Large Deviations Theory: Basic Principles and Applications to Communication Networks

Michele Pagano

Dipartimento di Ingegneria dell'Informazione
Università di Pisa, Via Caruso, I-56122 Pisa, Italy
michele.pagano@iet.unipi.it

Abstract. The *theory* of large deviations refers to a collection of techniques for estimating properties of rare events such as their frequency and most likely manner of occurrence. Loosely speaking, LDT can be seen as a refinement of the classical limit theorems of probability theory and it is useful when simulation or numerical techniques become increasingly difficult as a parameter of interest tends to its limit.

The first part of this tutorial deals with the behaviour of the empirical mean of IID RVs, the most natural framework to introduce the basic concepts and theorems of LDT and to highlight their heuristic interpretation.

Then, the large deviation principle for the single server queue is presented and its implications on network dimensioning are discussed. Finally, the tutorial overviews the application of LDT to rare event simulation, for the choice of the optimal change of measure in Importance Sampling.

Keywords: LDT, Rare Events, Contraction Principle, Queues, LRD.

1 Introduction

In the framework of teletraffic engineering, many challenging issues have arisen in the last two decades as a consequence of the fast growth of network service demand. The search for *global* network architectures, which should handle heterogeneous applications and different quality of service (QoS) guarantees [1], has determined a widespread interest for novel performance evaluation techniques, able to cope with the increasing size (and complexity) of telecommunication systems. The need for new mathematical approaches is also related to the adoption of more sophisticated traffic models, the so-called Long Range Dependent (LRD) processes, able to take into account the long memory features of real traffic [2,3].

In case of stringent QoS requirements, network performance are determined by events with a small probability of occurring, but with severe consequence when they occur. Since these events are linked to *large deviations* from the normal behaviour of the system, the so-called theory of large deviations (LDT) represents a natural candidate for analysing *rare events in large systems*.

In a nutshell, LDT studies the tails of distributions of certain random variables. Since, by definitions, probabilities of rare events are involved, it is also known as the theory of rare events. As a matter of fact, LDT only applies to certain types of rare events,

D. Kouvatsos (Ed.): Next Generation Internet, LNCS 5233, pp. 301–330, 2011.

caused by a large number of unlikely things occurring together (*conspiracy*), rather then a single event of small probability. For instance, winning a lottery is not a large deviations event, since it is determined by a single trial that cannot be broken into more than one sub-event [4].

Unlike classical limit theorems, LDT also provides a nice qualitative theory to *understand* rare events and the typical way they occur (*most likely path*). Indeed, the probability of a rare event is often reduced to a deterministic optimisation problem. If a cost is assigned to each sample path that would cause the rare event, its probability only depends on the *cheapest path*, i.e., on the cheapest way the event can happen. This concept is described in [4] as the *strong law of rare events*: if there is a unique cheapest path, then as the *asymptotic parameter* gets large, conditioned on the occurrence of the rare event, with overwhelming probability the system followed the cheapest path for any bounded interval of time before the rare event occurred.

This deeper insight into the system behaviour can be successfully employed, for instance, to design proper control systems and to speed-up the simulation of rare events. Indeed, a control affects the probability of the rare event iff it affects the cheapest path; as a consequence, the time scale on which the control should operate is implicitly determined by the *most likely time* of occurrence of the rare event. In the framework of simulation, unlike crude Monte Carlo, the application of speed-up techniques generally requires some additional information about the behaviour of the system, such as the one provided (although in an asymptotic and eventually approximate form) by the LDT.

As pointed out by many authors [5], there is no *real* theory of large deviations and often the same result may be reached in different (and apparently unrelated) ways. Hence, as a whole LDT refers to a set of basic definitions, that by now are standard, and a variety of tools for the analysis of small probability events in completely different frameworks (such as statistical mechanics, information theory, parameter estimation and traffic engineering, just to name a few application fields).

The aim of this tutorial, which is heavily based on [6], is to introduce the basic LDT concepts, highlighting their heuristic interpretation from an engineering perspective and focusing on their applications (or, at least, on some of them) in the framework of queueing systems and computer networks. In more detail, the rest of the paper is organised as follows. Section 2 describes the key LDT principles, starting from simple practical examples and generalising the results to more abstract frameworks. Then Section 3 deals with the application of LDT to the single server queue, focusing on two well-known asymptotic regimes: large-buffer and many-sources asymptotics, while a few more advanced topics (queueing performance in presence of LRD traffic and LDT-based changes of measures) are sketched in Section 4. Finally, hints on further readings conclude the tutorial.

2 Basic LDT Results

The theory of large deviations is concerned with the asymptotic estimation of probabilities of rare events. In its basic form, the theory considers the limit of normalisations of $\log \mathbb{P}(A_n)$ for a sequence of events with asymptotically vanishing probability. Although the topic may be traced back to the early 1900s (see [5] for more detailed historical notes, interpretations and references), its general abstract characterisation by means of a *large deviation principle* was formalised only in 1966 by Varadhan [7], who is considered one of the founders of the *modern* theory of large deviations, together with Donsker (in the West) as well as Freidlin and Wentzell (in the East).

The following subsections will review the basic concepts that by now are standard, starting from the case of independent, identically distributed (IID) random variables (RVs) and introducing some more advanced tools (such as the large deviation principle and the contraction principle), which will be applied in the following to queueing systems.

2.1 Large Deviations of IID RVs

Let us consider the most classical topic of probability theory, namely the behaviour of the empirical mean of IID RVs. Before stating the general result (Cramér's theorem), let us consider some simple examples (see Chapter 2 in [6] for further details).

Sums of Standard RVs. Let $X_i \in \mathcal{N}(0,1)$[1] and consider the empirical mean

$$M_n = \frac{1}{n} S_n \qquad \text{where} \quad S_n = \sum_{i=1}^{n} X_i \ . \tag{1}$$

Since $M_n \in \mathcal{N}(0, 1/n)$, it is easy to show that:

1. for any $a > 0$

$$\lim_{n\to\infty} \mathbb{P}(M_n \geq a) = 0 \qquad \text{(Weak Law of Large Numbers)} \tag{2}$$

2. for any interval A

$$\lim_{n\to\infty} \mathbb{P}\left(\sqrt{n}M_n \in A\right) = \frac{1}{\sqrt{2\pi}} \int_A e^{-\frac{1}{2}x^2}\, dx \qquad \text{(Central Limit Theorem)} \tag{3}$$

3. for any $a > 0$

$$\mathbb{P}(M_n \geq a) = \frac{1}{\sqrt{2\pi}} \int_{a\sqrt{n}}^{\infty} e^{-\frac{1}{2}x^2}\, dx \tag{4}$$

and therefore

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(M_n \geq a) = -\frac{a^2}{2} \ , \tag{5}$$

which is a typical large deviations result.

Roughly speaking, according to (3), the *typical* value of M_n is of the order of $1/\sqrt{n}$, but with small probability (of the order of $e^{-n\,a^2/2}$, as suggested by (5)), M_n takes relatively large values.

It is well known from elementary probability theory that (2) and (3) remain valid as long as $\{X_i\}$ are IID RVs of zero mean and unit variance and can be easily modified

[1] As usual, $\mathcal{N}\left(\mu, \sigma^2\right)$ will denote a Gaussian RV with mean μ and variance σ^2.

in case of IID RVs with mean μ and variance σ^2. Instead, as far as (5) is concerned, the limit still exists (under quite general assumptions), but its value depends on the specific distribution of X_i. This is precisely the content of Cramér's theorem. In order to understand the kind of approximations involved in LDT, it is useful to derive a result similar to (5) in a slightly less trivial framework.

Sums of Bernoulli RVs. Let $X_i \in \mathcal{B}(p)$, i.e., $\mathbb{P}(X_i = 1) = p = 1 - \mathbb{P}(X_i = 0)$; in this case M_n can be seen as the proportion of heads in n independent tosses of a biased coin, which has probability p of coming up heads. Suppose that n is large and consider the probability that M_n exceeds a, for some $a > p$. Through direct calculation[2] (for notational convenience, suppose that $na < n$ is an integer):

$$\begin{aligned}
\mathbb{P}\left(M_n \geq a\right) &= \sum_{j=na}^{n} \binom{n}{j} p^j (1-p)^{n-j} && (S_n \text{ has a Binomial distribution}) \\
&\approx \binom{n}{na} p^{na} (1-p)^{n(1-a)} && \text{(Principle of the largest term)} \\
&= \frac{n!}{(na)!\,(n-na)!}\, p^{na}(1-p)^{n(1-a)} \\
&\approx \frac{1}{\sqrt{2\pi n(1-a)a}}\; a^{-na}\,(1-a)^{-n(1-a)}\, p^{na}(1-p)^{n(1-a)} \\
&\approx \left(\frac{a}{p}\right)^{-na} \left(\frac{1-a}{1-p}\right)^{-n(1-a)} \\
&= e^{-n\left(a\log\frac{a}{p} + (1-a)\log\frac{1-a}{1-p}\right)} ,
\end{aligned}$$

where the Stirling's formula was used to approximate the binomial coefficient:

$$n! \approx \sqrt{2\pi n}\, n^n e^{-n} .$$

Hence, an expression similar to (5) can be written also in case of Bernoulli RVs:

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(M_n \geq a\right) = a\log\frac{a}{p} + (1-a)\log\frac{1-a}{1-p} \triangleq H(a;p) \tag{6}$$

and $H(a;p)$ is known as the relative entropy, or Kullback-Leibler divergence, of the probability distribution $(a, 1-a)$ with respect to the probability distribution $(p, 1-p)$.

It is worth noticing that a *single term* in the sum is sufficient to determine its correct exponential decay rate (in n). It turns out that this feature is characteristic of many situations where LDT is applicable and is known as *principle of the largest term*, which is often expressed in the probability context by the phrase *"rare events occur in the most likely way"*.

[2] It is straightforward to verify that the largest term in the sum corresponds to $j = na$.

LDT Rate Function. The limit (5) depends on the specific distribution of X_i through the so-called *rate function* Λ^*, which is defined as the Fenchel-Legendre transform of the Cumulant Generating Function. Before stating the general LDT result for sums of IID RVs, it is worth introducing the definition of rate function and its main properties.

Let $\Lambda(\theta)$ denote the Logarithmic Moment Generating Function or Cumulant Generating Function[3] of a real-valued RV X, i.e.,

$$\Lambda(\theta) \overset{\Delta}{=} \log M(\theta) = \log \mathbb{E}\left(e^{\theta X}\right) \tag{7}$$

where

$$M(\theta) \overset{\Delta}{=} \mathbb{E}\left(e^{\theta X}\right) \tag{8}$$

is the Moment Generating Function of X.

