Applications of Large Deviations to Performance Analysis with Long-Range Dependent Traffic

Nick Duffield
AT&T Labs-Research
Florham Park, New Jersey

AT&T

Outline

- Quick tour of Long-Range Dependence
- Quick tour of Large Deviations
- Recovery times in bufferless systems
- Queueing asymptotics
  - ◇ large buffers
  - ◇ many sources
- Measurement-based Admission Control
  - ◇ measurement of effective bandwidths, etc.
  - ◇ analysis of errors
- Bibliography

Heavy Tailed Distributions

Definition. *A positive random variable is* **heavy-tailed** *if it has infinite variance. (If not,* **light-tailed***)*

Operationally: X_i are i.i.d. samples from heavy-tailed dist.: variance of n samples diverges as $n \to \infty$. (Contrast light-tailed: converges a.s. to Var$X$).

Prime Example: Pareto distribution: $\mathbb{P}[X > x] \sim kx^{-\alpha}$ as $x \to \infty$. Heavy tailed if $\alpha < 2$, no mean if $\alpha < 1$.

Signatures:

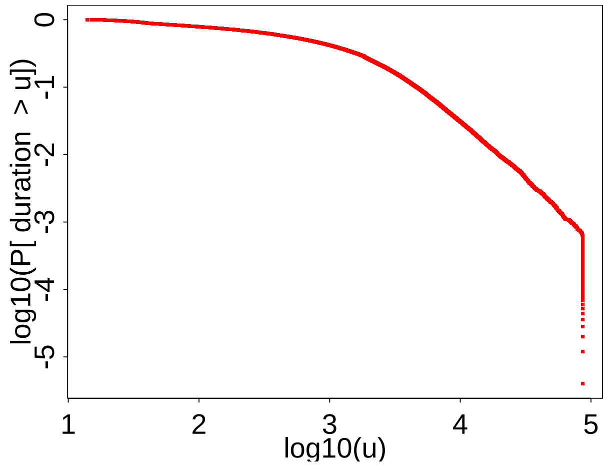
- CCDF has slope $-\alpha$ on log-log scale.
- high variability of samples
- poor averaging properties over samples
- small number of samples carry large proportion of the mean (e.g. 80-20 rules)

Some Empirical Heavy Tailed Distributions

- file sizes
- durations and bytes of TCP connections
- durations and bytes of modem calls
- number of TCP connections per modem call

Example

Modem call durations



Slope in right hand portion gives $\alpha \approx 1.5$

Values truncated at 24 hours

Boundedness: Many actual variables are bounded

- e.g. disk size bounds file sizes
- sampling properties can be close to heavy-tailed distribution
- "closeness" depends on quantities of interest
  - ◇ e.g. $\mathbb{P}[X \text{near bound}] \ll 1$
  - ◇ e.g. $\mathbb{E}[X \mid X \text{near bound}] \ll \mathbb{E}[X]$

Self-Similarity

Definition. *A process* $(Y(t) : t \in \mathbb{R})$ *is* **self-similar** *if for some* **Hurst parameter** H*, then for each* $s > 0$ *the processes* $Y(s\cdot)$ *and* $s^H Y(\cdot)$ *have the same finite dimensional distributions.*

Discrete time increment process $X_i = Y_i - Y_{i-1}$, with aggregates

$$X_i^{(m)} = \frac{1}{m} \sum_{j=(m-1)i+1}^{mi} X_j$$

then Y self-similar gives

$$(X_1^{(m)}, \ldots, X_n^{(m)}) \stackrel{d}{=} m^{H-1}(X_1, \ldots, X_n)$$

Consequences: $\mathrm{Var}(X_1^{(m)}) = m^{2H-2}\mathrm{Var}(X_1^{(1)})$

- decays slower than m^{-1} if $H > 1/2$
- contrast e.g. X_i. i.i.d. ($\mathrm{Var}(X^{(m)}) \propto 1/m$)
- signature: $\log \mathrm{Var}(X^{(m)})$ vs. $\log m$ has slope $2H - 2$

Example: $H \in (0, 1)$, fractional Brownian Motion (fBM), Gaussian process Y with stationary increments and autocovariance function

$$d(s, t) = t^{2H} + s^{2H} - 2|t - s|^{2H}$$

Long-Range Dependence

Increment process of fBM $X_i = Y_i - Y_{i-1}$ is fractional Gaussian Noise. Autocorrelation function

$$\mathrm{Cor}(X_j, X_{j+k}) \sim (2H-1)k^{2H-2}, \quad (k \to \infty)$$

Definition. *A second-order stationary process* $(X_k; k \in \mathbb{Z})$ *is* **long range dependent** *if*

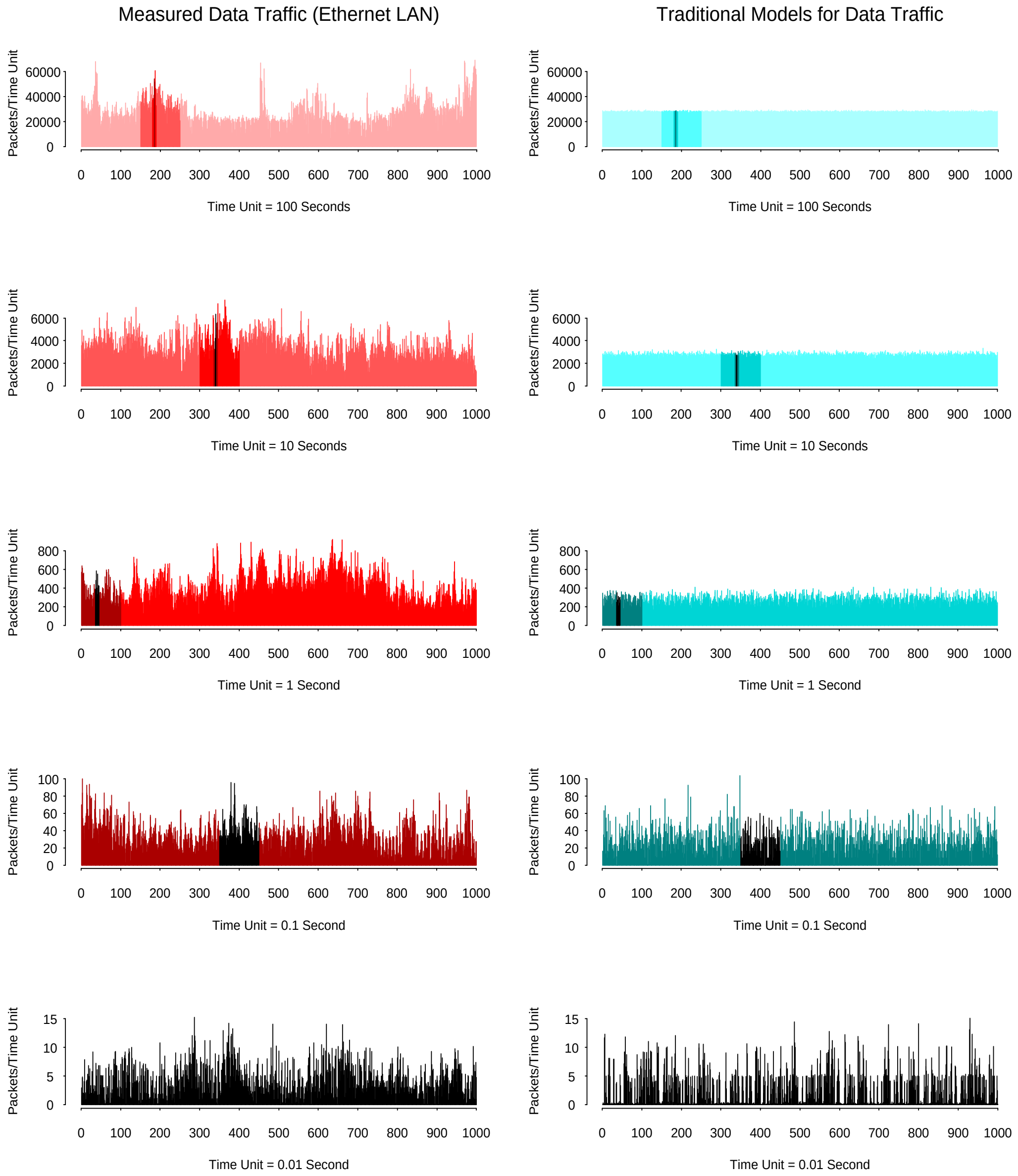
$$\sum_{k \geq 0} \gamma_k \text{ is divergent, where} \qquad \gamma_k = \mathrm{Cor}(X_j, X_{j+k})$$

(Otherwise **short-range dependent***). (Use* $\int dt$ *when* $t \in \mathbb{R}$*).*

Consequences and Signatures

- slow decay of correlations
- long-lived trends
- poor temporal averaging properties

Some Evidence for LRD in Teletraffic

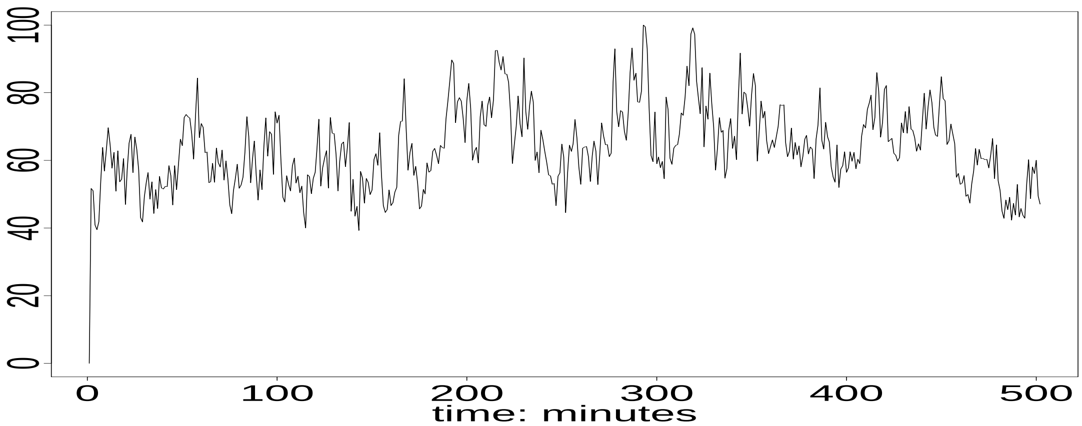


Time-scale Issues:

