An Introduction to Large Deviations for Teletraffic Engineers

J.T. Lewis (DIAS) Neil O'Connell (TCD)
Raymond Russell (DIAS)

October 6, 1995

Introduction

What you will learn from this tutorial:

- What Large Deviation Theory is about
- Coin tossing: how to explore Large Deviations using your PC
- Cramér's Theorem and the Rate-Function
- The connection with the Central Limit Theorem (why "Large" Deviations?)
- How to calculate the rate function: bypassing combinatorics with Chernoff's formula
- The connection with Shannon entropy
- How to deal with more general cases: Varadhan's Theorem and the scaled CGF
- New rate-functions from old: the Contraction Principle
- Large Deviations in Queuing Networks: effective bandwidths
- Bypassing modelling: estimating the scaled CGF

What Large Deviation Theory is about

Large Deviation theory is a theory of rare events; it can be used to estimate the tails of probability distributions. In recent years, it has been used extensively in teletraffic theory. The aim of this tutorial is to introduce engineers to the basic ideas of the theory.

Coin tossing: how to explore Large Deviations using your PC

If you have a little skill in programming, you can very quickly get a good feel for the basic ideas of Large Deviation theory by carrying out on your PC the experiments we are going to describe. Even if you can't program – and you can't rope anyone into programming for you, you will find it useful to read this section: consider what we are going to describe as a thought-experiment, and we will supply the results.

Imagine a coin-tossing experiment, where we toss a coin n times and record each result. There are 2 possible outcomes for each toss, giving 2^n possible outcomes in all. What can we say about the total number of heads? Well, there are $n+1$ possible values for the total, ranging from 0 heads to n heads; of the 2^n possible outcomes,

$$\left(\begin{array}{c} n \\ r \end{array}\right)$$

result in r heads. If the coin is fair, every outcome is equally likely, and so the probability of getting r heads is

$$\left(\begin{array}{c} n \\ r \end{array}\right) \frac{1}{2^n}.$$

Thus the distribution of heads is made up of $n+1$ atoms with weights given by the combinatorial factors; to calculate the probability of the average number of heads per toss lying in a particular range, we add up the weight of each of the atoms which fall inside that range. If we let M_n be the average number of heads in n tosses, then

$$\mathbb{P}[x < M_n < y] = \sum_{\{r : x < \frac{r}{n} < y\}} \left(\begin{array}{c} n \\ r \end{array}\right) \frac{1}{2^n}.$$

Exercise: Write a function/procedure to take an integer n and two floating-point numbers x and y and return the value of the expression above. Use this function/procedure to write a program to produce histograms of the distribution of M_n for selected values of n.

We have done this for n=16, n=32, n=64 and n=128; we can see clearly the Law of Large Numbers at work: as n increases, the distribution becomes more and more sharply peaked about the mean, 1/2, and the tails become smaller and smaller.

Exercise: Pick some point x greater than 1/2 and write a program to calculate, for a range of values of n, the logarithm of the probability of M_n exceeding x.

We have chosen x=0.6 and produced a plot of $\log \mathbb{P}[M_n > x]$ against n for n up to 1000. It is clear that, although things are a little jumpy initially, the plot becomes linear for large n. Repeat the experiment for a different value of x and you will see that the same thing happens: no matter what value of x greater than 1/2 you take, the plot will always be linear for n large. How quickly it becomes linear, and what the asymptotic slope is, depends on the value of x, but the graph of $\log \mathbb{P}[M_n > x]$ against n is always linear for large n. Let's call this asymptotic slope $-I(x)$.

Exercise: Repeat the experiment for a range of values of x from 1/2 to 1, measure the asymptotic slope in each case, and plot the values of $I(x)$ you get against x. Do the same thing for $\log \mathbb{P}[M_n < x]$ for a range of values of x from 0 to 1/2.

You have made a discovery:

THE TAIL OF THE DISTRIBUTION OF THE
AVERAGE NUMBER OF HEADS IN n TOSSES
DECAYS EXPONENTIALLY AS n INCREASES

The plot you have made tells you the local rate at which a tail decays as a function of the point from which the tail starts: you have built up a picture of the *rate-function* $I(x)$.

Exercise: Plot the graph of the function $x \ln x + (1-x)\ln(1-x) + \ln 2$ against x and compare it with your previous plot.

We see that the two plots fit: we have guessed a formula for $I(x)$, the rate-function for coin-tossing. One of the goals of Large Deviation theory is to have a systematic way of calculating the rate-function; we will show you later one way of achieving this.

To summarise: we have found that, for coin tossing, the tails of the distribution of M_n, the average number of heads in n tosses, decay exponentially fast

$$\mathbb{P}[M_n > x] \approx e^{-nI(x)}, x > 1/2,$$

$$\mathbb{P}[M_n < x] \approx e^{-nI(x)}, x < 1/2,$$

as n becomes large; in fact, the approximation is quite good for surprisingly small values of n.