Let $\Lambda^*(x)$ be the *convex dual* or *Fenchel-Legendre transform* of $\Lambda(\theta)$:

$$\Lambda^*(x) \overset{\Delta}{=} -\log\left(\inf_\theta e^{-\theta x} M(\theta)\right) = \sup_\theta \left(\theta x - \log M(\theta)\right) = \sup_\theta \left(\theta x - \Lambda(\theta)\right) \tag{9}$$

Figure 1 gives a graphical interpretation of the previous definition: $\Lambda^*(x)$ is the smallest amount by which the straight line $x\,\theta$ (with slope x) has to be pushed down so as to lie below the graph of $\Lambda(\theta)\ \ \forall\theta \in \mathbb{R}$.

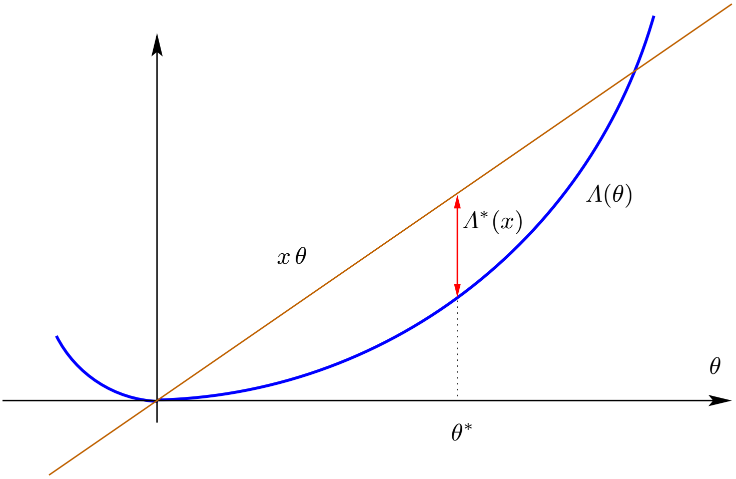

Fig. 1. Graphical interpretation [8] of the Fenchel-Legendre transform

The most relevant properties of Λ^* are recalled (the corresponding proofs are given, for example, in [5]) in the following:

1. $\Lambda^*(x)$ is convex, i.e., $\forall\,\lambda \in \left[0,1\right]$:

$$\Lambda^*\left(\lambda x_1 + (1-\lambda)x_2\right) \le \lambda\Lambda^*(x_1) + (1-\lambda)\Lambda^*(x_2)$$

[3] Indeed, the cumulants of X are just the derivatives of $\Lambda(\theta)$ evaluated at $\theta = 0$.

2. $\Lambda^*(x)$ is non-negative
3. $\Lambda^*(x)$ has its minimum for $x = \mu \triangleq \mathbb{E}(X)$ and $\Lambda^*(\mu) = 0$
4. $\Lambda^*(x)$ is lower semicontinuous, i.e., the level sets $\{x : \Lambda^*(x) \leq \alpha\}$ are all closed for $\alpha \in \mathbb{R}$
5. If the supremum in (9) is attained at a point θ^* in the interior of the interval where $M(\theta)$ is finite, then $M(\theta)$ is differentiable at θ^*, so that
$$\Lambda^*(x) \;=\; -\log \mathbb{E}\left(e^{\theta^*(X-x)}\right) \;=\; \theta^* x - \Lambda(\theta^*)$$
6. Let c be the greatest lower bound for a RV X, i.e.,
$$\mathbb{P}(X < c) \;=\; 0 \qquad \text{and} \qquad \mathbb{P}(X \leq c + \epsilon) \;>\; 0 \quad \forall \epsilon > 0 \ .$$
Then
(a) $\Lambda^*(x) \;=\; \infty \quad$ for $x < c$
(b) $\Lambda^*(c) \;<\; \infty \quad \Longleftrightarrow \quad \mathbb{P}(X = c) \;>\; 0$

Table 1 gives the expressions of Λ and Λ^* for some common distributions, highlighting the similarities with the preliminary examples reported in this section. For instance, in the case of Bernoulli RVs, $\Lambda^*(x) = H(x, p)$ and (6) justifies the name of *rate function* given to Λ^* in the LDT framework: indeed it is the function that specifies the rate of convergence for the Weak Law of Large Numbers.

Cramér's Theorem. Cramér's theorem (1938) is the most general result for IID RVs, stated in generic large deviations form.

Table 1. Examples of rate functions

X	$\Lambda(\theta) = \log \mathbb{E}\left(e^{\theta X}\right)$	$\Lambda^*(x) = \sup\limits_{\theta \in \mathbb{R}} \left(\theta x - \Lambda(\theta)\right)$
$\mathcal{N}(\mu, \sigma^2)$	$\theta\mu + \frac{1}{2}\theta^2\sigma^2$	$\dfrac{1}{2\,\sigma^2}\,(x-\mu)^2$
$\mathcal{B}(p)$	$\log\left(1 - p \,+\, pe^{\theta}\right)$	$\begin{cases} x \log \dfrac{x}{p} + (1-x) \log \dfrac{1-x}{1-p} & 0 \leq x \leq 1 \\ \infty & \text{otherwise} \end{cases}$
$\text{Exp}(\lambda)$	$\log \dfrac{\lambda}{\lambda - \theta}$	$\begin{cases} x\lambda - 1 - \log(x\lambda) & x > 0 \\ \infty & \text{otherwise} \end{cases}$
$\text{Poisson}(\lambda)$	$\lambda\left(e^{\theta} - 1\right)$	$\begin{cases} \lambda + x\left(\log \frac{x}{\lambda} - 1\right) & x > 0 \\ \lambda & x = 0 \\ \infty & \text{otherwise} \end{cases}$

Theorem 1 (Cramér's theorem). *Let X_i be IID (real valued) RVs and define*

$$S_n = \sum_{i=1}^{n} X_i \ .$$

Let $\Lambda(\theta)$ denote the Logarithmic Moment Generating Function of X_i, i.e.,

$$\Lambda(\theta) \ = \ \log \mathbb{E}\left(e^{\theta X_i}\right)$$

and let Λ^* be the convex conjugate of Λ:

$$\Lambda^*(x) \ \overset{\Delta}{=} \ \sup_{\theta} \left(\theta x - \Lambda(\theta)\right) \ . \tag{10}$$

For all closed sets F,

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{S_n}{n} \in F\right) \leq - \inf_{x\in F} \Lambda^*(x) \qquad \textit{Upper Bound for closed sets} \tag{11}$$

and, for all open sets G,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{S_n}{n} \in G\right) \geq - \inf_{x\in G} \Lambda^*(x) \qquad \textit{Lower Bound for open sets} \tag{12}$$

i.e., for any set $B \subset \mathbb{R}$:

$$\begin{aligned} - \inf_{x\in B^o} \Lambda^*(x) \ &\leq \ \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{S_n}{n} \in B\right) \\ &\leq \ \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{S_n}{n} \in B\right) \leq \ - \inf_{x\in \bar{B}} \Lambda^*(x) \end{aligned} \tag{13}$$

where B^o denotes the interior of B and $\bar{B}$ its closure.

A complete proof of the theorem goes beyond the scope of this tutorial and can be found, for instance, in [6] or, in a more general form, in [5]. However, it is quite useful to draw here some general remarks:

1. In the theorem, no conditions, not even existence of the mean, are required for the RVs X_i.
2. The **Lower Bound** (12) is *local* (the bound for open balls implies the bound for all open sets) and its proof uses an *exponential change of measure* [9] argument, as in Importance Sampling (more on Importance Sampling and LDT-based changes of measures in section 4.2)

$$\frac{d\mu_\theta}{d\mu}(x) \ = \ e^{\theta x - \Lambda(\theta)} \ = \ \frac{1}{M(\theta)}\, e^{\theta x} \tag{14}$$

where μ and μ_θ denote the law of the original and tilted RVs respectively.

In order to derive a bound on the probability that the sample mean S_n/n lies in $\left(x - \delta, x + \delta\right)$ we seek a tilt parameter θ^* that makes the mean of the tilted distribution equal to x. From a heuristic point of view, this tilted RV captures the idea of being close in distribution to X_i, conditional on having a value close to x.

Indeed, the tilted measure μ_θ identifies the *most likely way* by which the mean of a large sample turns out to be close to x. More precisely, conditional on the sample mean S_n/n being in $(x-\delta, x+\delta)$, the empirical distribution of $X_1, \ldots, X_n$ approaches μ_θ as $n \to \infty$.

3. The **Upper Bound (Chernoff's Bound)** holds for all closed sets $F \subset \mathbb{R}$ and *all* n, not just on a logarithmic scale in the limit as $n \to \infty$. This means that (11), presented as a *classical* LDT upper bound, in case of sums of IID RVs can be strengthened as follows:
$$\frac{1}{n} \log \mathbb{P}\left(\frac{S_n}{n} \in F\right) \leq -\inf_{x \in F} \Lambda^*(x) \ . \tag{15}$$
4. When **the limit exists** (i.e., limsup and liminf are equal), the Cramér's Theorem implies that
$$\mathbb{P}\left(\frac{S_n}{n} \in B\right) \approx e^{-n \inf\limits_{x \in B} \Lambda^*(x)} \ . \tag{16}$$
The last approximation highlights three important features of LDT:
   (a) The asymptotic probability that the sample mean lies in B tends to zero *exponentially fast (in n).*
   (b) Λ^* gives the *exact* (ignoring terms that are subexponential in n) decay rate of the family of probabilities $\mathbb{P}(M_n \in B)$ and is commonly known as *rate function.*
   (c) The speed of convergence essentially depends on *one point*, denoted in the following as $\hat{x}$, the so-called *dominating point* of the set B, i.e., the point where the rate function $\Lambda^*(x)$ attains its infimum (*principle of the largest term*). For instance, the three sets in fig. 2 (which refers to the sums of exponential RVs with mean 1) have the same probability in the large deviations limit.
5. Since Cramér's theorem *only* gives *logarithmic asymptotics*, (16) implies that
$$\mathbb{P}\left(\frac{S_n}{n} \in B\right) = \phi(n)\, e^{-n\Lambda^*(\hat{x})}$$
for some subexponential (at ∞) function $\phi(\cdot)$
$$n^{-1} \log \phi(n) \to 0 \qquad \text{as } n \to \infty \ .$$
For instance, $\phi(n)$ can be any polynomial function n^α or even $\exp(n^{1-\epsilon})$; this means that the LDT approximation may be very inaccurate and better results are sometimes available (for instance, the Bahadur-Rao exact asymptotics [10] for Normal RVs). On the other hand, in many cases LDT represents the only available analytical tool for the analysis of complex systems.
6. Cramér's theorem has a multivariate counterpart [5] dealing with the large deviations of the empirical mean of IID random vectors X_i in $\mathbb{R}^d$. In that case, the

