Large deviations for performance analysis

Adam Shwartz

Electrical Engineering
Technion—Israel Institute of Technology
Israel

Based on:

A. Shwartz and A. Weiss:

Large deviations for performance analysis

queues, communication and computing.

Chapman-Hall, 1995.

DIAS, Dublin, December 1995.

$\boxed{\text{OUTLINE}}$

1. Theory.

- Large Deviations: an overview
  - Large Deviations for random variables
  - The rate function—Poisson random variables
  - The contraction principle
  - Probabilistic interpretation of optimal points
- Jump Markov Processes
  - Definitions
  - Most likely path
  - Large Deviations
  - Boundary problems
  - Wentzell-Friedlin results

Applications

- The M/M/1 queue
- The AMS model
  - Buffer overflow
  - Control schemes
- Priority queues

$\boxed{\text{LARGE DEVIATIONS}}$

- Compute probabilities of rare events by transforming into variational problems

For $X_1, X_2, \ldots$ i.i.d., define the "Rate Function"

$$\ell(\alpha) \triangleq \sup_{\theta} \left\{ \theta \cdot \alpha - \log \mathbb{E} e^{\theta \cdot X_1} \right\}. \tag{1}$$

Theorem 1 *Let $X_i \in \mathbb{R}$. Set $S_n \triangleq n^{-1} \sum_{1}^{n} X_k$. Then*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\mathrm{S_n} \in \mathrm{C}\right) \leq -\inf_{\mathrm{x} \in \mathrm{C}} \ell(\mathrm{x}) \tag{2}$$

for every closed set C, while for every open set G,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\mathrm{S_n} \in \mathrm{C}\right) \geq -\inf_{\mathrm{x} \in \mathrm{G}} \ell(\mathrm{x}). \tag{3}$$

ℓ is positive, convex, minimum $(= 0)$ at $\mathbb{E}\mathrm{X}_1$.

Remark 1 *no conditions!*
Upper bound holds for finite n, with constant 2.
In $\mathbb{R}^{\mathrm{d}}$, finite moment generating function suffices.
Upper bound is proved by Chebycheff, lower bound by a change of measure.

$\boxed{\text{Poisson Random Variables}}$

In $\mathbb{R}$ we can compute: for Poisson (λ) random variables, if $x \geq 0$,

$$\ell(x) = x\left(\log\left(\frac{x}{\lambda}\right) - 1\right) + \lambda \tag{4}$$

Let $\langle x, y\rangle$ denote the inner product in $\mathbb{R}^{\mathrm{d}}$.

Let e_j, $j = 1, \ldots, k$ be vectors in $\mathbb{R}^{\mathrm{d}}$.

Let y_{ij}, $j = 1, \ldots k$, $i = 1, 2, \ldots$ be independent, with $y_{ij} \stackrel{\mathcal{L}}{=}$ Poisson (λ_j).

Let $X_i \stackrel{\triangle}{=} \Sigma_{j=1}^{k} y_{ij} e_j$. The rate function for the sample mean of the X_i is

$$\ell(a) = \sup_{\theta}\left(\langle\theta, a\rangle - \sum_{j=1}^{k} \lambda_j \left(e^{\langle\theta, e_j\rangle} - 1\right)\right). \tag{5}$$

Remark 2 *ℓ behaves like $x \log x$ when $|a|$ is large, and vanishes if and only if $a = \Sigma_j \lambda_j e_j$—the mean.*

ℓ equals $+\infty$ if for every representation $a = \Sigma_1^k \beta_j e_j$ we have $\beta_j < 0$ for some j: probability 0 that a combination of jumps moves in direction a.

General Principles

The Large Deviations Principle

Definition 1 *A real valued function I on $\mathcal{X}$ is a rate function if I is non-negative and lower semicontinuous. If in addition the sets $\{x : I(x) \le \alpha\}$ are compact then I is called a good rate function.*

Definition 2 *A sequence z_n satisfies the large deviations principle with rate function I if*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\mathrm{z_n} \in \mathrm{C}\right) \le -\inf_{\mathrm{x}\in\mathrm{C}} \mathrm{I}(\mathrm{x}), \qquad (6)$$

for every closed set C, while for every open set G,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\mathrm{z_n} \in \mathrm{C}\right) \ge -\inf_{\mathrm{x}\in\mathrm{G}} \mathrm{I}(\mathrm{x}). \qquad (7)$$

Thus Theorem 1 claims that sample means of i.i.d. random variables satisfy the Large Deviations Principle.