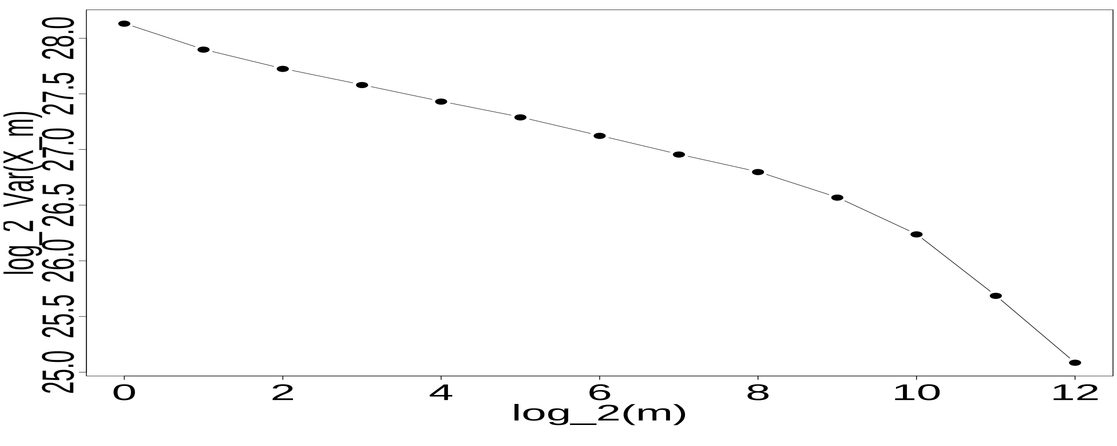
- Range of time scales over which self-similarity is manifested?
- LRD behavior relevant for queueing time-scales?

Modeling Aggregate Traffic as LRD

Time series; 1 minute average rates during 8 hours



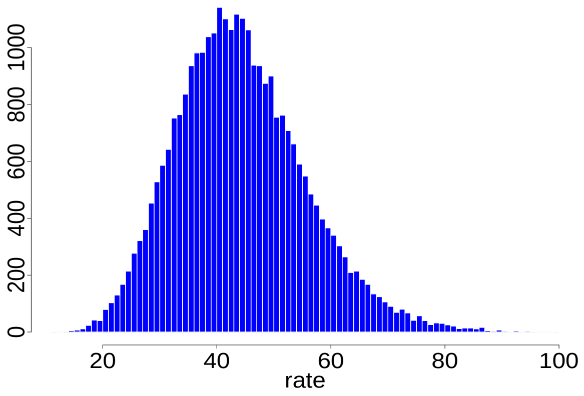
Variance-Time plot:

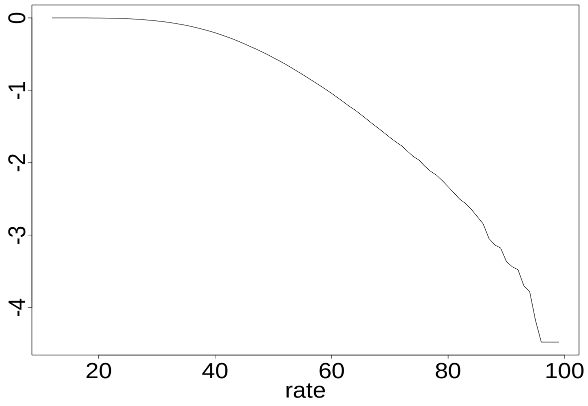


- slope(s) correspond to values of $H > 1/2$
- limitations of variance-time plots
- more structure than simple self-similarity
    - ◇ recent applications of wavelet analysis

Modeling Aggregate Traffic as LRD

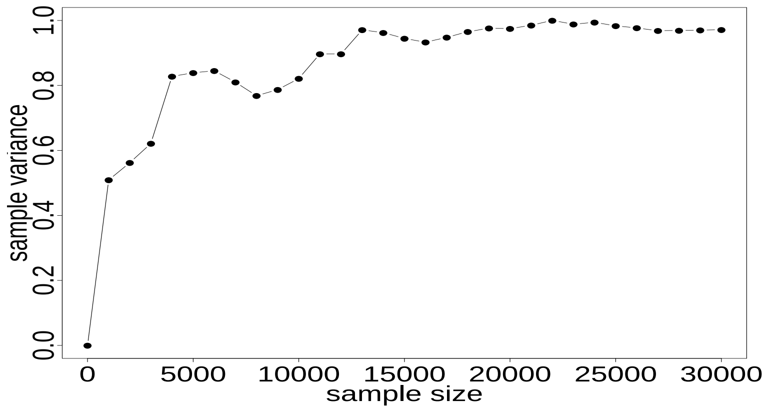
Marginal Distribution





(left plot: histogram) upper tail fatter than Gaussian, but:
(right plot: log tail probability) fast decaying tail

Growth of variance with sample size



Similar behavior with aggregate source video traces

AT&T

On-Off Model and and its Limits

On-Off Model: $(X_i(t); i = 1, 2)$ i.i.d. alternating on-off processes. (Simple model of source data traffic).

(ON,OFF) periods, mean μ_i, CCDF $F_i^c(t) = k_i L_i(t) t^{-\alpha_i}$; $1 < \alpha_i < 2$, L_i slowly varying

Aggregate in space and time:

$$W_{T,M}(s) = \int_0^{Ts} \sum_{i=1}^{M} X_i(u) du$$

Take limits:

- space, then time; set $H = (3 - \min\{\alpha_i\})/2$, $B^H(\cdot) =$ fBM, Hurst parameter H (LRD, but light-tailed marginals)

$$\overset{\text{fd}}{\lim_{T\to\infty}} \overset{\text{fd}}{\lim_{M\to\infty}} \frac{W_{T,M}(t) - TMt\frac{\mu_1}{\mu_1+\mu_2}}{T^H L^{1/2}(T) M^{1/2}} = \sigma^2 B^H(t)$$

- time first, $\Lambda(\cdot)$ stable Levy motion (heavy-tailed marginals)

$$\overset{\text{fd}}{\lim_{T\to\infty}} \frac{1}{T^{1/\min\{\alpha_i\}}} \int_0^{Tt} (X(u) - \mathbb{E}[X(u)]) du = k\Lambda(t)$$

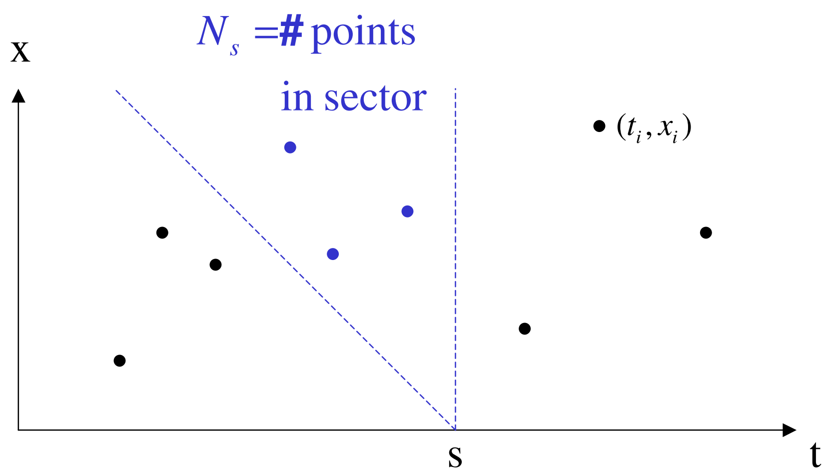
Previous Aggregate Traffic Traces: closer to idealization of first limit

$M/G/\infty$; the Cox model

- Poisson process of arriving sessions, rate r
- i.i.d session durations, cdf G, mean $\gamma < \infty$.
- N_t: number of sessions present at time t

Represent as Poisson process on $\mathbb{R}$ with marks in $\mathbb{R}_+$: (point,mark)$=(t, x)$ means arrival at t with lifetime x.