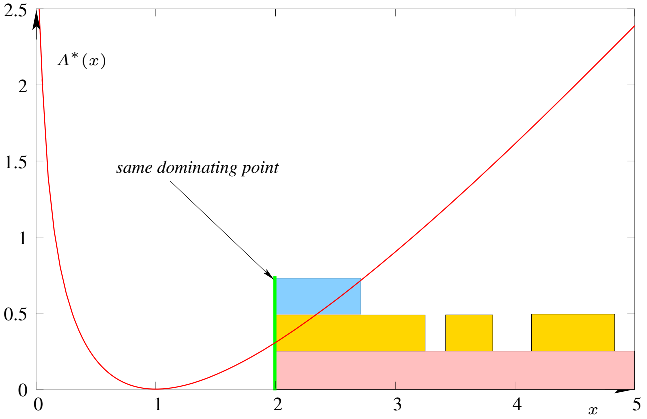

Fig. 2. Application of Cramér's theorem to sums of Exp(1) RVs

definition of the logarithmic cumulant generating function is the straightforward generalisation of (7):

$$\Lambda(\theta) \overset{\Delta}{=} \log M(\theta) = \log \mathbb{E}\left(e^{\langle \theta, X_i \rangle}\right) , \tag{17}$$

where

$$\langle \theta, x \rangle = \sum_{j=1}^{d} \theta_j x_j$$

is the usual scalar product in $\mathbb{R}^d$ and x_j denotes the j^{th} component of x.

2.2 General Principles of Large Deviation Theory

The general theory of large deviations has a beautiful and powerful formulation due to Varadhan [7], based on the so-called *Large Deviation Principle* (LDP), which leads to asymptotic results similar to (13), but under more general conditions.

Let S_n be any sequence of RVs, not necessarily the partial sum of IID RVs. In general, Cramér's theorem cannot be invoked *as is* (for instance if the RVs are correlated, as it often happens in computer networks); however, the "scaled" sequence[4] S_n/n may happen to show the *same asymptotic behaviour* proved for the partial sums of IID RVs. In LDT terms, this means that the sequence S_n/n satisfies an LDP.

The following subsections introduce the definition of LDP (at first in $\mathbb{R}^d$ and then its generalisation in Hausdorff spaces) and the main tools that can be used *to build* an LDP.

[4] In some cases (see section 4.1 for a relevant application in the field of network performance) it will be necessary to change the scaling factor and consider the sequences S_n/v_n for an adequate choice of the deterministic scaling factors v_n.

Large Deviations Principle in $\mathbb{R}^d$. In its abstract formulation [5], the *large deviation principle* characterises the limiting behaviour of a family of Borel probability measures on a Hausdorff space in terms of a *rate function*.

As in [6], to make the concept more intuitive to non-specialists, preliminary definitions of rate function (not simply the convex conjugate of the Logarithmic Moment Generating Function) and LDP are given in the framework of $\mathbb{R}^d$-valued RVs.

In the following $\mathbb{R}^*$ will denote the extended real numbers, $\mathbb{R} \bigcup \{\infty\}$.

Definition 1 (Rate function). *A function $I : \mathbb{R}^d \to \mathbb{R}^*$, is a* rate function *if*

1. *$I(x) \geq 0$ for all $x \in \mathbb{R}^d$*
2. *I is lower semicontinuous, i.e., the level sets $\{x : I(x) \leq \alpha\}$ are all closed, for $\alpha \in \mathbb{R}$*
3. *It is called a* good rate function *if in addition the level sets are all compact*

The definition of lower semicontinuity implies that I is allowed to jump down, but not to jump up; indeed, a function I is lower semicontinuous (according to a definition equivalent to the previous one) iff

$$\text{whenever } x_n \to x \qquad \liminf_{n\to\infty} I(x_n) \;\geq I(x) \;.$$

It is easy to verify, for instance, that Λ^* in Cramér's theorem is a *good rate function*, where the term "good" has been introduced to highlight that some LDT results (such as the widely used contraction principle) only hold if the rate function has this additional property (i.e., if the level sets are not only closed, but also compact).

Definition 2 (Large Deviations Principle). *Let $(X_n, n \in \mathbb{N})$ be a sequence of RVs taking values in $\mathbb{R}^d$. X_n satisfies a* large deviations principle *in $\mathbb{R}^d$ with rate function I if I is a rate function and, for any measurable set $B \subset \mathbb{R}^d$*

$$\begin{aligned} -\inf_{x\in B^o} I(x) \;&\leq\; \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(X_n \in B\right) \\ &\leq\; \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(X_n \in B\right) \;\leq\; -\inf_{x\in \bar{B}} I(x) \end{aligned} \tag{18}$$

where B^o denotes the interior of B and $\bar{B}$ its closure.

For example, Cramér's theorem states that the empirical mean S_n/n of IID RVs satisfies an LDP with good rate function Λ^* given by (10). Further examples of LDP will be given in the following sections.

Gärtner-Ellis Theorem. The *Gärtner-Ellis theorem* defines under which hypotheses the sequence S_n/n satisfies an LDP (for instance, an LDP can be derived for dependent random processes, such as Markov chains and autoregressive processes) and says how to calculate the corresponding rate function.

Roughly speaking, the generalisation of Cramér's theorem to any sequence of RVs mainly relies on the existence of a sufficiently "well-behaved" non trivial *limiting scaled cumulant generating function*

$$\Lambda(\theta) = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E} e^{\theta S_n} \tag{19}$$

and, given the existence of the exponential moments of S_n, this essentially requires that the autocorrelation of the increments of S_n decays sufficiently fast. For instance, this result can be used to prove an LDP for a queue with a weakly dependent input flow [6]. To state the theorem properly (for sake of generality, for random vectors in $\mathbb{R}^d$), it is useful to recall the following definition:

Definition 3 (Essential smoothness). *A convex function $\Lambda : \mathbb{R}^d \rightarrow \mathbb{R}^*$ is* essentially smooth *if*

1. *$(\mathcal{D}_\Lambda)^o$ in non-empty*
2. *$\Lambda(\cdot)$ is differentiable throughout $(\mathcal{D}_\Lambda)^o$*
3. *$\Lambda(\cdot)$ is* steep, *i.e., for any sequence θ_n in $(\mathcal{D}_\Lambda)^o$ which converges to a boundary point of $\mathcal{D}_\Lambda$*
$$\lim_{n\to\infty} |\nabla\Lambda(\theta_n)| = +\infty$$

where $\mathcal{D}_\Lambda$ denotes the effective domain *of $\Lambda(\cdot)$, i.e.,*

$$\mathcal{D}_\Lambda = \left\{\theta : \Lambda(\theta) < \infty\right\}$$

Theorem 2 (Gärtner-Ellis Theorem). *Let S_n be a sequence of random vectors in $\mathbb{R}^d$ with cumulant generating functions:*

$$\Lambda_n(\theta) = \log \mathbb{E}\left(e^{\langle\theta, S_n\rangle}\right) \quad . \tag{20}$$

Assume that:

1. *The limiting scaled cumulant generating function*
$$\Lambda(\theta) = \lim_{n\to\infty} \frac{1}{n} \Lambda_n(\theta) \tag{21}$$
exists in $\mathbb{R}^*$ for each $\theta \in \mathbb{R}^d$
2. *$\Lambda(\theta)$ is finite in a neighbourhood of $\theta = 0$, i.e., $0 \in (\mathcal{D}_\Lambda)^o$*
3. *Λ is essentially smooth and lower-semicontinuous.*

Then, the sequence S_n/n satisfies an LDP in $\mathbb{R}^d$ with good rate function Λ^*

$$\Lambda^*(x) \overset{\Delta}{=} \sup_{\theta\in\mathbb{R}^d} \left(\langle\theta, x\rangle - \Lambda(\theta)\right) \quad . \tag{22}$$

To illustrate the meaning of the Gärtner-Ellis theorem, it is useful to consider the empirical mean S_n/n of real-valued RV X_i, where

$$S_n = X_1 + X_2 + \cdots + X_n \quad ,$$

under different correlation structures (calculations may be found in [6]):

- **IID RVs:** it is trivial to prove that
$$\Lambda_n(\theta) \overset{\Delta}{=} \log \mathbb{E}e^{\theta S_n} = n \log \mathbb{E}e^{\theta X_i}$$
and hence the rate function given by (22) coincides with (10). This explains why the Gärtner-Ellis theorem is sometimes (for instance, in [6]) referred to as the *generalised Cramér's theorem.*

- **Additive functionals of Markov chains:** let $\left(\xi_n, n \in \mathbb{N}\right)$ be an irreducible Markov chain, taking values in a finite set E with transition matrix $P = \left\{p_{ij}\right\}$. Let f be a function from E to $\mathbb{R}$ and define $X_n = f(\xi_n)$; finally, let $Q(\theta)$ denote the (non-negative irreducible) $E \times E$ matrix whose $\{ij\}$-element is
$$q_{ij}(\theta) \;=\; e^{\theta f(i)}\, p_{ij}$$
and let $\rho(\theta)$ denote its spectral radius (Perron-Frobenius eigenvalue).
Then, S_n/n satisfies an LDP with
$$\Lambda(\theta) \;=\; \log \rho(\theta) \ .$$
In queueing applications, this result is quite relevant, since it permits to identify the rate function for Markov-modulated fluid sources (S_n/n represents the average data rate over n time slots).
- **Gaussian autoregressive processes:** the samples X_i are defined by the (stable) recursion
$$X_i = \sum_{k=1}^{r} a_k X_{i-k} + \epsilon_i \qquad i \in \mathbb{Z}$$
where the ϵ_i are independent standard normal RVs. The covariance structure of $(X_i,\ i \in \mathbb{Z})$ is usually described through its Fourier transform
$$\mathcal{S}_X(\omega) \;=\; \sum_{k=-\infty}^{\infty} \mathbb{E}\left(X_0\, X_k\right)\, e^{i\omega k}$$
which is called the power spectral density of the process. It is easy to show that $\mathcal{S}_X(\omega) \;=\; \left|A(\omega)\right|^2$, where
$$A(\omega) \;\triangleq\; 1 - \sum_{j=1}^{r} a_j e^{i\omega j} \ .$$
Then, S_n/n satisfies an LDP with rate function
$$I(x) \;=\; \frac{x^2}{2\mathcal{S}_X(0)} \ .$$
It is worth mentioning that different Gaussian processes having the same power spectral density at zero have the same rate function. The underlying assumption is that $\mathcal{S}_X(\omega)$ is finite and differentiable on $\left[-\pi, \pi\right]$. This basically requires that the correlations decay sufficiently fast; for LRD processes the spectrum has a singularity at zero and, to use LDT, it will require a different scaling in n (see section 4.1).