The function ℓ of Equation (5) is a good rate function if and only if the positive cone spanned by the e_j equals $\mathbb{R}^{\mathrm{d}}$.

TRANSFORMATIONS

LET f BE CONTINUOUS, DEFINE $y_n \triangleq f(z_n)$ AND

$$I'(y) \triangleq \inf \{I(z) : z \in \mathcal{X},\ y = f(z)\}, \tag{8}$$

WITH $I'(y) = \infty$ IF NO SUCH z EXISTS.

Theorem 2 (Contraction) *Suppose the large deviations principle holds for z_n, with good rate function I. Then it holds for y_n, with good rate function I'.*

Theorem 3 (Varadhan's Lemma) *Suppose the large deviations principle holds for z_n, and f is bounded and continuous. Then*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left(\mathrm{e}^{\mathrm{nf(z_n)}}\right) = \sup_{\mathrm{x}} \left(\mathrm{f(x)} - \mathrm{I(x)}\right). \tag{9}$$

HOW RARE EVENTS OCCUR

S IS A CONTINUITY SET IF $I(x)$ ACHIEVES THE SAME INFIMUM ON THE INTERIOR AND THE CLOSURE OF S.

Theorem 4 *Let F be closed. If z_n satisfies the large deviations principle with good rate function I, then I achieves its minimum at say $x^* \in F$, and*

$$\mathbb{P}\left(\mathrm{d(z_n, x^*)} \geq \epsilon \mid \mathrm{z_n} \in \mathrm{F}\right) \leq \mathrm{e}^{-\mathrm{n}\delta}, \quad \text{some } \delta > 0. \tag{10}$$

Large Deviations for Processes

Intuition

If $\mathbb{E}x < \alpha$, HOW do we get $\sum_{i=1}^{n} X_i > n\alpha$?

Claim 1 $\mathbb{P}\left(\sum_{i=1}^{n} X_i > n\alpha\right) \approx \mathbb{P}\left(\sum_{i=1}^{j} X_i > j\alpha, \text{ all } j \leq n\right)$.

More precisely, on an EXPONENTIAL SCALE the probabilities are the same.

LD tells us HOW the rare event occurred.

"**Proof**" Clearly $\geq$ holds. On the other hand, if $\sum_{i=1}^{n} X_i > n\alpha$ then there is a cyclic permutation for which $\sum_{i=1}^{j} X_i > j\alpha$, all $j \leq n$. Since all n permutations are equally likely,

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i > \alpha\right) \leq n \cdot \mathbb{P}\left(\sum_{i=1}^{j} X_i > j\alpha, \text{ all } j \leq n\right). \quad (11)$$

Take $\frac{1}{n}\log$ on both sides and get $\approx$.

Jump-Markov processes

Construction

Local specification: given

- List of possible jump directions $\{e_j,\ j = 1, \ldots, k\}$
- Rates of jump $\lambda_j(x)$ in direction j when at x

the process $x(t)$ behaves as follows:

Leave the point x as Poisson, rate $\lambda(x) \triangleq \sum_j \lambda_j(x)$.

When x is left, the process moves to $x + e_m$ with probability $\lambda_m(x)/\lambda(x)$.

Encode the information in the generator L (acting on functions)

$$(Lf)(x) \triangleq \sum_{j=1}^{k} \lambda_j(x) \left(f(x + e_j) - f(x)\right). \tag{12}$$

The generator determines the distribution of the process (up to initial conditions):

$$f(x(t)) = f(x(0)) + \int_0^t (Lf)(x(s))\, ds + \text{martingale}. \tag{13}$$

$\boxed{\text{SCALING}}$

THE SCALED PROCESS MOVES n TIMES FASTER WITH JUMPS SMALLER BY n. ITS GENERATOR IS

$$(L_n f)(z) \stackrel{\triangle}{=} \sum_{j=1}^{k} n \cdot \lambda_j(z) \left(f\left(z + \frac{e_j}{n} \right) - f(z) \right). \tag{14}$$