Poisson Random Measure, mean $r\ dt\ dG(x)$



- N_t Poisson, mean r
- $\mathrm{Cov}(N_s, N_t) = r\gamma G_e^c(|t-s|)$ where G_e is stationary excess distribution of density $\gamma^{-1}G^c(x)$
- G heavy tailed iff N LRD

Some Analyses of LRD Traffic Processes

Regime considered here: LRD rather than *spatial* heavy tails

- assume marginal rates not heavy-tailed. . .
  - ◇ not enough to preclude reasonable aggregation/multiplexing
  - ◇ assume exponential moments $\mathbb{E}[e^{\theta X}]$ finite for some $\theta > 0$
- temporal heavy-tails (call durations, event times, events/call)
  - ◇ limit theorems motivate LRD aggregate processes

Analyses: how to allocate capacity

- what is appropriate notion of quality
  - ◇ degradation: how rare, small and short-lived must it be?
- first-order: understand rate dynamics
  - ◇ how often does rate exceed a given "dangerous" level
  - ◇ recovery times back to "safe" levels
- second order: buffering
  - ◇ queue-length distribution
  - ◇ multiplexing gains in (partially) shared resources

Large Deviation Principles: Motivation

Motivation: How fast is convergence in the WLLN? X_i i.i.d. as X with $\mathbb{E}[X] < \infty$. Sums $S_n = \sum_{i=1}^n X_i$.

$$\mathbb{P}[|S_n - n\mathbb{E}[X]| > n\alpha] \overset{n\to\infty}{\longrightarrow} 0, \quad \forall\alpha > 0.$$

Example: $X_i = \{1, 0\}$ prob. $(p, 1-p)$. For $xn \in \mathbb{N}$

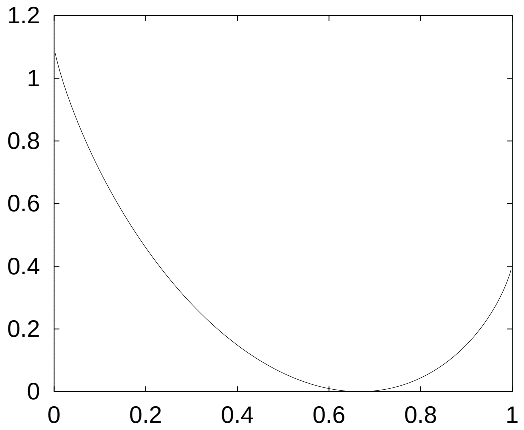
$$\mathbb{P}[S_n = nx] = \frac{n!}{(nx)!(n-nx)!} p^{nx}(1-p)^{n-nx}$$

Use Stirling $\log n! \sim n \log n$ for large n:

$$\log \mathbb{P}[S_n = nx] \sim -n \left(x \log \tfrac{x}{p} + (1-x) \log \tfrac{1-x}{1-p} \right)$$

Convergence: to 0 exponentially fast in n if $I(x) > 0$

Large Deviation Principles: Heuristics



$I(x) = x \log \frac{x}{p} + (1-x) \log \frac{1-x}{1-p}$
with $p = 2/3$.

I convex

$I(x) = 0$ at $x = p = \mathbb{E}[X]$

$I(x) = +\infty$ for $x < 0, x > 1$.

"Rare Events occur in the Most Likely Way"

Example: pick $x \in (0, \mathbb{E}[X])$ (e.g. $x = 0.2$).

Since $I(y) > I(x)$ for all $y < x$, then

$$e^{-nI(y)} \ll e^{-nI(x)}, \quad \text{for large } n$$

So if $S_n < xn$, then $S_n \approx xn$

Principle of the Largest Term (c.f. Laplace)

$$\overline{\lim_n} \frac{1}{n} \log(a_n + b_n) = \max\{\overline{\lim_n} \frac{1}{n} \log a_n, \overline{\lim_n} \frac{1}{n} \log b_n\}$$

Large Deviation Principles: Statement

Setting: $\mathcal{S}$ a topological space, Borel sets $\mathcal{B}$

Definition. *A* **rate function** I *is a lower semicontinuous mapping* $I : \mathcal{S} \to [0, \infty]$. *A* **good** *rate function has compact level sets* $\{x : I(x) \leq \delta\}$.

Definition. *A* **scale** *on* $\mathbb{R}$ *(or* $\mathbb{N}$*) is is a positive monotone function (or sequence) increasing to* $+\infty$.

Definition. *Family of random variables* X_n *in* $\mathcal{S}$ *(*$n \in \mathbb{R}$ *or* $\mathbb{N}$*) satisfies a* **large deviation principle** *with rate function* I *and* **scale** a *if each Borel set* B

$$\begin{aligned} -\inf_{x \in B^o} I(x) &\leq \liminf_{n\to\infty} \frac{1}{a(n)} \log \mathbb{P}[X_n \in B] \\ &\leq \limsup_{n\to\infty} \frac{1}{a(n)} \log \mathbb{P}[X_n \in B] \leq -\inf_{x \in B} I(x) \end{aligned}$$

Consequently: if I continuous on B

$$\lim_{n\to\infty} \tfrac{1}{a(n)} \log \mathbb{P}[X_n \in B] = -\inf_{x\in B} I(x)$$

Informally:

$$\mathbb{P}[X_n \approx nx] \approx e^{-a(n)I(x)}$$

Large Deviation Principles: Example

Definition. *The* **Legendre-Fenchel transform** *of a function f on $\mathbb{R}^d$ is a function f^* on $\mathbb{R}^d$:*

$$f^*(x) = \sup_\theta \{x \cdot \theta - f(\theta)\}$$

Theorem. [Cramer's Theorem]

- *Let X_i be i.i.d. in $\mathbb{R}^d$, sums $S_n = \sum_{i=1}^n X_i$.*
- *Assume the* **log moment generating function** $\lambda(\theta) = \log \mathbb{E}[e^{\theta X_1}]$ *is finite* $\forall \theta \in \mathbb{R}$.

Then S_n/n satisfies an LDP with scale n and rate function λ^*.

Sketch: upper bound. For $\theta > 0$

$$\mathbb{P}[S_n > nx] \leq e^{-\theta n x}\mathbb{E}[e^{\theta S_n}] = e^{-n(x\theta - \lambda(\theta))}$$

Vary θ for least upper bound (for $x > \mathbb{E}[X]$)

$$\mathbb{P}[S_n > nx] \leq e^{-n\lambda^*(x)}$$

Large Deviation Principles: Example

Definition. *Let S_i be a sequence of r.v's in $\mathcal{S}$. Their* **log moment generating function (MGF)** *in scale a is the limit (when it exists)*

$$\lambda(\theta) = \lim_{n\to\infty} \lambda_n(\theta) = \lim_{n\to\infty} \frac{1}{a(n)} \log \mathbb{E}[e^{\theta a(n) S_n}]$$

Definition. *A function $f : \mathbb{R}^d \to \mathbb{R}^d$ is* **steep** *if it is differentiable on the interior $\mathcal{D}^0$ of its effective domain $\mathcal{D} = \{x : f(x) < +\infty\}$, and $\|\nabla f(x_m)\| \to +\infty$ if $\{x_m\}$ converges to the boundary of $\mathcal{D}$.*

Theorem. [Sufficiency: Gärtner-Ellis] *Sequence of r.v.'s $\{S_i\}$ satisfies LDP with rate function λ^*, scale $a(n)$, if MGF $\lambda(\theta)$ exists $\forall\theta$ and is steep. (Steepness only for lower bound)*

Example: Markov Processes. Let X_i form a stationary irreducible MP on finite state space $\mathcal{S} = \{x_1, \ldots, x_{|\mathcal{S}|}\}$. Transition Matrix T. Unique stationary distribution q_i.

Define **twisted transition matrix** $T_{ij}(\theta) = T_{ij} e^{\theta x_j}$

$$\lambda(\theta) = \lim_{n\to\infty} \frac{1}{n} \log \sum_{ij} (T^n(\theta))_{ij} q_j = \log \rho(T(\theta))$$

(Perron-Frobenious: $\rho =$ spectral radius $=$ PF eigenvalue).

Large Deviation Principles: Example

Theorem. [Sanov] *Let $\mathcal{S}$ be a complete separable metric space, let X_i a sequence of i.i.d. r.v's in $\mathcal{S}$ with common distribution ν. The sequence of empirical measures $\rho_n = n^{-1} \sum_{i=1}^{n} \delta_{X_i}$ satisfies an LDP (in the weak topology) with scale n and good convex rate function*

$$I(\rho) = D(\rho, \nu) = \begin{cases} \int d\rho(x) \log \frac{d\rho}{d\nu}(x) & \rho \text{ abs. cont. w.r.t. } \nu \\ \infty & \text{otherwise} \end{cases}$$

$D(\rho, \nu)$ is called **relative entropy** (of ρ w.r.t. ν) or the Kullback–Leibler "distance"

Markov Processes: Donsker-Varadhan have an extension for the empirical measures when X_i are a Markov process

Recovery Times

Rate process (X_t); how long to recover from large excursion?

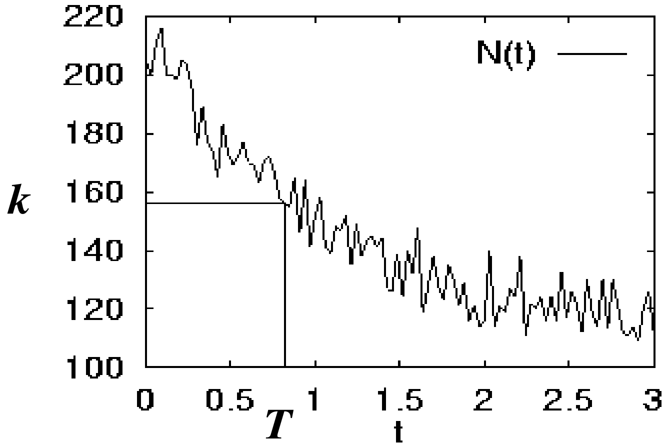
Example: Rate adaptive sources sharing link capacity C

- N_t sources active at time t
- each source has maximum transmission rate m
- allocated rate per source $\min\{m, C/N_t\}$
- quality acceptable for allocated rate $m_- < m$ per source
- how long are excursions of N_t above C/m_-?

Model: $\boldsymbol{M/G/\infty}$ (excursions of $M/M/\infty$ well understood: Aldous, Guillemin et. al.)

