = \lambda E(S(2))$.)

For the FIFO GI/GI/c queue, the tail asymptotics are not fully worked out. The complication arises due to the fact that in this case, the df of the tail depends on the value of ρ. For example, if $c = 2$ and $\rho < 1$, then $E(D) < \infty$ if $E(S^{3/2}) < \infty$, but if $1 < \rho < 2$, then $E(D) = \infty$ unless $E(S^2) < \infty$.

Queue length is different: importance of $e^{\sqrt{x}}$

For the single-server queue, it is now known that the heavy-tailed asymptotic for queue length $P(L > k)$, is not as simple as for delay; it depends on how heavy the service time tail is.

From distributional Little's law,

$$L \stackrel{\mathsf{d}}{=} N(W),$$

($\{N(t) : t \geq 0\}$ is a time-stationary version of the renewal arrival process and W is an independent copy of sojourn time).

Thus one might expect that when $F_e \in \mathcal{S}$, $P(L > k) \sim P(\lambda W > x)$ because $N(t)$ should only enter through its mean behavior. This turns out to be true only if F_e has a tail that is heavier than the Weibull tail $e^{\sqrt{x}}$.

Because of this, we say that a distribution is *moderately heavy-tailed* if it is heavy-tailed but lighter than (or same as) $e^{\sqrt{x}}$. This would include for example Weibull tails such as $\overline{F}(x) = e^{-\lambda x^{\alpha}}$, when $0.5 < \alpha < 1$, but not the lognormal. (Tail asymptotics are known for L when F_e is moderately heavy-tailed but the form is very complicated.)

This area of research led to the notion of *independent sampling* of a process (such as a Poisson counting process $N(t)$) at a random time T; $N(T)$, and figuring out what the tail asymptotics of $N(T)$ are when T is heavy-tailed.

$e^{\sqrt{x}}$ arises elsewhere too: Busy periods

In 1980 it was shown that the busy period B for the M/G/1 queue had a tail like $P(B > x) \sim (1-\rho)^{-1} P(S > (1-\rho)x)$, when S is regularly varying. (Note how it is not S_e that is here, only S itself.)

In 1999 it was observed that this asymptotic could not hold if S was moderatey heavy-tailed.

Since then, there has been progress in proving that the asymptotic does hold when S is subexponential and heavier than $e^{\sqrt{x}}$. Extensions to GI/GI/1 also exist.

Interestingly, for the M/G/1, a busy period in which the first service time is deterministic of length x, has reprentation

$$B(x) = x + \sum_{i=1}^{N(x)} B_i,$$

where the B_i are iid copies of B. Thus $\{B(x) : x \geq 0\}$ forms a compound Poisson process in which B has the same distribution as $B(S)$ with S chosen independently. One would expect this fact to be useful in proving tail asymptotics for B, via "independent sampling" but a proof along these lines has eluded us.

$e^{\sqrt{x}}$ arises again: Processor sharing

Recently there has been results showing that if S is suitably heavy tailed, then the sojourn time W for a M/G/1 queue under processor sharing (PS) is of the nice form

$$P(W > x) \sim P(S > (1-\rho)x).$$

Once again it turns out that S must have a heavier tail than $e^{\sqrt{x}}$ in order that this asymptotic holds.

One way to gain some intuition as to how such an asymptotic would arise:

It is well known that the *expected* stationary sojourn time for a customer with a service time of length x is given by $(1-\rho)^{-1}x$ and in general $E(W) = (1-\rho)^{-1}E(S)$.

Finally, we point out that the $e^{\sqrt{x}}$ is key to so-called "reduced load equivalence" results in fluid queues.

Multi-class networks: a mess

Suffice to say that in a multi-class setting, there are no general results. The problem is that even stability is a serious issue: The intuitive $\rho_i < 1$ conditions while necessary for stability are no longer sufficient. In one example, it appears that by making a given distribution (path length) heavier, the system becomes more stable! So it is possible that the stability region of a model can even depend upon component distributions.

This is a huge open area of research.