$\boxed{\text{MOST LIKELY BEHAVIOR}}$

FOR n LARGE, OVER SMALL TIME Δ THE CHANGE IS ROUGHLY $n^{-1}\Delta \sum_{j=1}^{k} \sum_{i=1}^{n\Delta} y_{ij}$, $y_{ij} \stackrel{\mathcal{L}}{=}$ POISSON $\lambda_j(x)$ INDEPENDENT, AND y_{ij} MOVES IN DIRECTION e_j. BY LLN, FOR LARGE n THE PROCESS MOVES IN DIRECTION $\sum_j \lambda_j(z) e_j$. SO DEFINE z_∞ BY

$$\frac{d}{dt} z_\infty(t) = \sum_{j=1}^{k} \lambda_j(z_\infty(t)) e_j. \tag{15}$$

Theorem 5 (Kurtz) *If* $\log \lambda_i(x)$ *are bounded and Lipschitz continuous, then there are* $C_i > 0$ *s.t.*

$$\mathbb{P}\left(\sup_{0 \le t \le T} |z_n(t) - z_\infty(t)| \ge \epsilon \;\middle|\; z_n(0) = x \right) \le C_1 e^{-nC_2(\epsilon)}, \tag{16}$$

C_2 *is linear for* ϵ *large and quadratic for* ϵ *small.*

LD FOR PROCESSES: INTUITION

THE QUESTION ABOUT RARE EVENTS IS: FOR $r \neq z_\infty$,

$$\mathbb{P}\left(\mathrm{z_n(t)\ follows\ the\ path\ r(t),\ 0 \le t \le T}\right)\ ? \tag{17}$$

DIVIDE TIME TO Δ INTERVALS.
BY MARKOV PROPERLY, WE NEED, FOR EACH k,

$$\mathbb{P}\left(\mathrm{z_n(t) \approx r(t),\ k\Delta \le t \le (k+1)\Delta} \,\middle|\, \mathrm{z_n(k\Delta) \approx r(k\Delta)}\right). \tag{18}$$

SO, ON EACH INTERVAL, NEED $z_n(t)$ TO FOLLOW THE SLOPE $r'(k\Delta) \approx [r((k+1)\Delta) - r(k\Delta)]/\Delta$ OF r.
SINCE INCREMENTS OF THE PROCESS ARE POISSON, REPRESENT THEM AS A SUM OF $n \cdot \Delta$ INDEPENDENT RV'S, AND RESULT FOR RV'S IMPLIES

$$\text{PROBABILITY IN } (18) \approx e^{-(n\cdot\Delta)\cdot\ell(r(k\Delta),r'(k\Delta))}. \tag{19}$$

Remark 3 *r is fixed, so approximation of derivative is good if Δ small. Obviously, need $n \to \infty$ before $\Delta \to 0$.*

By Markov property, the probability in (17)

$$
\begin{aligned}
&\mathbb{P}\left(\mathrm{z_n(t)\ follows\ the\ path\ r(t),\ 0 \le t \le T}\right) \\
&\qquad \approx \prod_k e^{-n\cdot\Delta\ell(r(k\Delta),r'(k\Delta))} \\
&\qquad = \exp\left(-n\sum_k[\ell(r(k\Delta),r'(k\Delta))\cdot\Delta]\right) \\
&\qquad \approx \exp\left(-n\int_0^T \ell(r(t),r'(t))\,dt\right).
\end{aligned}
$$

So, Large Deviations for processes:

$$\mathbb{P}\left(\mathrm{z_n(t)\ near\ r(t)}\right) \approx \exp\left(-nI_0^T(r)\right) \tag{20}$$

$$I_0^T(r) \triangleq \int_0^T \ell(r(t),r'(t))\,dt. \tag{21}$$

An "event" is a set of possible paths: so

$$\mathbb{P}_\mathrm{n}\left(\mathrm{event}\right) \approx \exp\left[-\mathrm{n}\inf_{\mathrm{r}\in\mathrm{event}} \mathrm{I}_0^\mathrm{T}(\mathrm{r})\right] \tag{22}$$

Moreover, if event occurs, it does near r^*— the minimizer in (22).