Given $N_0 = \ell$, plus possibly other information: **recovery time** from ℓ to k:

$$T_{\ell,k} = \inf\{t \geq 0 : N_t < k | N_0 = \ell, \ldots\}$$



Large Deviation Approximations

Recovery time for mean

$$\tau_{\ell,k} = \inf\{t \geq 0 : \mathbb{E}[N_t] < k | N_0 = \ell\}$$

Deterministic fluid approximation:

- scale $\ell \to n\ell$, $k \to nk$, arrival rate $r \to rn$; $n \to \infty$
- a.s. convergence of $(N_t | N_0 = n\ell)$ to $(\mathbb{E}N_t | N_0 = n\ell)$
- LDP for sample paths of $(N_t | N_0 = n\ell)$
- corresponding convergence of $T_{n\ell,nk}$ to $\tau_{\ell,k}$
- set $m = r\gamma$, $\gamma =$mean(G), $G_e(t) = \gamma^{-1} \int_0^t du\ G^c(u)$

Recovery of mean:

$$\mathbb{E}[N_t | N_0 = \ell] = m + (\ell - m) G_e^c(t)$$

Recovery time of mean:

$$\tau_{\ell,k} = G_e^{-1}\left(\frac{\ell - k}{\ell - m}\right)$$

Conditional variance also tractable

$$\text{Var}[N_t | N_0 = \ell] = m G_e(t) + \ell G_e(t) G_e^c(t)$$

Information and Conditioning

Assume known session ages $x_1, x_2, \ldots, x_{N_0}$

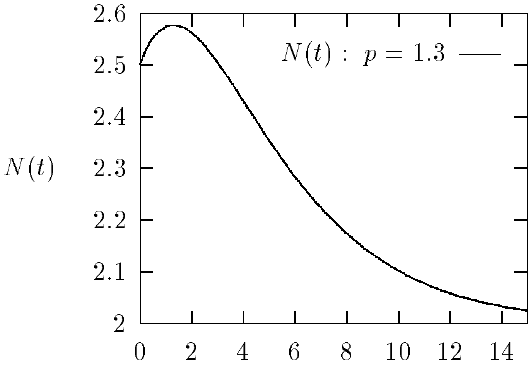
Conditioned Evolution:

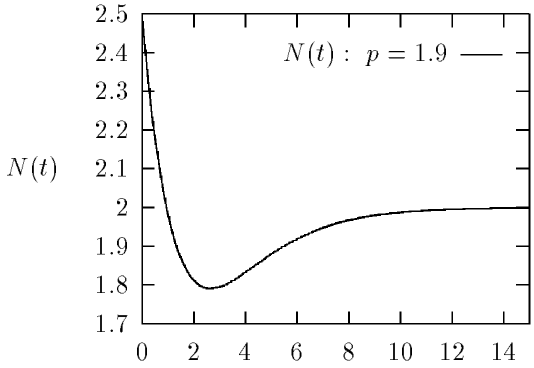
$$\mathbb{E}[N_t | N_0 = \ell, x_1, \ldots x_\ell] = mG_e(t) + \int \rho_\ell(dx) H_x^c(t)$$

$H_x^c(t) = \frac{G^c(x+t)}{G^c(x)}$ CCDF of residual life t, current age x
$\rho_\ell = \sum_{i=1}^{l} \delta_{x_i}$ empirical age measure

Effect of initial age distribution

e.g. lifetime distribution G Pareto, $G^c(x) \sim x^{-\alpha}$, $\alpha = 1.5$





(Left) initial ages Pareto, $\alpha = 1.3$, ages long
(Right) initial ages Pareto, $\alpha = 1.9$, ages long

Asymptotics of Age Distribution

Normalized empirical age measure ρ/m: LDP, scale m, rate function

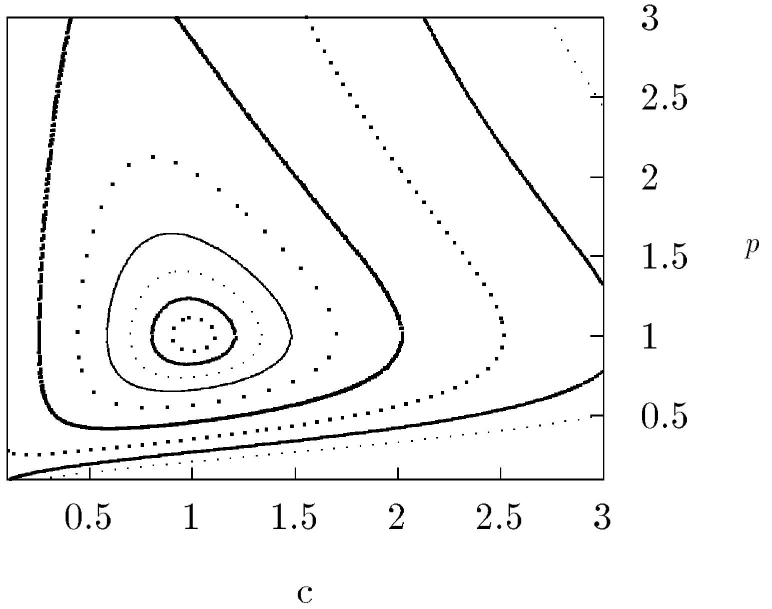
$$I(\nu) = J(\overline{\nu}) + \overline{\nu} D(\nu/\overline{\nu}, G_e)$$

- $\overline{\nu} =$ mass of ν
- $J(x) = x \log x + (1 - x)$ (r.f. for LDP of Poisson$(r)/r$)
- D is relative entropy (through Sanov's theorem)

Interpretation:

- LD in current age distribution and #sessions equally likely
- Can be worth exploiting current ages for controls

Example: contours of rate function $J(c) + cD(f(p), f(1))$ where $f(p)$ is Pareto, index p.



Arrivals, Service, Waiting, and Queueing

(Doubly infinite) sequence of customers, served FIFO at queue

- R_n service requirement of customer n ($n \in \mathbb{Z}$).
- S_n interarrival time between customers n and $n+1$
- V_n waiting time of customer n to start service

These are related in Lindley's equation: set $X_n = R_n - S_n$

$$V_n = \max\{0, V_{n-1} + X_{n-1}\}$$

Iterating with $Y_{m,n} = \sum_{i=m}^{n-1} X_i$:

$$V_n = \max\{0, Y_{n-1,n}, Y_{n-2,n}, \ldots, Y_{n-j+1,n}, Y_{n-j,n} + V_{n-j}\}$$

Regard $W_n = \sup_{j \geq 0} Y_{n-j,n}$ as the wait of customer n.

Theorem. *Assume (X_n) stationary and ergodic, with $\mathbb{E}[X_n] < 0$. Then (W_n) is the unique stationary process for which $W_n = \max\{0, W_{n-1} + X_{n-1}\}$. W_n is a.s. finite.*

W_0 is max of walk with steps (X_i). $E[X_i] < 0$ means -ve drift.

For singly infinite sequences $(V_n, X_n; n \geq 0)$ with init. cond. $V_0 = v$, V_n converges to W_0 in distribution as $n \to \infty$.

Large Deviations, Arrivals & Queueing

Setting:

- Queue, infinite buffer, service rate c
- Arrivals A_t in $[-t, 0)$ or $\{-t, \ldots, -1\}$. $A_0 = 0$.
- (Virtual) queue length $Q = \sup_{t>0}(A_t - ct)$ (c.f. Lindley)

Motivation:

- Generalize: from previous work using Markov Methods
  - ◇ classes of short and long-range dependent traffic models
- Relaxed assumptions:
  - ◇ LD axioms, rather than model classes
  - ◇ Assume A_t satisfies some LDP for large t (as for Markov)
- Observe: large Q needs sustained arrivals above mean
- Derive: LDP for Q from LDP for A.

LDP for A: picking the right scale

- Markov: $\lambda_t(\theta) = t^{-1} \log \mathbb{E}[e^{\theta t \cdot A_t/t}]$; limit MGF finite
  - ◇ MGF for A_t/t with scale t.
- Fractional Brownian motion, Hurst parm $H \in (1/2, 1)$.
  - ◇ $\mathbb{E}[A_t = at]$, Var$[A_t] = \sigma^2 t^{2H}$

$$t^{-1} \log \mathbb{E}[e^{\theta t \cdot A_t/t}] = a\theta - \theta^2 t^{2H-1}/2 \to \infty$$

  - ◇ Instead: MGF for A_t/t with scale $v(t) = t^{2-2H}$

$$v(t)^{-1} \log \mathbb{E}[e^{\theta v(t) \cdot A_t/t}] = a\theta + \sigma^2\theta^2/2 \quad \text{finite limit}$$

Large Buffer Asymptotics for Q

Assumptions: (Service rate c)

(i) Arrivals A_t with stationary increments, $t \in \mathbb{N}$.
(ii) Scale $v(t) = t^{\tilde{v}}$, $\tilde{v} \in (0,1)$ $(H = 1 - v/2 \in (1/2, 1))$
  - ◇ More general OK: e.g. $v(t)t^{-\tilde{v}} \to 1$ or v regularly varying

(iii) MGF λ exists for scale v, and is steep.
  - ◇ $\Rightarrow$ LDP for A_t/t, scale v, rate function λ^*

(iv) $\lambda(\theta) < c\theta$ some $\theta > 0$
  - ◇ Stability: $\Rightarrow \mathbb{E}[A_t] < ct$

Theorem. *Under these assumptions*

$\lim_{b\to\infty} \frac{1}{v(b)} \log \mathbb{P}[Q > b] = -\delta := -\inf_{x>0} x^{-\tilde{v}} \lambda^*(c+x)$

Lower bound:

$$
\begin{aligned}
b^{-\tilde{v}} \log \mathbb{P}[Q > b] &= b^{-\tilde{v}} \log \mathbb{P}[\cup_t \{A_t - ct > b\}] \\
&\geq b^{-\tilde{v}} \log \mathbb{P}[A_{b/x} - cb/x > b], \quad b/x \in \mathbb{N} \\
&= x^{-\tilde{v}} (b/x)^{-\tilde{v}} \log \mathbb{P}[A_{b/x} > (b/x)(c+x)]
\end{aligned}
$$

$$\liminf_{b\to\infty} b^{-\tilde{v}} \log \mathbb{P}[Q > b] \geq -\delta$$

Intepretation: Minimizer x^* picks out most likely time scale $t^* = b/x^*$ for excursion of $A_t - ct$ to b.