Large Deviations Principle in a Hausdorff space. In the study of queueing systems, it is sometimes useful to consider the large deviations of the sample mean of random processes (i.e., infinitely dimensional objects); for instance, Schilder's theorem gives an expression for the probability of the sample mean (which is now a *path*, i.e., a function of time) of n IID Gaussian processes being in some set $\mathcal{S}$. To include such results in the general theory, it is necessary to rephrase the large deviation principle in a more powerful way, making use of the classical abstract language of LDT.

Definition 4 (Large Deviations Principle). *Let $(\mu_n, n \in \mathbb{N})$ be a sequence of Borel probability measures on a Hausdorff space $\mathcal{X}$ and let $\mathcal{B}$ be the Borel σ-algebra.*
μ_n satisfies a large deviations principle *on $\mathcal{X}$ with rate function I if I is a rate function and, for all $B \in \mathcal{B}$*

$$\begin{aligned} -\inf_{x \in B^o} I(x) &\leq \liminf_{n\to\infty} \frac{1}{n} \log \mu_n(B) \\ &\leq \limsup_{n\to\infty} \frac{1}{n} \log \mu_n(B) \leq -\inf_{x\in\bar{B}} I(x) \end{aligned} \tag{23}$$

A few comments permit to better clarify the LDP concept in its general form:

1. If X_n is a sequence of RVs with distribution μ_n, then we may equivalently say that the sequence X_n satisfies the LDP.
2. If $\mathcal{X}$ is a space of functions indexed by $\mathbb{R}$ or $\mathbb{N}$, the LDP is usually called *sample path LDP.*
3. If X_n satisfies an LDP in a regular Hausdorff space $\mathcal{X}$ with rate function I, and with rate function J, then $I = J$ (uniqueness of the rate function).
4. A set $A \subset \mathcal{X}$ is called an I-continuity set if

$$\inf_{x\in A^o} I(x) = \inf_{x\in\bar{A}} I(x) \ .$$

For such a set (for instance, if $\mathcal{X} = \mathbb{R}$ and I is continuous, then all intervals are I-continuity sets), if it is measurable, then (23) becomes

$$\lim_{n\to\infty} \frac{1}{n} \log \mu_n(A) = -\inf_{x\in A} I(x) \ . \tag{24}$$

Starting from the existence of an LDP, it is possible to give a precise definition (see [6] for the proof) of one of the *most famous* LDT results, the *principle of the largest term.*

Indeed, if I is a good rate function and $A \subset \mathcal{X}$ is closed, then the infimum is attained at some $\hat{x} \in A$. This $\hat{x}$ is the most likely way for an event A to occur, since $I(\hat{x})$ dominates in $\mathbb{P}(X_n \in A)$.

Theorem 3 (Rare events occur in the most likely way). *Suppose X_n satisfies an LDP with good rate function I, and C is a closed set with*

$$\inf_{x\in C} I(x) = k < \infty \ .$$

This infimum must be attained; suppose it is attained in C^o and let B be a neighbourhood of $\{x \in C : I(x) = k\}$. Then

$$\mathbb{P}(X_n \notin B \mid X_n \in C) \to 0 \ . \tag{25}$$

The Contraction Principle. The contraction principle is one of the most useful tools in LDT; indeed, once we have an LDP for one sequence of RVs, we can *effortlessly*[5]

[5] At least in principle; in practise it might be quite difficult to establish the continuity of a given function, and to compute the resulting rate function.

establish LDPs for a whole other class of random sequences, obtained via continuous transformations.

For example, in queueing applications, starting from the LPD for the arrival process, if a quantity of interest can be written as a continuous function (in some Hausdorff space) of the arrivals, then it will be possible to deduce an LDP for that quantity.

Theorem 4 (Contraction Principle). *Let $\mathcal{X}$ be a Hausdorff space and suppose that X_n satisfies an LDP in $\mathcal{X}$ with* good rate function *I, and that $f : \mathcal{X} \to \mathcal{Y}$ is a continuous map to another Hausdorff space $\mathcal{Y}$.*
Then $f(X_n)$ satisfies an LDP in $\mathcal{Y}$, with good rate function

$$J(y) = \inf_{x \in \mathcal{X} : f(x) = y} I(x) \ . \tag{26}$$

Although the proof of the theorem is rather technical (mainly to prove that J is a good rate function), it is easy to give a heuristic justification taking into account the basic idea behind the LDT limits in the spirit of (16):

$$\mathbb{P}\left(f(X_n) \approx y\right) \approx \mathbb{P}\left(X_n \approx f^{-1}\left(\{y\}\right)\right) \approx$$

$$\approx e^{-n \ \inf_{x \in f^{-1}(\{y\})} I(x)} = e^{-n \ \inf_{x : f(x) = y} I(x)}$$

Unfortunately, the hypotheses of the contraction principle are too restrictive for its application in the framework of *many flows scaling*; hence in [6] a generalisation is given, in which Y_n is only exponentially equivalent to $f(X_n)$ (i.e., the probability they differ even by ϵ decays superexponentially for all $\epsilon > 0$) and f is continuous only on the subspace where the rate function is finite.

2.3 Sample Path Large Deviations

In many applications, interest lies in the probability that a *path* of a random process hits a particular set; the LDT tools are to be developed in an infinitely dimensional framework and quite often are rather abstract. For sake of brevity, only Gaussian processes are considered in this section, since they represent a widely used model for aggregated traffics [3].

More in detail, as an example of Sample Path LDT, the Schilder's theorem (for Brownian motion) is deeply discussed and then the result is extended to a wider class of Gaussian processes.

Schilder's Theorem. Schilder's theorem analyses the most likely paths of a standard Brownian motion $B(t)$, while a sample path LDP for a generic random walk is given by the Mogulskij's theorem [5].

Before stating the theorem, it might be useful to recall the main properties of the Brownian motion and the definition of absolutely continuous function.

Definition 5 (standard Brownian motion). *A standard Brownian motion is characterised by the following properties:*

- *$B(\cdot)$ is Gaussian, i.e., its finite-dimensional distributions are multivariate normal*
- $B(t) \in \mathcal{N}(0,t)$
- $B(0)=0$
- *$B(t)$ has* independent *increments, i.e., $(B(t+u)-B(u))$ is independent of $B(u)$ and*

$$B(t+u)-B(u) \in \mathcal{N}(0,t)$$

- *$B(\cdot)$ has continuous sample paths*

Definition 6 (Absolutely continuous function). *A function $f : [0,1] \to \mathbb{R}$ is* absolutely continuous *if for all $\epsilon > 0$ there exists a $\delta > 0$ such that for every finite collection of non-overlapping intervals $\{[s_i,t_i], 1 \le i \le N\}$*

$$\sum_{1\le i\le N} (t_i - s_i) < \delta \qquad \Rightarrow \qquad \sum_{1\le i\le N} |f(t_i)-f(s_i)| < \epsilon$$

Theorem 5 (Schilder's Theorem). *Let $(B(t), t \in [0,1])$ be a standard Brownian motion, taking values in $\mathcal{C}[0,1]$, the space of continuous functions $f : [0,1] \to \mathbb{R}$ equipped with the supremum norm:*

$$\|f\| = \sup_{0\le t\le 1} |f(t)| \ .$$

Then $\left(B^n(t) \overset{\Delta}{=} \frac{1}{\sqrt{n}}\, B(t), n \in \mathbb{R}^+\right)$ satisfies a sample path LDP in $\mathcal{C}[0,1]$ with good rate function

$$I(f) = \begin{cases} \dfrac{1}{2}\displaystyle\int_0^1 \dot{f}(t)^2 dt & \textit{if } f \textit{ is absolutely continuous and } f(0)=0 \\ \infty & \textit{otherwise} \end{cases} \tag{27}$$

A heuristic argument, based on the Cramér's theorem for Gaussian RVs, can lead to a simple (and instructive) justification of (27). A rigorous proof of the theorem and its extension to $[0,T]$ (for any $T<\infty$) can be found in [5].

Let $\boldsymbol{\Pi_K} f$ be the polygonalised version of f, i.e., the piecewise linear approximation of f at n/K, $(0 \le n \le K)$; then, assuming $f(0)=0$ (otherwise the probability is 0 since, by definition, $B(0)=0$; hence if $f(0)\neq 0$, the corresponding rate function should be $I(f)=\infty$)

$$\mathbb{P}\left(B^n(\cdot) \approx f(\cdot)\right) \approx \mathbb{P}\left(\boldsymbol{\Pi_K} B^n(\cdot) \approx \boldsymbol{\Pi_K} f(\cdot)\right) \ .$$

Since $B^n(\cdot)$ has independent increments, the latter can be written as

$$\prod_{i=0}^{K-1} \mathbb{P}\left(B^n\left(\frac{i+1}{K}\right) - B^n\left(\frac{i}{K}\right) \approx f\left(\frac{i+1}{K}\right) - f\left(\frac{i}{K}\right)\right)$$

and, taking into account that $B^n(t) \in \mathcal{N}\left(0, \frac{t}{n}\right)$,

$$\prod_{i=0}^{K-1} \mathbb{P}\left(\mathcal{N}\left(0, \frac{1}{nK}\right) \approx f\left(\frac{i+1}{K}\right) - f\left(\frac{i}{K}\right)\right) \ .$$

Since in the LDT limit

$$\mathbb{P}\left(\mathcal{N}\left(0, \frac{\sigma^2}{L}\right)\right) \approx e^{-L\frac{1}{2\sigma^2}x^2} \ ,$$

it is easy to show that

$$\begin{aligned}\frac{1}{n}\log \mathbb{P}\left(B^n(\cdot) \approx f(\cdot)\right) &\approx -\frac{K}{2}\sum_{i=0}^{K-1}\left(f\left(\frac{i+1}{K}\right) - f\left(\frac{i}{K}\right)\right)^2 \\ &= -\frac{1}{2}\sum_{i=0}^{K-1}\frac{1}{K}\left(\frac{f\left(\frac{i+1}{K}\right) - f\left(\frac{i}{K}\right)}{1/K}\right)^2 \\ &\xrightarrow{K\to\infty} -\frac{1}{2}\int_0^1 \dot{f}(t)^2\, dt \ ,\end{aligned}$$

which gives the expression of rate function $I(f)$ when $f(0) = 0$:

$$I(f) = \frac{1}{2}\int_0^1 \dot{f}(t)^2 dt \ .$$

The previous computations highlight that, at least informally, multivariate Cramér's theorem can be seen as a special (finite-dimensional) case of Schilder's theorem. Moreover, it is worth noticing that the cost of a path f is exclusively determined by the derivative along the path.