$\boxed{\text{THEOREMS FOR PROCESSES}}$

RECALL: LOCAL RATE AND RATE FUNCTION

$$\ell(x,y) \triangleq \sup_\theta \left(\langle \theta, y \rangle - \sum_{j=1}^k \lambda_j(x) \left(e^{\langle \theta, e_j \rangle} - 1 \right) \right). \quad (23)$$

$$I_0^T(r) \triangleq \int_0^T \ell(r(t), r'(t))\, dt. \quad (24)$$

Theorem 6 *Assume that for each i, $\log \lambda_i(x)$ is a bounded and Lipschitz continuous function. Then $I_0^T(\cdot)$ is a good rate function. Moreover, for every closed set F*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}_x (z_n \in F) \le - \inf \left\{ I_0^T(r) : r \in F,\ r(0) = x \right\}.$$

For every open set G, uniformly for x in compact sets,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}_x (z_n \in G) \ge - \inf \left\{ I_0^T(r) : r \in G,\ r(0) = x \right\}.$$

Remark 4 *The space of paths includes all right continuous functions on $[0,T]$ with $r(0) = x$.*
Theorem 6 holds in the Skorohod (J) topology as well as in the sup-norm topology.
There is generally no uniformity in the upper bound.
The theorem holds for each T, but not for $T = \infty$!

Consequences of integral form

Computation of probabilities leads to variational problem: for a set of paths S,

$$\inf\left\{I_0^T(r) : r \in S\right\} \tag{25}$$

- Integral form of I is exactly the form treated by Calculus of Variations.

  We can restrict the probabilistic question to Absolutely Continuous Paths (can prove all limits are AC), but NOT to smoother paths (e.g. piecewise C_1).

  Thus must solve Calculus of Variations problem over AC functions.
- Analytic solutions—sometimes, mostly in $\mathbb{R}$.
- If $\ell(x, y) = \ell(y)$ is independent of x then optimal path with $r(0) = r_0$ and $r(T) = r_1$ is a straight line (Jensen).
- Principle of optimality: if r^* optimal, then for $s < T$, $\{r^*(t),\ 0 \leq t \leq s\}$ is optimal for the problem ending at $r^*(s)$, and similarly for the end part. This leads to pasting solutions.

Easy example: the Free M/M/1

$$e_1 = +1 \quad \lambda_1(x) = \lambda \tag{26}$$

$$e_2 = -1 \quad \lambda_2(x) = \mu \tag{27}$$

By Theorem 5,

$$z_\infty(t) = z_\infty(0) + (\lambda - \mu)t. \tag{28}$$

The local rate is

$$\ell(x, y) = \ell(y) = y \log\left(\frac{y + \Delta}{2\lambda}\right) + \lambda + \mu - \Delta \tag{29}$$

where $\Delta \triangleq \sqrt{y^2 + 4\lambda\mu}$. To calculate

$$\mathbb{P}\left(\mathrm{z_n}(1) \geq \mathrm{a} \,\middle|\, \mathrm{z_n}(0) = 0\right)$$

we need to solve

$$\inf\left\{I_0^1(r) : r(0) = 0,\ r(1) \geq a\right\}.$$

Since ℓ is independent of position, the optimizer is a straight line. So, for $a > \lambda - \mu$, $r^*(1) = a$ and $I^* = \ell(a)$.

This is not typical ... (way too easy).

All queueing systems have "boundaries:" surfaces where rates are discontinuous. The latest theory is in Dupuis, P., R.S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*, Wiley 1995.

An easy solution to some cases: contraction.

$\boxed{\text{The M/M/1 process}}$

$$e_1 = +1 \quad \lambda_1(x) = \lambda \tag{30}$$

$$e_2 = -1 \quad \lambda_2(x) = \mu 1[x > 0]. \tag{31}$$

Assume stability: $\rho \triangleq \frac{\lambda}{\mu} < 1$. Kurtz's Theorem 5 holds, so the likely path

$$z_\infty(t) = \max\{z_\infty(0) - (\lambda - \mu) \cdot t, 0\}. \tag{32}$$

Lemma 1 *The map R below is continuous:*

$$(Rr)(t) \triangleq r(t) - \inf_{0 \le s \le t} r(s). \tag{33}$$

This maps real functions with $r(0) = 0$ to positive functions with $r(0) = 0$.

R maps the free M/M/1 process to the M/M/1!

Theorem 7 *The M/M/1 process with $\rho < 1$ satisfies the large deviations principle with good rate function*

$$I_0^T(r) = \int_0^T \ell(r(t), r'(t))\, dt \tag{34}$$

for positive paths, and ℓ is the same as for the free M/M/1 process except that $\ell(0,0) = 0$.