Impact of LD Scaling

Assume: corresponding upper bound, approximate

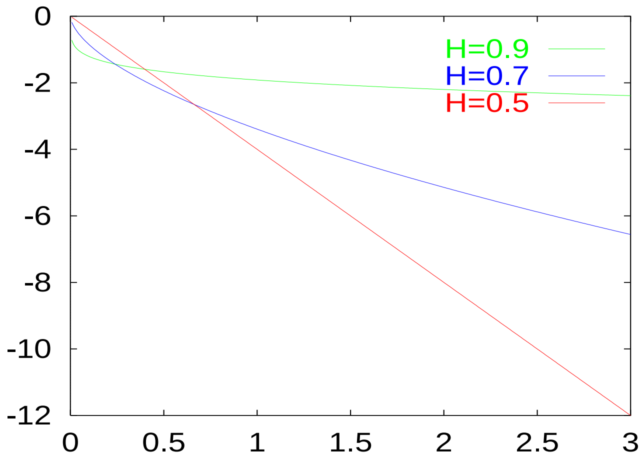
$$\mathbb{P}[Q > b] \approx e^{-\delta b^{-\tilde{v}}}$$

Example: A_t Markov, $v(t) = t$: $P[Q > b]$ decay exponential

Example: fBM. $\mathbb{E}[A_t] = at < ct$, $\mathrm{Var}(A_t) = \sigma^2 t^{2H}$, $\tilde{v} = 2 - 2H$, $\lambda(\theta) = a\theta + \sigma_2 \theta^2/2$

$$\begin{aligned} \lambda^*(x) &= (x-a)^2/(2\sigma^2) \\ \delta &= \left(\frac{(c-a)^H}{\sqrt{2}\sigma H^H (1-H)^{1-H}} \right)^2 \end{aligned}$$

Impact:
Weibull decay for $H > 1/2$; δ deceasing in H: longer queues.



$-\delta b^{2-2H}$ vs. b

Large Deviation Upper Bound

Discrete time: for some $t_b > 0$,

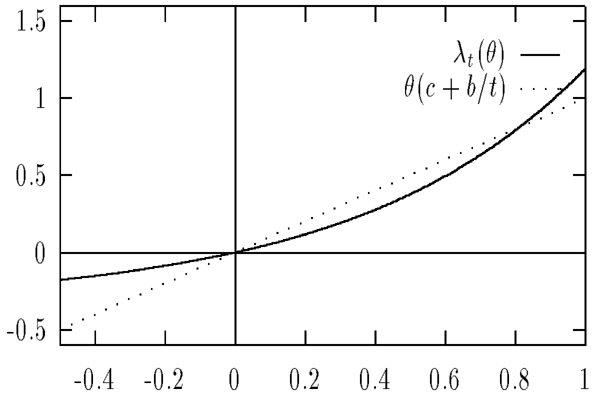
$$\mathbb{P}[Q > b] \leq t_b \sup_{t<b/d} \mathbb{P}[A_t > ct + b] + \sum_{t \geq t_b} \mathbb{P}[A_t > ct + b]$$

Use Chebychev: for $\theta > 0$

$$\begin{aligned} (1)\ \mathbb{P}[A_t > ct + b] &\leq e^{-v(t)(\theta(c+b/t)-\lambda_t(\theta))} \quad \text{and hence} \\ \mathbb{P}[A_t > ct + b] &\leq e^{-v(t)\lambda_t^*(c+b/t)} \end{aligned}$$

Could extend (1) to all θ

since $ct + b > at$.



First Term: $\leq t_b e^{-I(b)}$ where $I(b) = \inf_t v(t)\lambda_t^*(c + b/t)$

Will show that $\lim_{b\to\infty} I(b)/v(b) = \delta$.

LD Upper Bound: Continued

Second Term: by assumption (iv), pick $\theta > 0$ with $\lambda(\theta) < c\theta$ and hence $\lambda_t(\theta) < c\theta$ for $t > t_0$.

For $t_b > t_0$,

$$\sum_{t \geq t_b} \mathbb{P}[A_t > ct + b] \leq \sum_{t > t_b} e^{-\theta v(t) b / t_b}$$

Choose $t_b = \max[kb, t_0]$, (using incomplete gamma function):

$$\limsup_{b \to \infty} b^{-\tilde{v}} \log \sum_{t > t_b} e^{-\theta t^{-\tilde{v}} b / t_b} = -\theta k^{\tilde{v}}$$

Both Terms Together: using Principle of Largest Term

$$\begin{aligned}
\overline{\lim_b}\, b^{-\tilde{v}} \log \mathbb{P}[Q > b] &\leq \overline{\lim_b}\, b^{-\tilde{v}} \log \left(t_b e^{-I(b)} + \text{Second Term} \right) \\
&\leq -\min\{\underline{\lim}_b I(b)/v(b)\ ,\ \theta k^{\tilde{v}}\} \\
&= -\underline{\lim}_b I(b)/v(b) \quad (k \text{ large enough})
\end{aligned}$$

LD Upper Bound: Concluded

Remains: to show $\underline{\lim}_b I(b) b^{-\tilde{v}} = \delta := \inf_x x^{-\tilde{v}} \lambda^*(c+x)$

Define: $f_t(x) = \lambda^*_{t^{1/\tilde{v}}}(c + x^{1/\tilde{v}})$ for $x \geq 0$, else $+\infty$.

Properties of f, f_t

(i) f_t, f convex. ($\tilde{v} \in (0,1) \Rightarrow x^{1/\tilde{v}}$ convex on $[0,\infty)$, $\uparrow$)
(ii) f_t, f closed. (Legendre transform $\times I_{[0,\infty)}$)
(iii) $f_t \to f$ in $\mathcal{D}^o(f)$, $f_t^* \to f^*$ in $\mathcal{D}^o(f^*)$. (Since $\lambda_t \to \lambda$)
(iv) f^* increasing. ($= \sup_{x>0}\{x\theta - f(x)\}$: $f(x<0) = \infty$)
(v) $\delta = \inf_{x>0} f(x)/x$. (rescale $x \to x^{1/\tilde{v}}$)
(vi) $f^*(\delta) < 0$. ((v) means $\delta x \leq f(x)$ all x)
(vii) $f_t^*(\delta') < 0$ for any $\delta' < 0$, $t > t_{\delta'}$. (By (iii), (iv), (vi))
(viii) $\lim_x f_t(x)/x = \infty$. (Convexity $\lambda_t(c+x) \geq k_{1,t} + k_{2,t} x$)
(ix) $f_t^{**} = f_t$. (By (i), (ii))

Lower bound:

$$\begin{aligned}
\underline{\lim}_b b^{-\tilde{v}} I(b) &= \underline{\lim}_x \inf_t \tfrac{f_t(x/t)}{x/t} \\
(\text{any } \delta' < \delta) &= \underline{\lim}_x \min\{\min_{t \leq t_{\delta'}} \tfrac{f_t(x/t)}{x/t}, \inf_{t > t_{\delta'}} \{\delta' - \tfrac{f_t^{**}(\delta')}{x/t}\}\} \\
&\geq \delta' \qquad \text{(by (vii), (viii), (ix))}
\end{aligned}$$

Upper Bound:

$$\overline{\lim}_b \frac{I(b)}{b^{\tilde{v}}} \overset{b=tx}{\leq} \inf_{x>0} \overline{\lim}_b \frac{\lambda_t^*(b/x, c+x)}{x^{\tilde{v}}} \overset{\text{(iii)}}{=} \inf_{x \in \mathcal{D}^o(\lambda^*)} \frac{\lambda^*(c+x)}{x^{\tilde{v}}} = \delta$$

Effective Bandwidths

Linear LD Scaling:

- defined as $\sigma(\theta) = \lambda(\theta)/\theta$, $\theta > 0$
- incrsng $(0, \infty) \to (\mathbb{E}[A_1], \operatorname{ess\,sup}(A_1)) =$ (mean,peak)

$$\theta < \delta \Leftrightarrow \sigma(\theta) < c \Leftrightarrow \lim_{b\to\infty} b^{-1} \log \mathbb{P}[Q > b] > -\delta$$

Interpretation: $\sigma(\theta)$ as rate for loss prob $e^{-b\delta}$ from buffer b.

Additivity: across independent sources:

$$\sigma_{\sum_i A_i}(\theta) = \frac{\log \mathbb{E}[e^{\theta \sum_i A_{i,t}}]}{\theta} = \sum_i \frac{\log \mathbb{E}[e^{\theta A_{i,t}}]}{\theta} = \sum_i \sigma_{A_i}(\theta)$$

Linear Constraints: for admission; aggregate N_i connections of type i, service rate c. Asymptotic loss rate better than θ if

$$\sum_i N_i \sigma_i(\theta) < c$$

Pitfalls for LD with Large Buffers

Cox Model: $M/G/\infty$ Arrivals:

- Poisson process of arriving sessions, rate r
- i.i.d session durations, cdf G, mean $\gamma < \infty$.
- each session empties fluid into infinite buffer at unit rate
- buffer drain rate c

Intrinsic Large Deviation Scaling:

- Residual life $G_e(x) = \gamma^{-1} \int_x^\infty G^c(x)$
- Set scale $v(t) = -\log G^c(t)$
  - ◇ G, G_e exponential, $v(t) \propto t$
  - ◇ G Pareto, $G^c(t) \sim t^{-\alpha}$, $v(t) \sim (\alpha - 1)\log t$

Theorem. *Suppose G_e vanishes slower than exponential: $t^{-1}\log G_e^c(t) \to 0$ as $t \to \infty$. Then MGF of arrivals A_t/t exists in scale v:*

$$\lim_t \frac{1}{v(t)} \log \mathbb{E}[e^{\theta v(t) A_t/t}] = \lambda(\theta) = \begin{cases} r\gamma\theta & \theta < 1 \\ +\infty & \theta > 1 \end{cases}$$

(Limit at $\theta = 1$ depends on finer properties).