Generalised Schilder's Theorem. In [11], Schilder's theorem is extended to the general case of a (non-trivial) centred Gaussian process $A(t)$, with $A(0) = 0$ and stationary increments. The variance function of $A(\cdot)$ is denoted by $v(t)$; the standard Brownian motion $B(\cdot)$ is only a special case, with $v(t) = t$, for which the resulting expressions are relatively transparent (as shown in the previous section). A key role in the following will be played by the covariance function of $A(\cdot)$:

$$\Gamma(s,t) = \mathrm{Cov}\left(A(s), A(t)\right) = \frac{1}{2}\left(v(t) + v(s) - v(|t-s|)\right) \ . \tag{28}$$

An intrinsic difficulty of the generalisation of Schilder's theorem is that the rate function $I(\cdot)$ cannot be given explicitly. The case of Brownian motion is an exception: indeed, due to the *independence* of the increments, it was possible to derive an explicit formula for $I(f)$ (see the heuristic justification of the theorem). To state the general theorem, it is necessary to introduce a *path state* Ω and a *reproducing kernel Hilbert space* R, equipped with inner product $\langle\cdot,\cdot\rangle_R$ and norm $\|\cdot\|_R$.

The *path space* Ω is defined as

$$\Omega = \left\{\omega : \mathbb{R} \to \mathbb{R},\ \text{continuous},\ \omega(0) = 0,\ \lim_{t\to\pm\infty}\frac{\omega(t)}{1+|t|} = 0\right\}$$

equipped with the norm

$$\|\omega\|_\Omega = \sup_{t\in\mathbb{R}}\frac{|\omega(t)|}{1+|t|} \ .$$

$A(\cdot)$ can be realised on Ω under the assumption that $v(\cdot)$ increases slower than quadratically. For example, this is the case for fractional Brownian motion, one of the most relevant LRD traffic models, which is characterised by $v(t) = t^{2H}$, where $1/2 < H < 1$ (see section 4.1 for a discussion on LRD and its implications on traffic engineering).

In addition to Ω, a central role is played by a linear subspace of Ω, which consists of smoother functions than the typical paths of $A(\cdot)$ and which can be given a Hilbert space structure. This space, the *reproducing kernel Hilbert space* R, is defined starting from the set of functions $\left\{\Gamma\left(s,\cdot\right)\right\}$, equipped with the inner product

$$\langle \Gamma\left(s,\cdot\right), \Gamma\left(\cdot,t\right)\rangle_R \;=\; \Gamma\left(s,t\right) \;\; .$$

The space R is obtained by closing this set of functions with linear combinations, and completing with respect to the norm

$$\|\omega\|_R^2 \;=\; \langle \omega, \omega\rangle_R \;\; .$$

The inner product definition generalises to the *reproducing kernel property*:

$$\langle \omega, \Gamma\left(s,\cdot\right)\rangle_R \;=\; \omega(s) \qquad \omega \in R \;\; . \tag{29}$$

To give a heuristic understanding of the space R, let us consider a centred Gaussian distribution on $\mathbb{R}^d$. In this case, the space R is $\mathbb{R}^d$ itself, but equipped with an inner product such that the density of the distribution can be written as

$$f(x) \;=\; \mathit{const} \cdot \exp\left(-\frac{1}{2}\,\|x\|_R^2\right) \;\; .$$

Thus, minimising $\|\cdot\|_R$ corresponds to maximising the density.

Theorem 6 (Generalized Schilder's Theorem). *Let $A(\cdot) \in \Omega$ be a (non trivial) centred Gaussian process, with variance function $v(t)$.*
Then $\left(\frac{1}{\sqrt{n}}\,A(\cdot), n \in \mathbb{R}^+\right)$ satisfies a sample path LDP in Ω with good rate function

$$I(f) \;=\; \begin{cases} \frac{1}{2}\|f\|_R^2 & \mathit{if}\; f \in R \\ \infty & \mathit{otherwise} \end{cases} \tag{30}$$

In [10], the generalised Schilder's theorem is written in terms of the *sample-mean path*

$$\frac{1}{n}\sum_{i=1}^{l} A_n(\cdot)$$

of a sequence of IID centred Gaussian processes with variance function $v(t)$. Informally, the theorem gives an expression for the probability of the *sample-mean path* being in some set $\mathcal{S}$ (that represents a collection of paths):

$$\mathbb{P}\left(\frac{1}{n}\,\sum_{i=1}^{n} A_i(\cdot) \in \mathcal{S}\right) \;\approx\; \exp\left(-n \inf_{f\in\mathcal{S}} I(f)\right) \;=\; \exp\left(-\frac{n}{2}\inf_{f\in\mathcal{S}}\|f\|_R^2\right) \;\; .$$

Hence, the probability decays exponentially in n and the corresponding exponential decay rate equals the minimum of $I(f)$ over all $f \in \mathcal{S}$.

The minimising $\hat{f}(\cdot)$ corresponds to the *most likely path* in $\mathcal{S}$. Conditional on the sample-mean path being in the set $\mathcal{S}$, with overwhelming probability this happens via a path that is *close to* $\hat{f}$. In other words, as $n \to \infty$

$$\frac{1}{n} \log \mathbb{P} \left(\frac{1}{n} \sum_{i=1}^{n} A_i(\cdot) \in \mathcal{S} \right) \to -I \left(\hat{f} \right)$$

and the decay rate is fully dominated by the likelihood of the most likely element in $\mathcal{S}$.

Unfortunately, finding the minimum of $I(f)$ over all $f \in \mathcal{S}$ is, in general, a hard variational problem. Indeed, the optimisation should be done over all paths in $\mathcal{S}$ (which are infinitely dimensional objects), and, according to (29), the objective function $I(f)$ is only explicitly given if f can be written as a linear combination of covariance functions $\varGamma(s, \cdot)$.

3 Large Deviations for Queues

One of the primary issues in queueing theory is to analyse the (steady-state) buffer content distribution. This problem can be solved explicitly only in a few special cases, and the goal of LDT is to get approximate estimations of the parameters of interest that are sufficiently close to the actual values at least in some asymptotic conditions. In more detail, two asymptotic scalings are usually considered: the *large buffer* regime and the *many-sources* regime, and for both of them LDT permits to obtain logarithmic asymptotics. Although in the following only the overflow probability in stationary condition will be considered, it is worth mentioning that LDT may be applied to estimate other quantities such as the most likely way a queue became big, the exit probability, the distribution of busy periods and even the way steady state is reached (see, as a comprehensive illustrative example, the analysis of the M/M/1 queue in [12]).

The main result in this section is the LDP for the single server queue, which, at least under some restrictive hypotheses, can be established through direct calculation. More general results are achieved making use of abstract LDT tools, such as the contraction principle (in an adequately chosen Hausdorff space), once an LDP for the arrival process is known.

3.1 The Single Server Queue

The evolution of a (FIFO) single server queue is described, in both continuous and discrete time settings, by the Lindley's recursion

$$Q_{n+1} = (Q_n + X_{n+1})^+ \tag{31}$$

where $x^+ = \max(0, x)$ and the interpretation of the different entities depends on the specific settings.

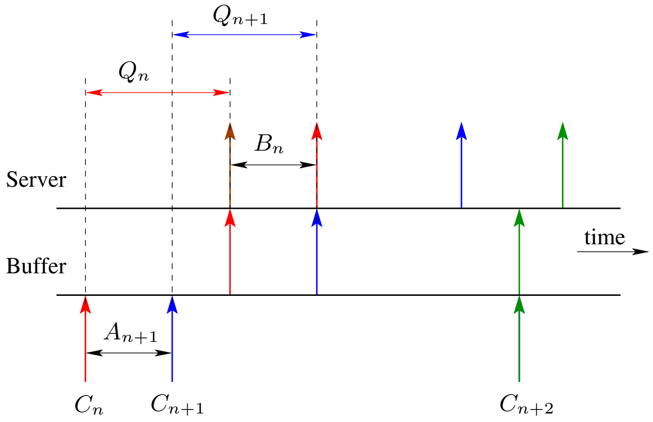

Fig. 3. Continuous time version of Lindley's recursion

In continuous time (see fig. 3), customers are labelled by the integers (C_n denotes the n^{th} arrival) and X_{n+1} is the difference between the service time B_n of C_n and the interarrival time A_{n+1} between C_n and C_{n+1}; in this case Q_n gives the waiting time of C_n, i.e., the time spent in the queue before commencing service.

In discrete time (slotted time model) the interpretation of (31) is straightforward: X_n is the difference between the amount of work A_n that arrives at the queue at time n (or, more in general, during the interval $\big(n-1, n\big)$) and the amount of work C_n that the server can process at time n; hence, Q_n represents the amount of work remaining in the queue just after time n. In the following we shall adopt the latter interpretation, but most of the results can be easily adapted to waiting time in the continuous time framework.

In the rather common case of constant rate server (which can be used, for instance, to model a transmission line), Lindley's recursion becomes

$$Q_{n+1} \;=\; (Q_n + A_{n+1} - C)^+ \tag{32}$$

Equation (32) may have different solutions, depending on boundary conditions (see [6,10] and references therein for a detailed analysis). For example, let us consider the queue size at time $n = 0$, subject to the boundary condition that the queue was empty at $n = -\infty$. It is easy to prove that

$$Q_0^{-\infty} \;=\; \sup_{n\geq 0}\; S_n - Cn \tag{33}$$

where S_n is the cumulative arrival process, i.e.,

$$S_n \;\stackrel{\Delta}{=}\; A_{-n+1} + A_{-n+2} + \cdots + A_{-1} + A_0$$

and, by convention, $S_0 = 0$. If the arrival process $(A_n, n \in \mathbb{Z})$ is stationary, then $Q_0^{-\infty}$ has the same distribution as $Q_n^{-\infty}$ for any $n \in \mathbb{Z}$ and this distribution is called the *steady state distribution* of queue size. Moreover, if $(A_n, n \in \mathbb{Z})$ is also ergodic and $\mathbb{E}A_0 < C$ (i.e., $\mathbb{E}X_0 < 0$), then the limit does not depend on the initial condition.