Remark 5 *ℓ is discontinuous in x at 0, and at $x = 0$ it is defined only for $y \geq 0$.*

Remark 6 *If, for some queueing system, there exists a continuous "reflection map," then we can use contraction. See Dupuis and Ishii, Tsoucas, etc.*

SOMETIMES WE CAN AVOID THE BOUNDARY:

Theorem 8 *For a stable M/M/1 process,*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}_0\,(\text{busy period} \geq \text{n}) = -\left(\sqrt{\mu} - \sqrt{\lambda}\right)^2$$

$$\lim_{n\to\infty} \mathbb{P}_0\left(\sup_{0\leq \text{t} \leq \text{n}} \frac{\text{x(t)}}{\text{n}} \geq \epsilon \;\middle|\; \text{busy period is} \geq \text{n}\right) = 0.$$

HERE THE PROCESS AND THE FREE PROCESS AGREE!

FLAT BOUNDARY THEORY

HERE THE SET $\{x : x_0 = 0\}$ IS THE BOUNDARY.

THE JUMP DIRECTIONS ARE $e_i(1)$ IN THE INTERIOR $(x_0 > 0)$ AND $e_i(0)$ ON THE BOUNDARY.

JUMP RATES ARE $\lambda_j(i; x),\ i = 0, 1.$

THE GENERATOR CHANGES WHEN $x_0 = 0$.

SCALING IS AS BEFORE: RATES INCREASE BY n, JUMPS DECREASE BY n.

e_j^0 IS THE COMPONENT OF e_j IN DIRECTION x_0.

THE NORMAL DRIFT AT x IS

$$\beta(x) = \sum_{j=1}^{k(1)} \lambda_j(1; x) e_j^0(1) \tag{35}$$

IF $\beta(x) > 0$ AT $x = (0, x_1, \ldots, x_d)$ THEN THE PROCESS SPENDS NO TIME ON THE BOUNDARY.
OTHERWISE, THE PROPORTION OF THE TIME IS $\pi(x)$

$$\pi(x) \sum_{j=1}^{k(0)} \lambda_j(0; x) e_j^0(0) + (1 - \pi(x))\beta(x) = 0. \tag{36}$$

$\pi(x)$ IS THE INVARIANT DISTRIBUTION AT 0 OF THE PROCESS RESTRICTED TO THE NORMAL DIRECTION.

MOST LIKELY PATH IS GIVEN BY USUAL ODE IN THE INTERIOR OR WHEN $\beta(x)$ IS POSITIVE; BY A π-WEIGHTED SUM ON THE BOUNDARY WHEN $\beta(x) < 0$.

$\ell(x, y)$ CHANGES ONLY ON $x_0 = 0$: IT IS ∞ IF $y < 0$, UNCHANGED IF $y > 0$. IF $x_0 = y = 0$ ℓ IS

$$\sup_\theta \left(\langle \theta, y \rangle - \max_m \sum_{j=1}^{k(m)} \lambda_j(m; x) \left(e^{\langle \theta, e_j(m) \rangle} - 1 \right) \right). \quad (37)$$

THE FORM OF THE RATE FUNCTION I IS UNCHANGED.

- ℓ MAY JUMP DOWN AT $y_0 = 0 \Rightarrow$ NOT CONTINUOUS, NOT CONVEX!
- ℓ STRICTLY CONVEX WHEN $x_0 > 0$. CONVEX WHEN $x_0 = 0$ IN $y_1, \ldots, y_d$, BUT NOT STRICTLY CONVEX: STRAIGHT LINES ARE NOT UNIQUE MINIMIZERS!

THE LDP HOLDS WITH THE SAME FORMULATION.

PROOF OF UPPER BOUND IS BASICALLY THE SAME.

FOR LOWER BOUND NEED TO

- APPROXIMATE BY CONSTANT COEFFICIENT PROCESS
- ESTIMATE HOW MUCH TIME IS SPENT ON BOUNDARY
- SHOW EXPONENTIAL CONVERGENCE OF TIME ON BOUNDARY TO STEADY STATE, UNIFORMLY IN A CLASS OF TWISTED MEASURES!

Exit problems

Clearly, the probability that the stable M/M/1 process exceeds level a at some time is 1:

$$\mathbb{P}\left(\sup_{\mathrm{t}\geq 0} \mathrm{x(t)} \geq \mathrm{a}\right) = 1 \tag{38}$$

for all a. But given this event, what is the most likely path?

Consider a "stable" process with a "stable point" q. Let G be a set containing q in its interior. The most likely path of the scaled process will lead to q. Define

$$I_G^* \triangleq \inf\left\{I_0^T(r) : r(0) = q,\ r(T) \in \partial G,\ T > 0\right\}. \tag{39}$$

Let $\tau_n \triangleq \inf\{t : z_n(t) \in \partial G\}$.