Note λ is **not steep**. Cannot use Gärtner-Ellis alone to establish LD lower bound for queue-length. (Many other approaches to bounds for this model).

Brief Look at Sample Path Methods

Single Server Queue:
Dembo-Zajic extend Mogulskii's to wider classes of processes.

Consider as before process $(X_i; i \in \mathbb{N})$; partial sums process

$$Z_n(t) = \frac{1}{n}\sum_{i=1}^{k} X_i; \quad \frac{k-1}{n} \le t < \frac{k}{n}$$

Assume that for all $0 = t_0 < t_1 < \ldots < t_m \le 1$, the vector

$$V_n = (Z_n(t_i) - Z_n(t_{i-1}) : 1 \le i \le m)$$

satisfies LDP with scale n and rate function

$$I_m(v_1, \ldots, v_m) = \sum_{1 \le i \le m}(t_i - t_{i-1})I(v_i)$$

I is r.f. for $Z_n(1)$. Then Z_n has LDP with rate function

$$\mathcal{I}(\phi) = \begin{cases} \int_0^1 I(\dot{\phi})ds & \phi \in \mathcal{AC}[0,1] \\ \infty & \text{otherwise} \end{cases}$$

Interpretation of Conditions:
Form of rate function (sum over disjoint intervals) to reflect asymptotic independence of increments of $Z_n(\cdot)$

MGF for linear scale: limit of $\lambda_t(\theta) = t^{-1}\log \mathbb{E}[e^{\theta A_t}]$ as $\to \infty$

$$\begin{aligned} \lambda_{2t}(\theta) &\approx \lambda_t(\theta) \\ \mathbb{E}[e^{\theta A_{0,2t}}] &\approx \mathbb{E}[e^{\theta A_{0,t}}]\mathbb{E}[e^{\theta A_{t,2t}}] \qquad \text{asymptotic independence} \end{aligned}$$

Queueing and Sample Path Analysis

Sample path LDP: holds with $I = \lambda^*$ (under mixing conditions).

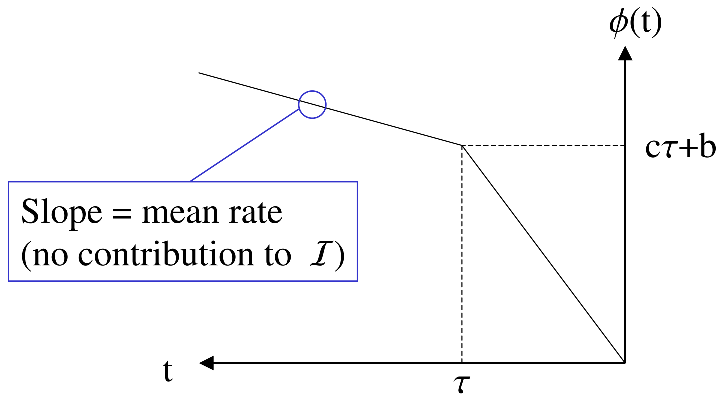
Queue length $Q = \sup_{t>0}\{\phi(t) - ct\}$

Sample path rate function for $\{Q = b\}$. Contraction Principle:

$$\begin{aligned}
J(b) &= \inf\{\int_0^\infty \lambda^*(\dot\phi) \mid \sup_{t>0}\{\phi(t) - ct\} = b\} \\
&= \inf_{\tau>0} \inf\{\int_0^\infty \lambda^*(\dot\phi) \mid \phi(\tau) - c\tau = b\} \\
&= \inf_{\tau>0} \tau\lambda^*(c + b/\tau) = b\delta
\end{aligned}$$

Convexity of λ^* $\Rightarrow$ approximate using piecewise linear paths.

Most likely path:



Further Developments

Large Deviations Bound and Limits For empirical means (Glynn-Whitt, Whitt, Kesidis-Walrand-Chang, Chang, Duffield-O'Connell, Norros)

- Departure Processes and Networks
  - ◇ decoupling conditions: when do streams not effect each others asymptotic properties in a shared resource
- Beyond FIFO: limited separation of resources
  - ◇ Generalized Processor Sharing, Polling, Round Robin

(O'Connell, Z. Zhang et al, Massoulie, Majewski, Bertsimas-Paschilidis-Tsitsiklis, Duffield, Weber)

A General Approach:

- LD hypothesis for exogenous traffic process X
- establish continuity of desired quantity Y as function of X
- use contraction principle to find LDP for Y
- issue: computability of rate function
  - ◇ high dimensional variation problems
  - ◇ scope for numerical-computational methods?

LRD processes

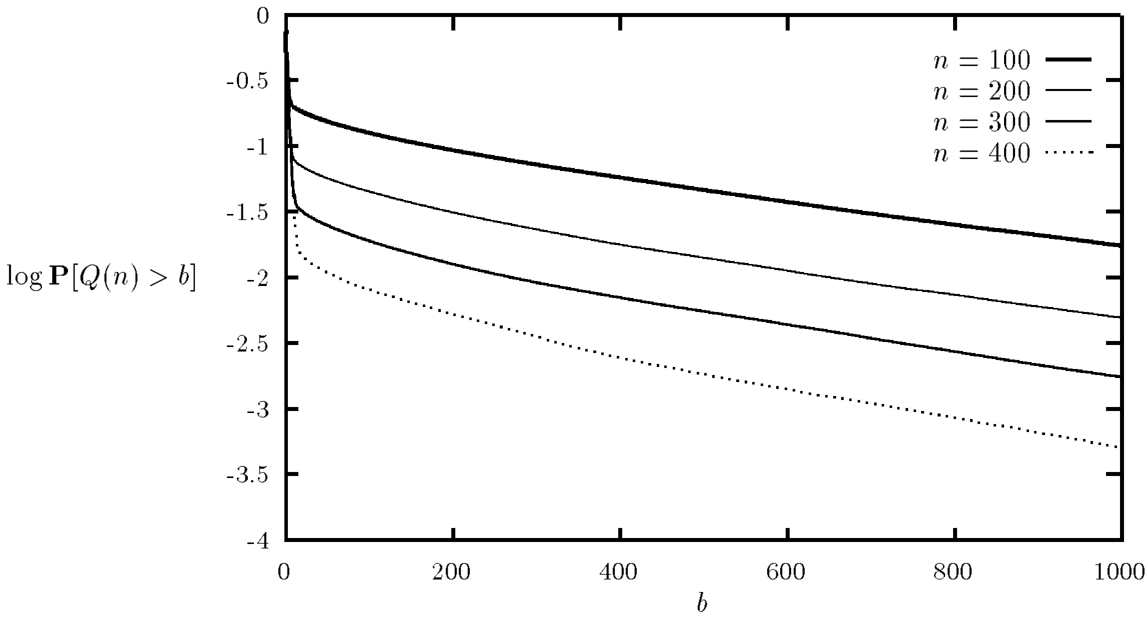
- Sample path LDP of LRD processes (Chang-Yao-Zajic)
- propagation of LRD by leaky bucket (Vamvakos-Anantharam)

Multiplexing and Aggregation

How good is the approximation $\mathbb{P}[Q > b] \approx e^{-\delta b^{\tilde{v}}}$?

Setting:

- aggregation $A_t(n) = \sum_{i=1}^{n} A_{i,t}$
- service rate cn (offered load independent of n)
- $Q(n) = \sup_{t>0}\{A_t(n) - cnt\}$



Observe

- $\mathbb{P}[Q(n) > nb]$ increases roughly linearly with n
- multiplexing gains across sources
- purely large buffer approximation worse as n increases

Many Sources LD Asymptotic

Theorem. *Under same assumptions as large buffer asymptotic*

$$\lim_n \tfrac{1}{n} \log \mathbb{P}[Q(n) > nb] = -I(b)$$

Lower Bound: Set $\mu_t(\theta) = \log \mathbb{E}[e^{\theta A_t}]$ $(= v(t)\lambda_t(\theta t / v(t))$.

For each t, $A_t(n)$ is i.i.d sum: $A_t(n)/n$ satisfies LDP, scale n, rate function $\mu_t^*(x) = v(t)\lambda_t^*(x/t)$.

$$\begin{aligned}
\underline{\lim}_n \tfrac{1}{n} \log \mathbb{P}[Q(n) > nb] &\geq \underline{\lim}_n \sup_t \tfrac{1}{n} \log \mathbb{P}[A_t(n) > n(ct+b)] \\
&\geq \sup_t \underline{\lim}_n \tfrac{1}{n} \log \mathbb{P}[A_t(n) > n(ct+b)] \\
&= -\inf_t \mu_t^*(ct + b^+) = -I(b^+)
\end{aligned}$$

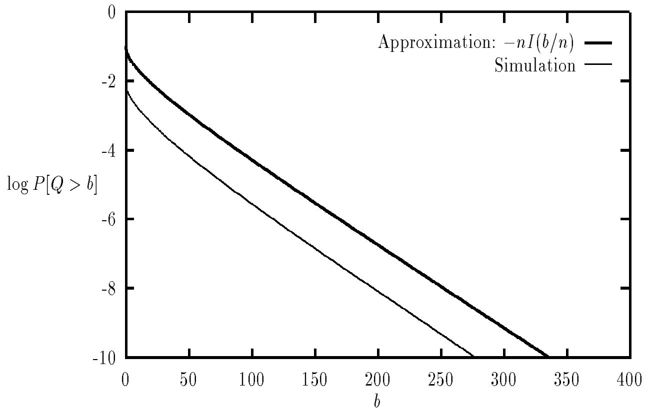
Upper Bound:

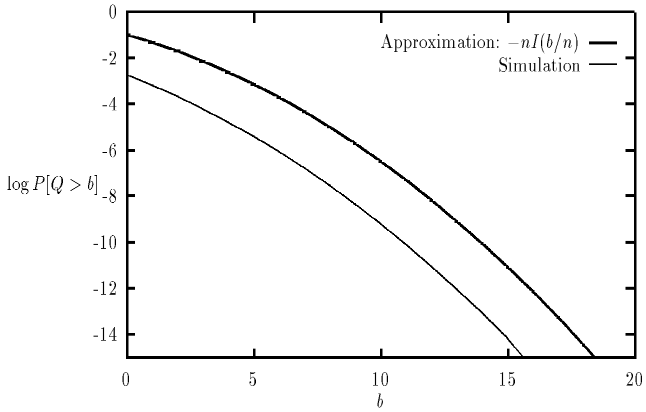
- proof similar to large buffer case
- role of scaling function v
  - ◇ require $\exists \lambda_t = \lim_t \lambda_t$ to control tail sum

Domain: can be established under weaker assumptions, including e.g. Cox model ($M/G/\infty$ with G heavy tailed).