Cramér's Theorem and the Rate Function

Harald Cramér was a Swedish mathematician who served as a consultant actuary for an insurance company; this led him to discover the first result in Large Deviation theory. The Central Limit Theorem gives information about the behaviour of a probability distribution near its mean while the risk theory of insurance is concerned with rare events out on the tail of a probability distribution. Cramér was looking for a refinement of the Central Limit Theorem. What he proved was this:

Cramér's Theorem *Let $X_1, X_2, X_3, \ldots$ be a sequence of bounded, independent and identically distributed random variables each with mean m, and let*

$$M_n = \frac{1}{n}(X_1 + \ldots + X_n);$$

denote the empirical mean; then the tails of the probability distribution of M_n decay exponentially with increasing n at a rate given by the rate-function $I(x)$:

$$P[M_n > x] \approx e^{-nI(x)}, x > m,$$
$$P[M_n < x] \approx e^{-nI(x)}, x < m.$$

Historically, Cramér used complex variable methods to prove his theorem. He gave $I(x)$ as a power-series; how this came about is the subject of the next section.

The connection with the Central Limit Theorem

Recall what the Central Limit Theorem tells us: if $X_1, X_2, X_3, \ldots$ is a sequence of independent and identically distributed random variables with mean μ and variance $\sigma^2 < \infty$, then the average of the first n of them, $M_n = \frac{1}{n}(X_1 + \ldots + X_n)$ is approximately normal with mean μ and variance σ^2/n. That is, its probability density function is

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

and the approximation is only valid for x within about $\sigma/\sqrt{n}$ of μ. If we ignore the prefactor in f and compare the exponential term with the approximation that Cramér's Theorem gives us, we see that the terms $(x-\mu)^2/2\sigma^2$

occupy a position analogous to that of the rate function. Let us look again at the coin tossing experiments: for x close to $1/2$, we can expand our rate-function in a Taylor series:

$$x \log x + (1-x)\log(1-x) + \log 2 = \frac{(x-\frac{1}{2})^2}{2 \times \frac{1}{4}} + \ldots.$$

The mean of each toss of a coin is $1/2$, and the variance of each toss is $1/4$; thus the rate-function for coin tossing gives us the Central Limit Theorem. In general, whenever the rate-function can be approximated near its maximum by a quadratic form, the Central Limit Theorem holds.

So much for the similarities between the CLT and Large Deviations; the name "Large Deviations" arises from the contrast between them. The CLT governs random fluctuations only near the mean – deviations from the mean of the order of $\sigma/\sqrt{n}$. Fluctuations which are of the order of σ are, relative to typical fluctuations, much bigger: they are *large deviations* from the mean. They happen only rarely, and so Large Deviation theory is often described as *the theory of rare events* – events which take place away from the mean, out in the tails of the distribution; thus Large Deviation theory can be described alternatively as the theory which studies the tails of distributions.

How to calculate the rate function: bypassing combinatorics with Chernoff's formula

One way of calculating the rate-function for coin-tossing is to apply Stirling's formula in conjunction with the combinatorial arguments we used earlier. Chernoff's formula gives the rate-function in terms of the *cumulant generating function* λ:

$$I(x) = \min_{t \in \mathbb{R}} \{xt - \lambda(t)\},$$

where λ is defined by

$$\lambda(t) := \log \mathbb{E} e^{tX_j}.$$

The cumulants of a distribution are closely related to the moments. The first cumulant is simply the mean, the first moment. The second cumulant is the variance, the second moment less the square of the first moment. The relationship between the higher cumulants and the moments is more complicated, but in general the k^{th} cumulant can be written in terms of the first k moments. The relationship between the moments and the cumulants is more clearly seen from their respective generating functions. The function $\phi(t) = \mathbb{E}e^{tX_1}$ is the moment generating function for the X's: the k^{th} moment of the X's is the k^{th} derivative of ϕ evaluated at $t = 0$:

$$\begin{aligned} \frac{d^k}{dt^k}\phi(t) &= \mathbb{E}[X_1^k e^{tX_1}] \\ \left.\frac{d^k \phi}{dt^k}\right|_{t=0} &= \mathbb{E}[X_1^k] = k^{\text{th}} \text{ moment} \end{aligned}$$

The cumulant generating function is defined to be the logarithm of the moment generating function, $\lambda(t) := \log \phi(t)$, and the cumulants are then just the derivatives of λ:

$$\begin{aligned} \left.\frac{d}{dt}\lambda(t)\right|_{t=0} &= m, \\ \left.\frac{d^2}{dt^2}\lambda(t)\right|_{t=0} &= \sigma^2, \ldots. \end{aligned}$$

So, what is the idea behind Chernoff's formula? Well, in order to calculate the Central Limit Theorem approximation for the distribution for M_n, we must calculate the mean and variance of the X's: essentially we use the first two cumulants to get the first two terms in a Taylor expansion of the the rate-function to give us a quadratic approximation. It is easy to see that, if we want to get the full functional form of the rate function, we must use all the terms in a Taylor series, that is, we must use all the cumulants. The cumulant generating function packages all the cumulants together, and Chernoff's formula shows us how to extract the rate-function from it.