By Freidlin-Wentzell, if $I^0_{-T^*}(r^*) = I_G^*$ then

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{\tau_\mathrm{n}-\mathrm{T}\leq \mathrm{t}\leq \tau_\mathrm{n}} |\mathrm{z_n(t)} - \mathrm{r^*(t} - \tau_\mathrm{n}\mathrm{(G))}|\right) = 1 \tag{40}$$

for all $T < T^*$. Moreover,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\log \tau_\mathrm{n}}{\mathrm{n}} \approx \mathrm{I_G^*}\right) = 1. \tag{41}$$

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{\tau_n - T\le t\le \tau_n} \left|z_n(t) - r^*(t-\tau_n(G))\right|\right) = 1 \qquad (42)$$

FOR ALL $T < T^*$. MOREOVER,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\log \tau_n}{n} \approx I^*_G\right) = 1 : \qquad (43)$$

THE LAST PART OF THE PATH LEADING TO EXIT IS ALONG THE SOLUTION TO A "FREE TIME" VARIATIONAL PROBLEM. THE EXIT TIME IS ROUGHLY $e^{nI^*_G}$.

THEORY ALLOWS TO STUDY RELATIVE PROBABILITIES OF EXIT LOCATIONS FROM G. ITS EXTENSIONS DEAL WITH RATES OF TRANSITION BETWEEN LOCALLY STABLE POINTS.

LET π_n BE THE INVARIANT PROBABILITY, AND $V(x)$ THE OPTIMAL (FREE TIME) COST TO REACH x FROM q. THEN FOR AN OPEN SET D NOT CONTAINING q,

$$\lim_{n\to\infty} \frac{1}{n} \log \pi_n(D) = -\inf_{x\in D} V(x) : \qquad (44)$$

THE STEADY STATE PROBABILITY OF D IS PROPORTIONAL TO THE PROBABILITY OF GOING FROM q TO SOME POINT IN D, SINCE MOST OF THE TIME THE PROCESS IS NEAR q.

Warning: uses and misuses

Given an application, suggested use:

- Formulate model
- Compute z_∞
- Formulate questions about rare events
- Solve variational problems
  - Analytically, if you can
  - Numerically, if you must
- Justify the mathematics
  - LDP or at least the variational formula for the set of interest
  - Calculus of Variations
- Write a paper ...

Exit: beyond M/M/1

one dimensional birth-death process

For a one-dimensional birth-death process (with position-dependent rates!), exit problem can be solved explicitly.
$e_1 = +1$, $e_2 = -1$, and $\lambda(x)$ and $\mu(x)$ are general smooth functions on $[0, a+1]$, with $\frac{\lambda(x)}{\mu(x)} < 1$.
The variational problem associated with exit

$$V(a) \stackrel{\triangle}{=} \inf \left\{ I_0^T(r) : r(0) = 0, \;\; r(T) = a, \;\; T > 0 \right\}. \quad (45)$$

$$\ell(x, y) \;=\; \sup_\theta \left(\theta y - g(x, \theta) \right) \quad (46)$$

$$g(x, \theta) \;=\; \lambda(x) \left(e^\theta - 1 \right) - \mu(x) \left(e^{-\theta} - 1 \right). \quad (47)$$

Solution is necessarily monotone increasing.
By generic calculations, optimizing θ^* (resp.) any optimizing path satisfy (Euler)

$$\theta^*(x) \;=\; \log \left(\frac{\mu(x)}{\lambda(x)} \right) \quad (48)$$

$$0 \;=\; r'(t) - \lambda(r(t)) \frac{\mu(r(t))}{\lambda(r(t))} + \mu(r(t)) \frac{\lambda(r(t))}{\mu(r(t))} \quad (49)$$

$$\;=\; r'(t) - \mu(r(t)) + \lambda(r(t)). \quad (50)$$

Exit of birth-death continued

Conclusions: optimal path given by

$$r'(t) = \mu(r(t)) + \lambda(r(t)), \;\; r(0) = 0 \tag{51}$$

Optimal cost of exit path

$$V(a) = \int_0^a \log\left(\frac{\mu(x)}{\lambda(x)}\right)\, dx. \tag{52}$$

In particular for the M/M/1 process,

$$r(t) \;=\; (\mu - \lambda)\cdot t \tag{53}$$

$$V(a) \;=\; a \cdot \log\frac{\mu}{\lambda} \tag{54}$$

$$T^* \;=\; \frac{a}{\mu - \lambda}. \tag{55}$$

Exit is caused by reversing the rates!
In particular, in steady state,

$$\mathbb{P}(\mathrm{z_n} \geq \mathrm{a}) \;\approx\; \exp\left(-na\log\frac{\lambda}{\mu}\right) \tag{56}$$