In the following, $(X_n, n \in \mathbb{Z})$ will be a stationary ergodic sequence of RVs with $\mathbb{E}X_0 < 0$ (stability condition) and Q will denote the unique equilibrium distribution, i.e., in case of constant rate server:

$$Q = \sup_{n \geq 0} S_n - Cn \ . \tag{34}$$

In other word, the steady-state buffer content (a reflected additive recursion, which can only assume non negative values) is distributionally equal to the supremum of a free (i.e., nonreflected) process with negative drift (the supremum is non-negative!).

To conclude this overview on Lindley's recursion, it is worth noticing that this framework, although it has been developed for a slotted system, can be directly extended (Reich's theorem) to the *steady-state queue length in continuous time*

$$Q = \sup_{t \geq 0} A(-t, 0) - Ct \ ,$$

where $A(s,t)$ denotes the amount of traffic offered to the system in $\left[s,t\right)$. Moreover, if the arrival process is time reversible, then

$$Q = \sup_{t \geq 0} A(t) - Ct \ . \tag{35}$$

3.2 LDT Asymptotics

Since LDT is an asymptotic theory, the solution of Lindley's recursion is analysed in some asymptotic conditions. In particular, two different regimes (large buffer and many sources) are discussed in the next subsections, presenting the key results and highlighting the ideas behind the proofs, which can be found in [6] together with illustrative examples.

Large-buffer regime. In the large-buffer regime, traditionally the most investigated limit (and not only in the field of LDT), the objective is to find asymptotic expansions of the queue size complementary probability $\mathbb{P}(Q > q)$ for $q \to \infty$.

In this section we will consider a (single server FIFO) queue with constant service rate C and arrival process $(A_t, t \in \mathbb{Z})$, A_t being the amount of work arriving at time t. In [6] an LDP for queue size is derived, at first, under the assumption that the A_t were IID and then weakening this assumption as in the Gärtner-Ellis theorem (see section 2.2). For sake of brevity, here only the more general statement is given, followed by some remarks on the proof and on the *interpretation* of the LDP.

Theorem 7 (LDP for queue size). *Let* $(A_t, t \in \mathbb{Z})$ *be a stationary random process, with* $\mathbb{E}A_0 < C$ *and let*

$$\Lambda_t(\theta) = \log \mathbb{E}e^{\theta S_t} \ .$$

Suppose that

1. *the limit*

$$\Lambda(\theta) = \lim_{t \to \infty} \frac{1}{t} \Lambda_t(\theta) \tag{36}$$

exists in $\mathbb{R}^*$ *for each* $\theta \in \mathbb{R}$

2. *$\Lambda(\theta)$ is essentially smooth, and finite in a neighbourhood of $\theta = 0$*
3. *$\Lambda_t\,(\theta)$ is finite for all t whenever $\Lambda(\theta) < \theta C$*

Then, for $q > 0$:

$$\lim_{l\to\infty} \frac{1}{l} \log \mathbb{P}\left(\frac{Q}{l} > q\right) \;=\; -I(q) \tag{37}$$

where

$$\begin{aligned} I(q) \;&=\; \inf_{t\in\mathbb{R}^+} \; t\,\Lambda^*\left(C + q/t\right) \\ &=\; \inf_{t\in\mathbb{R}^+} \sup_{\theta\geq 0} \; \theta\left(q + Ct\right) \;-\; t\,\Lambda(\theta) \\ &=\; q \,\sup\left\{\theta > 0 \;:\; \Lambda(\theta) < \theta C\right\} \end{aligned} \tag{38}$$

Some remarks may be useful to better understand the theorem.

1. Equation (37) is usually written in a visually simpler equivalent form

$$\lim_{q\to\infty} \frac{1}{q} \;\log \mathbb{P}\left(Q\; > q\right) \;=\; -I(1) \tag{39}$$

   The notation used in the theorem is "closer" to the standard formulation of an LDP. In any case there are two differences:
   (a) Upper and lower bounds happen to agree, so the theorem proves a limit.
   (b) It is a *restricted sort* of LDP, since the theorem only concerns intervals $\left[q, +\infty\right)$ and not general events.
2. The assumption that $\Lambda(\theta)$ is finite in a neighbourhood of the origin is necessary to guarantee the exponential decay of the queue size complementary probability. Completely different behaviours are associated to LRD traffic flows (section 4.1) as well as to heavy-tailed distributions [13].
3. The lower bound is proved by estimating the probability that the queue overflows over some fixed timescale. In other words, the approximation

$$\mathbb{P}\left(\sup_t \; S_t - Ct \geq q\right) \;\approx\; \sup_t \; \mathbb{P}\left(S_t - Ct \geq q\right) \tag{40}$$

   is justified (for large q) on a logarithmic scale.
4. The *most likely time* for the queue to fill up to some high level q is $l\hat{t}$, where $\hat{t}$ is the optimising parameter for $I(1)$ according to its definition in (38). Thus the most likely rate for the queue to build up is $1/\hat{t}$ and does not depend on q.
5. Another interpretation of the theorem is that $S_t - Ct$ is effectively a simple random walk with negative drift, in that

$$\mathbb{P}\left(\sup_t \; S_t - Ct \geq q_1 + q_2\right) \;\approx\; \mathbb{P}\left(\sup_t \; S_t - Ct \geq q_1\right) \mathbb{P}\left(\sup_t \; S_t - Ct \geq q_2\right)$$

   for large q_1 and q_2. Thus, the (weak) dependence of the A_t is invisible at the macroscopic scale (although it does contribute to the value of $I(1)$ through $\Lambda(\theta)$).
6. If the service is a RV C_t, it is possible to apply the theorem to the random process $A_t - C_t$ (rather than to A_t) and set $C = 0$. Under the usual assumption of independence between service and arrival processes, it is easy to show that (if the limits exist)

$$\Lambda(\theta) \;=\; \Lambda_A(\theta) + \Lambda_C(-\theta) \;\;.$$

Many-sources regime. An important limitation of large-buffer regime is that it does not give reasonably accurate results about overflow probability for small buffers, which can be desirable in case of applications with stringent delay constraints. Moreover, it might be useful to take into account that the input traffic can be often seen as the superposition of many IID streams.

These thoughts has led to the interest for the so-called many-sources regime. In this setting it is assumed that the number of source N grows large and, at the same time, the queueing resources (buffer and bandwidth) are scaled accordingly. In more detail, the buffer threshold is replaced by Nq and the service capacity by NC. Despite the fact that the load remains constant, it is clear that the overflow probability decays to 0.

Let $A_t^{(i)}$ denote the amount of work arriving from source i at time t. Assume that

1. for each i, $\left(A_t^{(i)},\ t \in \mathbb{Z}\right)$ is a stationary sequence of RVs
2. these sequences are independent of each other and identically distributed.

If the total amount of work arriving at the queue in the interval $\left(-t, 0\right]$ is denoted by S_t^N, the queue length at time 0 is given by

$$Q^N \ = \ \sup_{t \geq 0} \ S_t^N - NCt$$

and, in the spirit of LRD, we will consider the behaviour of $\mathbb{P}\left(Q^N \geq Nq\right)$ as the number of sources becomes large.

Theorem 8 (LDP for queue size with many sources). *Let S_t^1 be the amount of work produced by a typical source in the interval $\left(-t, 0\right]$ with $\mathbb{E}S_1^1 < C$ and let*

$$\Lambda_t\left(\theta\right) \ = \ \log \mathbb{E} e^{\theta S_t^1} \ .$$

Suppose that

1. *the limit*
$$\Lambda\left(\theta\right) \ = \ \lim_{t \to \infty} \ \frac{1}{t} \ \Lambda_t\left(\theta\right) \tag{41}$$
exists, and is finite and differentiable in a neighbourhood of the origin
2. *for all t, $\Lambda_t\left(\theta\right)$ is finite for θ in a neighbourhood of the origin*

Then

$$\begin{aligned} -I(q+) \ &\leq \ \liminf_{N \to \infty} \ \log \mathbb{P}\left(Q^N > Nq\right) \\ &\leq \ \limsup_{N \to \infty} \ \log \mathbb{P}\left(Q^N > Nq\right) \ \leq \ -I(q) \end{aligned} \tag{42}$$

where

$$I(q) \ = \ \inf_{t \in \mathbb{N}} \Lambda_t^*\left(q + Ct\right) \ = \ \inf_{t \in \mathbb{N}} \sup_{\theta \in \mathbb{R}} \ \theta\left(q + Ct\right) - \Lambda_t\left(\theta\right) \tag{43}$$

It is interesting to point out a few differences with respect to the previous theorem:

1. Expression (42) involves both an upper and a lower bound, as in classical LDT statements. If $\Lambda_t(\cdot)$ is continuous for each t, then the two bounds agree and we obtain a straightforward limit.
2. Once again the proof makes use of the principle of the largest term and of the most likely way in which the rare event may occur. However, in the many-sources limit, the optimising $\hat{t}$ is simply the most likely time to overflow and typically depends on q in a non-linear way.
3. The assumption (41) is a way to control the distribution of S_t^1 for large t and is needed to prove the upper bound (but not the lower bound), although it does not appear in the result ($I(q)$ depends only on Λ_t).
4. The previous condition does not allow for LRD sources; however, even in that case the probability of large queues still decays exponentially in the number of sources (but, as shown in section 4.1, a different scaling will be required in the definition of the limit expression for Λ).

3.3 Continuous Mapping Approach

It is quite complicated to apply the previous approach (based on direct calculation of the rate function for the overflow probability) to other queue parameters and to more complex network scenarios. An interesting alternative is represented by the use of the contraction principle, once the quantity of interest is expressed as a continuous function of all random inputs. In this way it is possible to analyse different service disciplines (priority queue, processor sharing), consider finite buffers and transient behaviours.

More in detail, let A denote any random influence on the network (in general it can be seen as a vector of arrival and service processes). Many relevant quantities (such as the queue size or the departure process at some queue) can be written as functions $f(A)$. In a nutshell, the continuous mapping approach consists of the following steps:

1. Consider a sequence of queueing networks indexed by L, in which the L^{th} network has a vector of inputs A^L, a version of A which is speeded up in time and scaled down in space (the exact scaling depends on the specific framework).
2. Prove a sample path LDP for A^L *in some topological space*.
3. Show that f is continuous *on that space*.
4. Use the contraction principle to derive an LDP for $f(A^L)$.
5. Simplify the resulting rate function (typically, the rate function for this LDP will be given as the solution to a variational problem).

A big advantage of this procedure is that, once a sample path LDP is proved for A^L, we can obtain LDPs for different quantities, under the assumption that they can be written as a continuous function of the inputs.

Another useful consequence of the application of the contraction principle is that we can not only estimate the probability of a rare event, but also find the most likely path to that event.