The connection with Shannon entropy

The context in which the word 'entropy' appears most often outside the physical sciences is Information Theory. Many people have heard of Shannon entropy and are familiar with the formula $-\sum_i p_i \log p_i$. Although it would be very interesting to give an exposition here of the ideas involved in Information Theory, that would sidetrack us somewhat. However, we shall at least apply Cramér's Theorem to derive one of the basic results used in Information Theory and show how Shannon information is related to a Large Deviation rate-function.

Suppose we draw n letters at random from a finite alphabet $A = \{a_1, \ldots a_r\}$. Let us call the word we form in this way ω, and let $\nu_{\mathbf{n}}(\omega)$ be the vector whose r components are the relative frequencies with which each of the letters appear:

$$\nu_n(\omega) := \left(\frac{n_1(\omega)}{n}, \ldots \frac{n_r(\omega)}{n} \right),$$

where $n_j(\omega)$ is the number of times the letter a_j appears in the word ω. The vector ν_n takes values in the space X of probability vectors $\mathbf{m} = (m_1, \ldots m_r)$, where $m_j \geq 0$ and $m_1 + \ldots + m_r = 1$. One such probability vector is the vector $\mathbf{p}$ whose j^{th} component p_j is the true probability of drawing the letter a_j. Since the letters are drawn independently of one another, the probability of getting the word ω is

$$\mathbb{P}^p[\omega] = p_1^{n_1(\omega)} \ldots p_r^{n_r(\omega)}.$$

Sanov's Theorem states that

$$\mathbb{P}^p[\nu_n(\omega) \text{ is close to } \mathbf{m}] \approx e^{-nH(\mathbf{m}|\mathbf{p})},$$

where the rate-function $H(\mathbf{m}|\mathbf{p})$ is given by

$$H(\mathbf{m}|\mathbf{p}) = m_1 \log \frac{m_1}{p_1} + \ldots + m_r \log \frac{m_r}{p_r}.$$

The fact that the letters are drawn independently allows us to apply Cramér's Theorem to prove the existence of the rate-function, but we must use Chernoff's formula to calculate the functional form of $H(\mathbf{m}|\mathbf{p})$. Since H is a function of a vector, the cumulant generating function is a function of a vector $\mathbf{t}$:

$$\begin{aligned}
\lambda(\mathbf{t}) &= \ln \mathbb{E} e^{(t_1 \frac{n_1}{n} + \ldots + t_r \frac{n_r}{n})}. \\
&= \log \left(p_1 e^{t_1} + \ldots + p_r e^{t_r} \right)
\end{aligned}$$

To compute $H(\mathbf{m}|\mathbf{p})$, we must calculate the Legendre transform of λ:

$$H(\mathbf{m}|\mathbf{p}) = \max_{\mathbf{t}}\{\, t_1 m_1 + \ldots + t_r m_r - \lambda(\mathbf{t})\,\}.$$

Exercise: Use differential calculus to show that $\mathbf{t}$ must satisfy

$$\frac{\partial\lambda}{\partial t_k} = \frac{p_k e^{t_k}}{p_1 e^{t_1} + \ldots + p_r e^{t_r}} = m_k,$$

and so

$$t_k = \log\frac{m_k}{p_k} + \lambda(\mathbf{t}).$$

Substitute this into the expression to be minimised to show that

$$H(\mathbf{m}|\mathbf{p}) = m_1 \log\frac{m_1}{p_1} + \ldots + m_r \log\frac{m_r}{p_r}.$$

Going back to the statement of Cramér's Theorem, we see that the distribution of M_n is concentrated near m, the place where the rate-function vanishes.

Execise: Show that it follows from Cramér's Theorem that, as n increases,

$$\lim_n \mathbb{P}[m - \delta < M_n < m + \delta] = 1,$$

for any $\delta > 0$.