$$\mathbb{P}(\mathrm{x(t)} \geq \mathrm{na}) \;\approx\; \exp\left(-na\log\frac{\lambda}{\mu}\right) = \left(\frac{\lambda}{\mu}\right)^{na} \tag{57}$$

which is exact!

More about the M/M/1

Probability of hitting a point

In discrete models in $\mathbb{R}^{\mathrm{d}}$, the number of points in a neighborhood grows as n^d (as $n \to \infty$). Probabilities of rare events decay exponentially. Therefore,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(|\mathrm{z_n(T)} - \mathrm{y}| \le \frac{1}{2\mathrm{n}} \;\middle|\; \mathrm{z_n(0) = x} \right) \tag{58}$$

$$= \inf \left\{ I_0^T(r) : r(0) = x,\ r(T) = y \right\}. \tag{59}$$

Transient behavior

How does this probability behave as a function of T? What is the optimal path?

For short T the optimal path is a straight line.

For large T it is optimal to empty first, then fill up to y.

Hence probability is convex up to some T^*, then constant.

There is no optimal path which spends zero time at $z = 0$.

THE BASIC ANICK-MITRA-SONDHI MODEL

n MARKOV FLUID SOURCES, RATE 1 EACH

LINK CAPACITY EXCEEDS MEAN DEMAND: $\frac{\lambda}{\lambda+\mu} < C$.

$$z_n(t) \triangleq \frac{1}{n} \cdot \text{NUMBER OF ``ON'' SOURCES} \tag{60}$$

$$b_n(t) \triangleq \frac{1}{n} \cdot \text{BUFFER CONTENTS} \tag{61}$$

z_n IS MARKOV. DYNAMICS OF BUFFER:

$$\frac{d}{dt} b_n(t) = \begin{cases} z_n(t) - C & \text{IF } z_n(t) > C \text{ OR } b_n(t) > 0 \\ 0 & \text{IF } z_n(t) \leq C \text{ AND } b_n(t) = 0 \end{cases} \tag{62}$$

$\boxed{\text{How to fill the buffer to level } B?}$

$\boxed{\text{Principle of optimality:}}$

First make $z_n = C$ optimally.

Then (area under $z_n(t) - C$) $= B$ optimally.

First variational problem (q steady state):

$$\inf \left\{ I_0^T(r) : r(0) = q,\ r(T) = C,\ T > 0 \right\} \tag{63}$$

Free time! Second variational problem:

$$\inf \left\{ I_0^T(r) : r(0) = C,\ \int_0^T (r(t) - C)\, dt = B \right\} \tag{64}$$

Optimizer is a hyperbolic cosine: concave.

From variational problem to probabilities

What do we get from the solution
of the variational problem?

1. $\mathbb{P}\,(\text{buffer starts filling}) \approx \mathrm{e}^{-\mathrm{n}\mathrm{I}_\mathrm{c}^*}$,

where I_c^* solves level crossing problem.

2. If r_c^* is optimal, that is $I_c^* = I_c(r_c^*)$, then

$$\mathbb{P}\,(z_n \text{ near } r_c^* \mid \text{buffer starts filling}) \approx 1. \qquad (65)$$

3. $\mathbb{P}\,(\text{buffer overflows}) \approx \mathrm{e}^{-\mathrm{n}\mathrm{I}_\mathrm{b}^*}$,

where I_b^* solves overflow problem.

4. If r_b^* is optimal, that is $I_b^* = I_b(r_b^*)$, then

$$\mathbb{P}\,(z_n \text{ near } r_b^* \mid \text{buffer overflows}) \approx 1. \qquad (66)$$

That is, we obtain the probability and the path of the unlikely event.

$\boxed{\text{Asymptotics: small buffer}}$

Approximate scaled rates.