Asymptotic Behavior: have shown before that $I(b)/v(b) \to \delta$

Examples





Asymptotic Form: SRD case: $I(b)/b \to \delta$ with $\lambda(\delta) = s\delta$

$$I(b) - \delta b \to \nu := - \lim_{t\to\infty} t(\lambda_t(\delta) - s\delta)$$

In examples: relative to large buffer asymptotic

- (Left) $\nu > 0$, economies of scale
- (Right) $\nu < 0$ negative economies of scale

Sign of $\boldsymbol{\nu}$: related to correlation structure.

Example: 2-state Markov traffic, discrete time

$$\text{Transition Matrix} \quad \begin{pmatrix} 1-a & d \\ a & 1-d \end{pmatrix}$$

$a + d < 1 \Leftrightarrow A_t$ increments +vely correlated $\Leftrightarrow \nu > 0$

Generally: A_t increments associated $\Rightarrow \nu > 0$

Further Developments

Can relax i.i.d. assumptions; only require $\forall t > 0,\ \exists \lim_n n^{-1} \log \mathbb{E}[e^{\theta A_t(n)}]$.

Many Sources Asymptotic: Weiss, Botvich-Duffield, Duffield, Courcoubetis-Weber, Simonian-Guibert

Large Deviation results:

- $\sigma(t,\theta) = (t\theta)^{-1} \log \mathbb{E}[e^{\theta A_t}]$ as a generalized effective bandwidth. Local properties of admission curve. (Kelly)
- Bandwidth-buffer trade-offs, optimality for multiplexing, many sources asymptotic for output processes (Duffield-Low)
- Bahadur-Rao corrections to many-sources asymptotic (Montgomery-de Veciana)
- Finite-buffer corrections (Likhanov-Mazumdar)
- Sample-path properties (Mandjes, Wischik)

Other Approaches:

- Gaussian Processes: Extreme-value methods reproduce shape-function approximation for single streams (Choe-Schroff)

Measurement Based Admission Control

Motivation: Statistical Multiplexing in Networks.

- balance opposing constraints
  - ◇ use quality (loss, delay, . . .)
  - ◇ network utilization
- connection admission control
  - ◇ limit admission to preserve quality of admitted sources

Requirements

- characterization of traffic
- rule to determine quality from characterization

Finding Traffic Characterizations

- from modeling?
  - ◇ model parameters not usually supplied
  - ◇ static models not robust w.r.t underlying changes in traffic
- from measured traffic parameters
  - ◇ need not be declared explicitly
  - ◇ can adapt to changes in traffic

Large Deviations and MBAC

Traffic Characterization with MGF

- no specific model class assumed
- efficient description of small probabilities
- robust w.r.t variation in traffic
- computationally feasible

Estimating the MGF

- traffic sample $(X_i)_{i=1,2,\ldots T}$ (rates over T slots)
- divide into n blocks of length t

$$Y_{t,j} = \sum_{i=(j-1)t+1}^{tj} X_i$$

- estimate MGF by

$$\widehat{\lambda}_t(\theta) = t^{-1} \log n^{-1} \sum_{i=1}^{n} e^{\theta Y_{j,t}}$$

- sampling issues
  - ◇ fewer sample points for larger t: increased sample variance
  - ◇ measure aggregate or individual connections?
  - ◇ effect of natural statistical variability of estimates

Admission Control:

Admit sources while estimated loss rate $<$ target ε

Predict loss rates with estimated $\widehat{\delta}$ or shape function $\widehat{I}$.

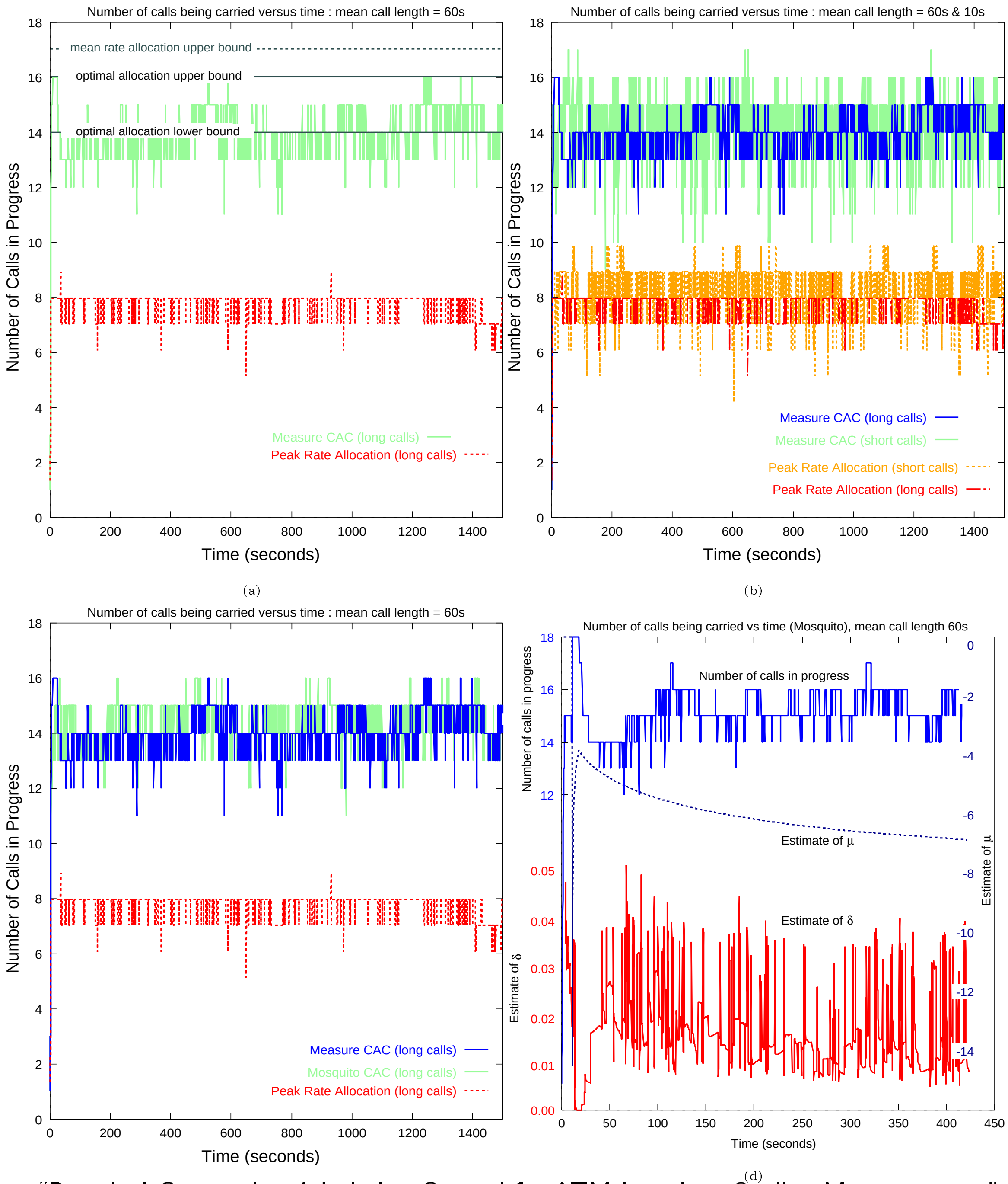
- large buffer regime, linear scale: admit to buffer size b while
$$\exp(-\widehat{\delta} b) < \varepsilon$$
- shift to account for multiplexing gains (martingale bound)
$$\exp(-\widehat{I}(0) - \widehat{\delta} b) < \varepsilon$$
- many source asymptotic
$$\exp(-\widehat{I}(b)) < \varepsilon$$

Enhancement:

Use declared peak rate p to bound $\widehat{\lambda}$ above as

$$\widehat{\lambda}'(\theta) = \begin{cases} \widehat{\lambda}(\theta) & \theta \leq \theta_0 \\ \widehat{\lambda}(\theta_0) + p(\theta - \theta_0) & \theta > \theta_0 \end{cases}$$

where $\theta_0 = -\log$(minimum observable cell-loss)/ buffer_size



From "Practical Connection Admission Control for ATM based on On-line Measurements", by Lewis, Russell, Toomey, McGurk, Crosby and Leslie, to appear in Computer Communications

Sampling Properties in MBAC

Potential source of errors:

- poor modeling of details of loss process
  - ◇ e.g. network element more complex than single buffer
- loss ratio only approximated
  - ◇ LD gives only logarithmic asymptotics
  - ◇ loss prediction in pre-asymptotic region?
  - ◇ infinite buffer idealization?
- Even if these are OK: role of sampling error

Impact of Sampling Errors

- Certainty Equivalence: assumes measured parameters are true
- samples are random: so is number of admitted connections
- two sources of randomness in admitted traffic
  - ◇ variability of individual connections
  - ◇ variability of number connections admitted
- attained loss rates > expected loss rates
- analyze how CAC must "aim high" to attained desired loss
  - ◇ Grossglauser-Tse analyzed Gaussian bufferless case:
  - ◇ measurement uncertainty double effect of bandwidth variance in heavy traffic limit