Large-buffers revisited. In spite of its apparent simplicity, the continuous mapping approach requires some technical work to identify the proper space in which $f(A)$ is actually continuous as well as to simplify the rate function. All these issues are deeply analysed in [6], at first identifying proper continuous queueing maps and then analysing, in two separate chapters, the two classical scaling regimes. Since the main goal of this tutorial is to give an introduction to LDT for non specialists, we will simply focus on the application of the contraction principle to derive an LDP for queues with large buffers.

In a single server queue with deterministic service rate C, the queue size at time 0 can be written as a function of the cumulative arrival process

$$Q_0 \;=\; \sup_{t \geq 0} S_t - C \cdot t \;\stackrel{\Delta}{=}\; f(A)$$

where A denotes the entire input process $(S_t,\ t \geq 0)$. Since it will be more convenient to work in continuous time, we introduce its polygonalised version (with step 1), $\tilde{A} = \boldsymbol{\Pi}_{\mathbf{1}} A$, defined for $t \in \mathbb{R}^+$. At this point it is meaningful to define the scaled processes:

$$\tilde{A}^L(t) \;=\; \frac{1}{L}\tilde{A}(Lt) \tag{44}$$

and the continuous-time version (Reich's theorem) of the queue size function:

$$\tilde{f}(\tilde{A}) \;=\; \sup_{t \in \mathbb{R}^+} \tilde{A}(t) - C \cdot t \ . \tag{45}$$

It is easy to verify through direct substitution that

$$\tilde{f}(\tilde{A}^L) \;=\; L^{-1} f(A) \;=\; L^{-1} Q_0$$

and this means that

$$\mathbb{P}\left(\tilde{f}(\tilde{A}) > b\right) \;=\; \mathbb{P}\left(Q_0 > Lb\right) \ . \tag{46}$$

Let us assume that $\tilde{A}^L(t)$ satisfies an LDP in *some topological space* with *good rate function* I, i.e.,

$$\frac{1}{L}\log \mathbb{P}\left(\tilde{A}^L \in B\right) \approx -\inf_{a \in B} I(a) \ . \tag{47}$$

If $\tilde{f}$ is continuous on *that space*, then the contraction principle gives estimates for the left hand side of (46) and hence for the overflow probability:

$$\frac{1}{L}\log \mathbb{P}\left(Q_0/L > b\right) \approx -J(b) \tag{48}$$

where

$$J(b) = \inf_{a : f(a) > b} I(a) \ . \tag{49}$$

The expression of the rate function J justifies that the probability of a rare event can be estimated by considering only the *optimal manner* for that event to occur. In other words, the most likely way for the rare event $\{Q_0 > Lb\}$ to occur is when the input process $\tilde{A}^L$ is close to the optimising a, which represents the *most likely path to overflow*.

In the previous discussion, we assumed the existence of a topological space in which A^L satisfies an LDP and on which f is continuous. As stated in [6] (where some instructive counterexamples are also shown), this involves a trade-off and, in general, the selection of the proper topological space depends on the application. It turns out that a suitable choice for the single server queue (and for many other systems) is the space C_μ, defined as the set of continuous functions $x : \mathbb{R}^+ \to \mathbb{R}$, for which $x(0) = 0$ and

$$\lim_{t\to\infty} \frac{x(t)}{t+1} = \mu \ , \tag{50}$$

equipped with the topology induced by the *scaled uniform norm*

$$\|x\| = \sup_{t\in\mathbb{R}^+} \left| \frac{x(t)}{t+1} \right| \ . \tag{51}$$

Indeed, the queue size function (45) is continuous on C_μ, where μ represents the mean arrival rate (the polygonalised version of the cumulative arrival process corresponds to $x(t)$ in the definition (50)).

It is worth mentioning that in the many-sources regime it is necessary to work in a *larger space* and use the extended version of the contraction principle. The problem is related to the definition of the mean arrival rate in (50); without going into details (see [6] for the definition of a "proper" space), a simple example is enough to highlight the trouble. Indeed, let us consider N constant rate flows, where each rate is drawn independently from $\mathcal{N}\left(\mu, \sigma^2\right)$; then the limit

$$\lim_{t\to\infty} \frac{S_t^N}{t+1}$$

is not necessarily μ, since it is a RV $\in \mathcal{N}\left(\mu, \sigma^2/N\right)$.

4 Applications of LDT to Networks

The results of the previous sections can be applied to more complex scenarios and to general problems related to network dimensioning and planning. Just to show the heterogeneous capabilities of LDT, two completely different issues will be addressed in this section: the analysis of LRD traffic flows and the use of LDT to speed-up the simulation of rare events through Importance Sampling, which is based on a change of measure argument similar to the one employed in the proof of Cramér's theorem.

4.1 Long Range Dependence and Large Deviations

In the early 1990s, researchers at AT&T [2] claimed, on the basis of a huge collection of high-quality traffic measurements, that Internet traffic presents Long Range Dependence. The main consequences were the search for new traffic models, able to take into account this feature in a parsimonious way, and the analysis of queueing performance under the new traffic paradigm. The first issue has led to the introduction of self-similar (or, more in general, asymptotically self-similar) traffic models, among which the most popular is fractional Brownian motion. The main drawback of these models is the lack of analytical results for queueing performance; indeed, even in the case of a single server queue, only asymptotic results are available.

Basic definitions. For sake of completeness, we recall here the main definitions related to Long Range Dependence and Self-similarity (see, for example, [14] for a complete overview).

Definition 7 (Long Range Dependence). *Let $(X_n,\ n\ \in \mathbb{Z})$ be a second order stationary process with autocorrelation function $\rho(k)$ and power spectral density $\mathcal{S}_X(\omega)$. X_n is* Long Range Dependent *(LRD) iff (the following properties are all equivalent):*

- *$\rho(k)$ decreases as a non summable power law when k tends to infinity*

$$\rho(k)\ \sim k^{-\alpha} \quad \textit{as}\ \ k \to \infty \qquad \textit{where}\ 0 < \alpha < 1$$

- *$\mathcal{S}_X(\omega)$ diverges as an integrable power law near the origin*

$$\mathcal{S}_X(\omega)\ \sim \omega^{-\beta} \quad \textit{as}\ \ \omega \to 0 \quad \textit{where}\ 0 < \beta < 1\ \textit{and}\ \beta = 1-\alpha$$

- *The variance of the aggregated process decays more slowly than the sample size*

$$\mathit{Var}\left(\frac{1}{n}\sum_{i=0}^{n-1} X_i\right)\ \sim n^{-\alpha} \qquad \textit{as}\ \ n \to \infty$$

One related phenomenon is self-similarity: roughly speaking, a dilated portion of the sample path of a self-similar process cannot be (statistically) distinguished from the whole. Indeed, self-similar processes have fluctuation at every time-scale, and the Hurst parameter relates the size of fluctuations to their time-scale according to (52).

Definition 8 (Self-similarity for continuous time processes). *Let $(Y_t,\ t \in \mathbb{R})$ be a continuous time process. Y_t is self-similar with self-similarity parameter H (Hurst parameter) iff*

$$c^{-H} Y_{ct} \overset{(d)}{=} Y_t \qquad \forall\, c > 0 \tag{52}$$

i.e., if for any $k \geq 1$, for any $t_1, t_2, \ldots, t_k \in \mathbb{R}$ and for any $a > 0$

$$\left(Y_{at_1}, Y_{at_2}, \ldots, Y_{at_k}\right) \quad \textit{and} \quad \left(a^H Y_{t_1}, a^H Y_{t_2}, \ldots, a^H Y_{t_k}\right)$$

have the same distribution

Typically self-similar processes are used to characterise the cumulated workload over a given time interval; in this framework the most popular and well-known self-similar model, widely adopted [3] for its parsimonious structure, is fractional Brownian motion (fBm).

Definition 9 (fractional Brownian motion). *A standard fractional Brownian motion $(Z_H(t), t \in \mathbb{R})$ with Hurst parameter H is characterised by the following properties:*

- *$Z_H(\cdot)$ is Gaussian, i.e., its finite-dimensional distributions are multivariate normal*
- *$Z_H(t) \in \mathcal{N}\left(0, |t|^{2H}\right)$*
- *$Z_H(\cdot)$ has stationary increments, i.e., $Z_H(u+t) - Z_H(u) \sim Z_H(t)$*
- *$Z_H(0) = 0$*
- *$Z_H(\cdot)$ has continuous sample paths*

From the above definition, it follows that Brownian motion is only a special case (for $H = 1/2$) of fBm; in that case, the analysis was much simpler since the increments were not only stationary, but also independent. Instead, for $H \neq 1/2$ the increments of $Z_H(t)$ are correlated and, if $1/2 < H < 1$, they exhibit Long Range Dependence.

Implications of LRD for Queues. Let $X\,(s,t] = X(t) - X(s)$ denote the amount of work arriving at a single server queue (with deterministic service rate C) in the time interval $\big(s,t\big]$, where $(X(t),\ t \in \mathbb{R})$ is a LRD process with drift μ (where $\mu < C$ to assure the stability of the queue) and $\mathsf{Var}X(-t,0] \ \sim\ \sigma^2 t^{2H}$.