$z =$ fraction of "on" sources.

$$\lambda(1-z) \to \lambda(1-C), \quad \mu z \to \mu C. \tag{67}$$

"Cost" $I^*(B)$ of filling B is

$$I^*(B) = c_2\sqrt{B} + O(B). \tag{68}$$

Asymptotics: large buffer

Most area obtained with $r(t) = \max_s r(s) = D$

where cost per buffer growth is minimal.

$$I^*(B) = c_3 B + c_4 + o(1). \tag{69}$$

$\boxed{\text{Control schemes}}$

Drop bits (lease significant) or packets (tagged)

When to start? when to stop?

$\boxed{\text{Rate based control:}}$

Reducing input when $z > D$ has no effect!

SO—DROP BITS according to BUFFER SIZE

Solution technique:

Split problem to pieces with smooth rates

Use principle of optimality on each piece

To match the pieces use

$\boxed{\text{Principle of smooth fit}}$

(Erdmann corner condition)

$\boxed{\text{Bit dropping schemes}}$

Trim above watermark ($\rho < 1,\ B > B_1$)

$$\frac{d}{dt}b_n(t) = \begin{cases} z_n(t) - C & \text{if } z_n(t) > C \text{ and } 0 < b_n(t) < B_1 \\ \rho z_n(t) - C & \text{if } b_n(t) > B_1 \\ 0 & \text{if } z_n(t) \le C \text{ and } b_n(t) = 0. \end{cases} \tag{70}$$

Principle of optimality: find cheapest way

$z_n(t) = C$, then

$b_n(t)$ exceeds B_1, and find resulting z_n

now make buffer B with input $\rho z_n(t)$.

Calculations—same as for the basic model

except possibly for values of some parameters!

More bit dropping schemes

Trim all packets above watermark

Once watermark exceeded, trim incoming as well as buffered bits until buffer empties!

Dynamics: IN EACH BUSY PERIOD

If watermark exceeded, trim packets in buffer by ρ, and

$$\frac{d}{dt} b_n(t) =$$

$$\begin{cases} z_n(t) - C & \text{if } b_n(t) > 0 \text{ and } b_n(s) < B_1, \\ & \text{all } s < t \text{ in current busy period} \\ \rho z_n(t) - C & \text{if } b_n(t) > 0 \text{ and } b_n(s) > B_1, \\ & \text{some } s < t \text{ in current busy period} \\ 0 & \text{if } z_n(t) \le C \text{ and } b_n(t) = 0 \end{cases}$$

Need additional coordinate to make this Markov.

Analysis in segments.

$\boxed{\text{More bit dropping schemes}}$

$\boxed{\text{Two watermark}}$

Start trimming if higher mark exceeded

Stop if went under lower watermark.

$$\frac{d}{dt} b_n(t) =$$

$$\begin{cases} z_n(t) - C & \text{if } b_n(t) > 0 \text{ and } b_n(s) < B_2, \\ & \text{all } s < t \text{ in current busy period} \\ z_n(t) - C & \text{if } 0 < b_n(t) < B_1 \\ \rho z_n(t) - C & \text{if } b_n(t) > B_1 \text{ and } b_n(s) > B_2, \\ & \text{some } s < t \text{ in current busy period} \\ 0 & \text{if } z_n(t) \leq C \text{ and } b_n(t) = 0 \end{cases} \tag{71}$$

Need additional coordinate

to make this Markov: trim status.

Analysis in segments.

$\boxed{\text{Sample calculation}}$

Solving the variational problem

Level crossing: First variational problem (q steady state):

$$\inf\left\{I_0^T(r) : r(0) = q,\ r(T) = C,\ T > 0\right\}.$$

Solve the necessary conditions:

$$\ell(r, r') - r'\ell(r, r') = 0.$$

Free time! In this case we obtain

$$r'(t) = \mu(r(t)) - \lambda(r(t)),$$

a linear equation! Moreover,

$$I_c^* = \int_q^C \log\left(\frac{\mu(r)}{\lambda(r)}\right) dr.$$

In this case have explicit solutions.

Second variational problem:

$$\inf \left\{ I_0^T(r) : r(0) = C,\ \int_0^T (r(t) - C)\, dt = B \right\}$$

Introduce the buffer size as a constraint.

Use Lagrange Multipliers.

Solution: usually numerical techniques

But can obtain explicit asymptotics.

Priority queues

Models:

- Preemptive priority queue (Two customer types)
- Join the shorter queue
- Serve the longer queue

The problems: non-flat boundaries.

The questions: exit paths.

The results: can compute and justify, but

- Not in all cases
- Using computer algebra software