LD analysis of sampling errors in MBAC

Model CAC to bufferless systems

- bufferless resource, capacity nc, target overflow prob $e^{-n\varepsilon}$
- connections seeking admission offer i.i.d. bandwidth samples X_i, distribution measure ν
- use samples to form empirical distribution

  $\widehat{\rho}_n = n^{-1} \sum_1^n \delta_{X_i}$ empirical measure
- empirical mean $\hat{a}_n = a_{\hat{\rho}_n}$ where $a_\rho = \int \rho(dx)x$
- empirical MGF $\hat{\mu}_n = \mu_{\hat{\rho}_n}$ where $\mu_\rho = \log \int \rho(dx)e^{\theta x}$
- LD estimate of overflow rate of N admitted sources

$$P[\sum_{i=1}^{N} X_i \geq nc] \approx e^{-(N\widehat{\mu}_N)^*(n \max[c,\widehat{a}_N])}$$

- admit

$$\hat{N}_n = \inf\{N : (N\widehat{\mu}_N)^*(n \max[c, \widehat{a}_N]) < n\varepsilon\} - 1$$

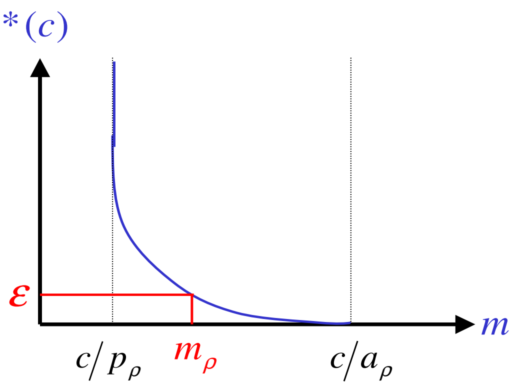
- will show LDP for $\hat{N}_n$

LDP for #Admitted Connections $\widehat{N}_n$

- rewrite $\hat{N}_n = \lfloor n\hat{m}_n \rfloor$ where $\hat{m}_n = m_{\hat{\rho}n}$ (with empirical peak rate $p_\rho = \sup \mathrm{supp}(\rho)$)

$$m_\rho = \begin{cases} \text{! soln. in } (c/p_\rho, c/a_\rho) \text{ of } (m\mu_\rho)^*(c) = \varepsilon & \text{if } \varepsilon < ((c/p_\rho)\mu_\rho)^*(c) \\ c/p_\rho & \text{otherwise} \end{cases}$$

When X_i are bounded can show that $\rho \mapsto m_\rho$ is weakly continuous.



- Sanov's Theorem: Empirical measures ρ_n satisfy LDP in weak topology with scale n and rate function (relative entropy)

$$\rho \mapsto D(\rho, \nu) = \int d\rho \log \frac{d\rho}{d\nu}$$

- Contraction Principle: $\widehat{N}_n/n$ has LDP, scale n, rate function

$$J(x) = \inf_{\rho : m_\rho = x} D(\rho, \nu)$$

- Simplfy: can show $J(x) = \inf_\theta \lambda_\theta^*(e^{(c\theta - \varepsilon)/x})$ where

$$\lambda_\theta(\phi) = \log \mathbb{E}[e^{\theta e^{\theta X_i}}]$$

LDP for Admitted Aggregate

Setting:

- Admitted traffic: Y_i, ident. dist. as X_i but independent
  - ◇ stationary connections, values Y_i long after admission
- admitted aggregate $\widehat{S}_n = \sum_{i=1}^{\widehat{N}_n} Y_i$

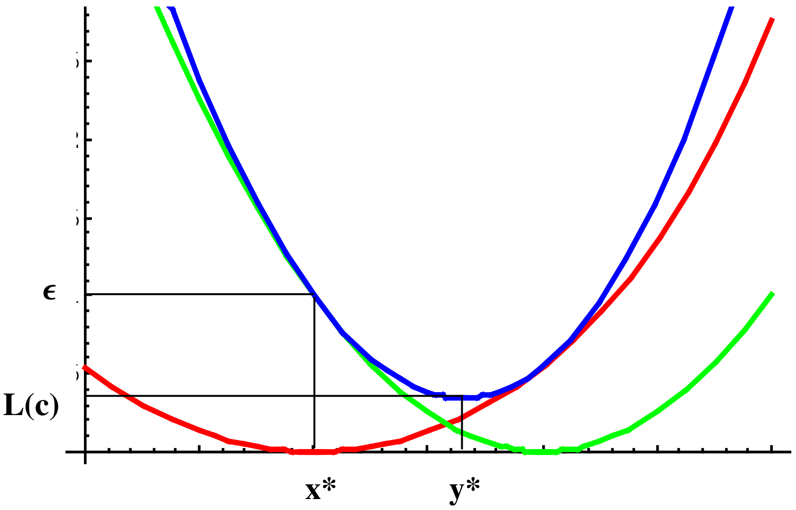
Theorem. *Assuming X_i have bounded support, $\widehat{S}_n/n$ satisfies LDP, scale n, rate function*

$$L(x) = \inf_y \{J(y) + y\mu_\nu^*(x/y)\}$$

Interpretation:

- $\mathbb{P}[\widehat{S}_n \approx nx] \approx \sup_y e^{-nJ(y)} e^{-ny\mu_\nu^*(x/y)}$
- $e^{-nJ(y)}$ is $\approx$ prob. to admit ny connections
- $e^{-ny\mu_\nu^*(x/y)}$ is $\approx$ prob. these have aggregate rate $\approx nx$
- extremum picks out most likely way for this to happen

Illustration (Bernoulli Connections)



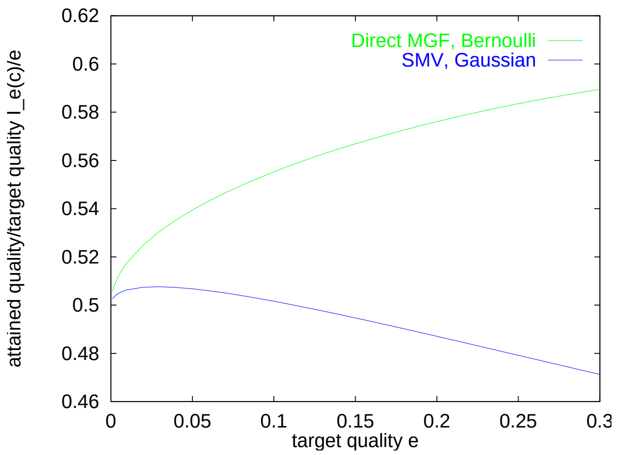
Connections Admitted:
J(x*) = 0; with true MGF,
would admit x* flows

Admitted Aggregate:
yμ*(c/y) = ϵ (target loss rate)
when y = x*

Sum J(x) + y μ*(c/y) has min
value I(c) < ϵ at some y* > x*

- measurement error
- over-optimistic admission of connections: $y^* > x^*$
- higher than expected loss $e^{-L(c)} > e^{-\varepsilon}$

Behaviour for small target loss



◇ $L(c)/\varepsilon \xrightarrow{\varepsilon \to 0} 1/2.$

◇ recovers Grossglauser-Tse

◇ no consistent sign of diff

AT&T

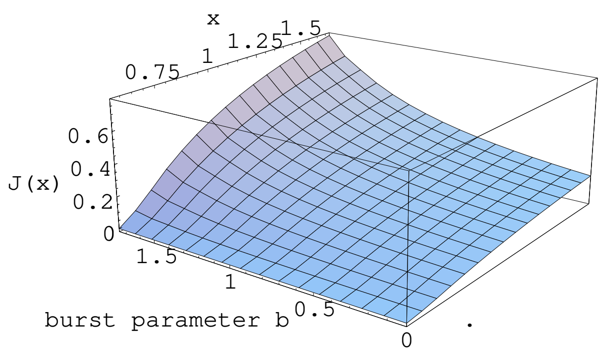
Extensions

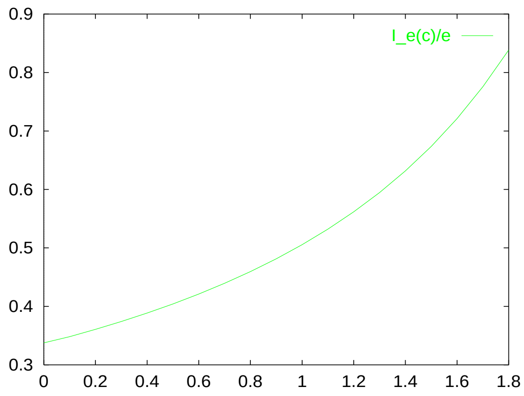
Buffered Resources

- scale buffer nb, use many sources asymptotic
- connections i present sample arrivals process $(A_{i,t})$
- decomposes over time-scales using empirical shape function
- admit $\widehat{N}_n = \inf_t \widehat{N}_{n,t}$: #connections admitted to bufferless resource $n(ct+b)$ from samples $\{A_{i,t}, i = 1, 2, \ldots\}$.
- can establish LDP for #connections and admitted aggregate

Impact of Sample Correlations

- Motivation: samples may be correlated
  - ◇ across connections, sample from same connection
- model X_i samples as Markov process
- Donsker-Varadhan for empirical measure of Markov process





- e.g. 2-state Markov: parameter $b < 1 \Leftrightarrow$ +ve correlations
  - ◇ (left) flatter sampling rate function J: $\widehat{N}$ more variable
  - ◇ (right) decreases $I(c)/\varepsilon$: increases attained loss

Bibliography

Currently under compilation; will be available at

`http://www.research.att.com/~duffield/pubs/lund`