In the large-buffer regime (section 3.2), the LDP for the queue size basically states that

$$\lim_{q\to\infty} \frac{1}{q} \log \mathbb{P}\,(Q > q) \ = \ -\delta \tag{53}$$

with the underlying assumption of the existence of a *sufficiently well-behaved* limiting cumulant generating function

$$\Lambda\,(\theta) \ = \ \lim_{t\to\infty} \frac{1}{t}\,\Lambda_t\,(\theta) \ = \ \lim_{t\to\infty} \frac{1}{t} \log \mathbb{E} e^{\theta X(-t,0]} \ .$$

If the limit exists, then the Taylor-Maclaurin expansion implies that

$$\mathsf{Var}X(-t,0] \ \sim\ t\Lambda''(0) \ . \tag{54}$$

This is not the case for LRD processes since $\mathsf{Var}X(-t,0] \ \sim \ \sigma^2 t^{2H}$. However, a variant of (53) still holds when there is some sequence $(v_t,\ t \in \mathbb{N})$ taking values in $\mathbb{R}^+$, with $v_t/\log t \to \infty$, such that the limit

$$\Lambda\,(\theta) \ = \ \lim_{t\to\infty} \frac{\Lambda_t\,(\theta v_t/t)}{v_t} \ = \ \lim_{t\to\infty} \frac{1}{v_t} \log \mathbb{E} e^{\theta X(-t,0]\, v_t/t} \tag{55}$$

exists, and is finite and differentiable in a neighbourhood of the origin. In that case, (54) becomes

$$\mathsf{Var}X(-t,0] \ \sim\ \frac{t^2}{v_t}\Lambda''(0) \tag{56}$$

and a natural choice for LRD traffic is to put $v_t \ = \ t^{2(1-H)}$. If, under this scaling, the limit $\Lambda(\theta)$, defined by (55), exists and is well-behaved, then the queue size does not decay exponentially; instead

$$\lim_{q\to\infty} \frac{1}{q^{2(1-H)}} \log \mathbb{P}\,(Q > q) \ = \ -\delta \tag{57}$$

where

$$\delta \ = \ \inf_{t>0}\ t^{2(1-H)}\,\Lambda^*\,(C + 1/t) \ . \tag{58}$$

The special case of Gaussian processes has been deeply investigated for the analytical tractability and for the relevance in traffic modelling (the *physical* motivations and all the underlying difficulties are discussed, for instance, in [10,15]) of such processes. In that framework, logarithmic as well as exact asymptotics are known, although the latter (which go beyond the LDT set-up) are much harder to obtain [16,17]. For instance, if $X(t)$ is an fBm with drift μ, variance parameter σ and Hurst parameter H, i.e.,

$$X(t) \ = \ \mu t \ + \ \sigma Z_H(t) \ ,$$

the LDP (57) can be rewritten as follows:

$$\lim_{q\to\infty} \frac{1}{q^{2(1-H)}} \log \mathbb{P}\left(Q > q\right) = -\gamma^2/2 \tag{59}$$

where

$$\gamma = \frac{(C-\mu)^H}{\sigma}\,\kappa \qquad \text{and} \qquad \kappa = \frac{1}{H^H\,(1-H)^{1-H}}\;. \tag{60}$$

Hence, as a function of q, $\mathbb{P}\left(Q > q\right)$ decays in a Weibullian way[6] i.e., roughly as $\exp\left(-q^{2-2H}\right)$ and if $H \in \left(1/2, 1\right)$, the decay is slower than exponential.

It is worth mentioning that the same asymptotic expression (called *basic approximation* in [11]) for the overflow probability can be obtained directly from the solution of the Lindley's recursion, taking into account the principle of the largest term (in the spirit of approximation (40)) and the Chernoff bound (the optimising $\hat{t}$ represents the *most likely time-scale of overflow*). The application of the generalised Schilder's theorem (that gives the sample path LDP for a general Gaussian process and hence permits to identify the most likely path to overflow) has been extended to heterogeneous traffic flows (for Gaussian processes superposition means just adding the variance functions) as well as to more complex queueing systems, such as priority queues, generalised processor sharing schedulers [15] and tandem queues [18]. Such results, which could be justified, at least in principle, invoking the contraction principle applied to the proper continuous function $f(A)$, are indeed quite accurate over the full range of buffer sizes and even for quite high traffic levels [11].

4.2 LDT and Rare Event Simulation by Means of Importance Sampling

Importance Sampling (IS) is a popular technique devised to build unbiased estimators not suffering from the smallness of the probability of interest. This is achieved by changing the law of the process so that to favour the occurrence of the target rare event and taking this change into account by reweighting the estimation according to the *likelihood ratio*, which, in measure-theoretic terms, is the Radon-Nikodym derivative of the original law with respect to the new one [9].

The efficiency of an IS-based algorithm depends on the choice of a "proper" *change of measure* to reduce the variance of the estimate. It is well known that the optimal change of measure (*zero-variance* pdf) involves the knowledge of the probability we want to estimate and therefore cannot be practically adopted. The issue is commonly tackled by restricting potential IS measures to a parametric class and determining the optimal change of measure within this restricted class[7]. The most common approach is represented by the use of a so-called exponential change of measure (ECM), already introduced in section 2.1 as a technique to prove the lower bound in Cramér's theorem.

Roughly speaking, LDT states that a target rare set is most likely to be reached by following the path $\hat{f}$ that minimises the corresponding rate function. Thus, simulating

[6] The exact asymptotics of $\mathbb{P}\left(Q > q\right)$, in which a crucial role is played by the so-called Pickands constant, factorise into the *same* Weibullian term and a hyperbolic prefunction [10].

[7] Since the topic requires by itself a complete tutorial, in the following only the basic ideas are sketched; see [9] for all the relevant definitions.

the system under the change of measure that favours that path is the quickest way to reach the rare set. For random walks (and hence for the G/G/1 framework), the previous heuristic idea is formally justified by the following theorem:

Theorem 9 (Siegmund, Lehtonen/Nyrhinen). *An IS estimator for the probability that a random walk with negative drift exceeds some positive level x is* asymptotically optimal*, iff it is built according to the ECM, where the twisting parameter $\theta^* > 0$ is chosen such that $\Lambda(\theta^*) = 0$, where $\Lambda(\cdot)$ denotes the cumulant generating function of the increments.*

In conclusion, the most likely way in which the random walk can cross level x is by moving linearly at rate $\Lambda'(\theta^*)$, which is exactly the new drift associated to the ECM (14). The previous theorem has a very nice interpretation for an M/M/1 queue: under the optimal ECM, the arrival and service rates are simply twisted. Unfortunately a similar result cannot be extended to Jackson queueing networks and, in general, state-dependent heuristics are required [19].

Finally, it is worth noticing that, when the input traffic is fBm, a change of measure based on the most likely path is *not asymptotically efficient* [20] even for the single server queue. An intuitive explanation is that the main contribution to the asymptotics of the second order moment of the IS estimator is determined by paths which give a very small contribution to the overflow probability, but for which the likelihood ratio is very large. Asymptotic optimality can be achieved by the use of more refined IS techniques [21,22], but at the cost of a higher computational complexity. An alternative approach, which retains the simplicity of ECM-based IS with a lower variance of the estimates (although the algorithm is not asymptotically efficient), is the so-called Bridge Monte-Carlo method, based on the idea of expressing the overflow probability in terms of the *bridge* of the input process [23].

5 Conclusions

The theory of large deviations is a powerful tool for the analysis and simulation of rare events. For lack of space and for its specific target, this tutorial could only introduce some basic principles and show how the general ideas may be used in the framework of computer networks. The interested reader (see also [4] for a more detailed review of the literature) can find in [5] a general introduction to the theory and in [7] a condensed and rigorous overview of the main results. A good compromise between general theory (in the first part of the book) and applications (to performance evaluation in communication and computer architectures) is represented by [12], while [6], the key reference for this tutorial, focuses on queueing systems. The book, starting from the elementary case of IID arrivals, shows how abstract LDT theorems permit to extend the results to very general scenarios and finally deals with more tangible concepts, such as effective bandwidths, scaling properties (which are used, for instance in [24], to analyse the effect of TCP on network stability) and hurstiness, an LDP-oriented characterisation of Long Range Dependence. Finally, [10] is not a book on large deviations, but the analysis of Gaussian queues represents a natural framework to derive and heuristically justify some of the main LDT results.

References

1. Xiao, X., Ni, L.M.: Internet QoS: A Big Picture. IEEE Network 13(2), 8–18 (1999)
2. Leland, W.E., Taqqu, M.S., Willinger, W., Wilson, D.V.: On the self-similar nature of Ethernet traffic (extended version). IEEE/ACM Trans. Netw. 2(1), 1–15 (1994)
3. Norros, I.: On the use of fractional Brownian motion in the theory of connectionless networks. IEEE Journal of Selected Areas in Communications 13(6), 953–962 (1995)
4. Weiss, A.: An introduction to large deviations for communication networks. IEEE Journal on Selected Areas in Communications 13(6), 938–952 (1995)
5. Dembo, A., Zeitouni, O.: Large deviations techniques and applications, 2nd edn. Applications of Mathematics, vol. 38. Springer, Heidelberg (1998)
6. Ganesh, A., O'Connell, N., Wischik, D.: Big Queues. Lecture Notes in Mathematics. Springer, Heidelberg (2004)
7. Varadhan, S.R.S.: Large Deviations and Applications. SIAM, Philadelphia (1984)
8. Bucklew, J.A.: Large Deviation Techniques in Decision, Simulation and Estimation. Wiley, Chichester (1990)
9. Heidelberger, P.: Fast simulation of rare events in queueing and reliability models. ACM Trans. Model. Comput. Simul. 5(1), 43–85 (1995)
10. Mandjes, M.: Large deviations for Gaussian queues. Wiley, Chichester (2007)
11. Addie, R., Mannersalo, P., Norros, I.: Most probable paths and performance formulae for buffers with Gaussian input traffic. European Transactions on Telecommunications 13(3), 183–196 (2002)
12. Shwartz, A., Weiss, A.: Large Deviations for Performance Analysis. Chapman & Hall, Boca Raton (1995)
13. Zwart, A.P.: Queueing Systems with Heavy Tails. PhD thesis, Eindhoven University of Technology (2001)
14. Beran, J.: Statistics for Long-Memory Processes. Chapman & Hall/CRC (1994)
15. Mannersalo, P., Norros, I.: A most probable path approach to queueing systems with general Gaussian input. Computer Networks 40(3), 399–411 (2002)
16. Narayan, O.: Exact asymptotic queue length distribution for fractional Brownian traffic. Advances in Performance Analysis 1(1), 39–63 (1998)
17. Dȩbicki, K.: A note on LDP for supremum of Gaussian processes over infinite horizon. Statistics & Probability Letters 44(3), 211–220 (1999)
18. Mandjes, M., Mannersalo, P., Norros, I.: Gaussian tandem queues with an application to dimensioning of switch fabrics. Computer Networks 51(3), 781–797 (2007)
19. Zaburnenko, T.S.: Efficient heuristics for simulating rare events in queuing networks. PhD thesis, University of Twente (2008)
20. Baldi, P., Pacchiarotti, B.: Importance Sampling for the Ruin Problem for General Gaussian Processes. Technical report, Universités de Paris 6 & Paris 7 (2004)
21. Dieker, A.B., Mandjes, M.R.H.: Fast simulation of overflow probabilities in a queue with Gaussian input. ACM Transactions on Modeling and Computer Simulation 16(2), 119–151 (2006)
22. Dupuis, P., Wang, H.: Importance Sampling, Large Deviations and differential games. Technical report, Lefschetz Center for Dynamical Systems, Brown University (2002)
23. Giordano, S., Gubinelli, M., Pagano, M.: Bridge Monte-Carlo: a novel approach to rare events of Gaussian processes. In: Proc. of the 5th St.Petersburg Workshop on Simulation, St. Petersburg, Russia, pp. 281–286 (2005)
24. Raina, G., Wischik, D.: Buffer sizes for large multiplexers: TCP queueing theory and instability analysis. In: EuroNGI Conference on Next Generation Internet Networks, Rome (April 2005)