In Sanov's Theorem, the rate-function is $H(\mathbf{m}|\mathbf{p})$; this vanishes if and only if $\mathbf{m} = \mathbf{p}$. It follows that

$$\lim_n \mathbb{P}^p[\nu_n(\omega) \text{ is close to } \mathbf{p}] = 1.$$

Now define Γ_n to be the set of words of length n for which $\nu_n(\omega)$ is close to $\mathbf{p}$. Then

$$\mathbb{P}^p[\Gamma_n] \approx 1.$$

The set Γ_n consists of the most probable words. We may decide that these are the only ones we need to code; this decision can yield a worth-while saving in effort if the probability vector $\mathbf{p}$ is not the uniform vector $\mathbf{u} = (\frac{1}{r}, \ldots, \frac{1}{r})$. Notice that $\mathbb{P}^u$ is just normalised counting measure:

$$\mathbb{P}^u[A] = |A|/|\Omega_n|$$

. In particular, $\mathbb{P}^u[\Gamma_n] = |\Gamma_n|/|\Omega_n|$. We can use Sanov's Theorem applied to $\mathbb{P}^u$ to estimate the size of Γ_n:

$$\mathbb{P}^u[\Gamma_n] = \mathbb{P}^u[\nu_n(\omega) \text{ is close to } \mathbf{p}] \approx e^{-nH(\mathbf{p}|\mathbf{u})}.$$

Exercise: Use Sanov's Theorem to show that

$$|\Gamma_n| \approx e^{nh(\mathbf{p})},$$

where $h(\mathbf{p}) := -\Sigma_j p_j \ln p_j$ is the Shannon entropy of $\mathbf{p}$.

Since $|\Omega_n| = e^{n \ln r}$ and $h(\mathbf{u}) = \ln r > h(\mathbf{p})$ if $\mathbf{p} \neq \mathbf{u}$, it follows that Γ_n is substantially smaller than Ω_n when $\mathbf{p}$ is not uniform.

How to deal with more general cases than coin tossing

So far we have only talked about I.I.D. processes. What about more general processes, such as Markov chains? Well, just as there is an analogue for non-independent random variables of the CLT, giving a normal approximation to the the distribution of M_n, so there is an analogue of Cramér's Theorem, giving the Large Deviation estimates for the distribution of M_n.

Again, suppose that $X_1, X_2, X_3, \ldots$ is a sequence of real-valued random variables which is *mixing* (a very descriptive term – it means that X's which are widely separated are approximately independent of each other: no matter what the value of X_1 is, its influence on the value of X_n is negligible since it is well "mixed" in with the randomness of all the X's in between). Let $M_n = \frac{1}{n}(X_1 + \ldots + X_n)$ be the average of the first n of them; then, again, we have a rate function I which describes the dominant behaviour of the distribution of M_n:

$$\mathbb{P}[M_n \in C] \approx e^{-n \min_{x \in C} I(x)}.$$

How do we calculate the rate-function here? Varadhan's Theorem gives us the answer:

$$I(x) = \min_{t \in \mathbb{R}} \{xt - \lambda(t)\}$$

where λ in now the *scaled cumulant generating function* (SCGF) defined by

$$\lambda(t) := \lim_{n \to \infty} \frac{1}{n} \ln \mathbb{E} e^{ntM_n}.$$

The proof is somewhat technical, but the idea behind it is very simple. Write the expectation $\mathbb{E}$ in the definition of λ as an integral with respect to the distribution of M_n:

$$\lambda(t) = \lim_{n \to \infty} \frac{1}{n} \log \int_{\mathbb{R}} e^{ntx} d\mathbb{P}[M_n = x];$$

since $\mathbb{P}[M_n > x]$ behaves like $e^{-nI(x)}$, we can say

$$\begin{aligned} \lambda(t) &= \lim_{n \to \infty} \frac{1}{n} \log \int_{\mathbb{R}} e^{ntx} e^{-nI(x)} dx \\ &= \lim_{n \to \infty} \frac{1}{n} \log \int_{\mathbb{R}} e^{n(tx - I(x))} dx. \end{aligned}$$

For n large, the integral is dominated more and more strongly by the maximum value of the integrand, which is $e^{n \max_x (tx - I(x))}$, and so we expect that

$$\begin{aligned} \lambda(t) &= \lim_{n\to\infty} \frac{1}{n} \log e^{n \max_x (tx - I(x))} \\ &= \max_x (tx - I(x)). \end{aligned}$$

The last quantity is known as the Legendre transform of I. The Legendre transform is like the Fourier transform in that, for an appropriate class of functions, the transformed function contains exactly the same information as the original function and so the transform is invertible . The Legendre transform is invertible on the class of convex functions and is inverted by repeating it; thus if I is convex, then its double transform I^{**} is just I itself.

Thus the scaled cumulant generating function λ is the Legendre transform of the rate-function and, if the latter is convex, it is the Legendre transform of the SCGF. The convexity of the rate-function is not that unusual a condition; very often, the argument that establishes the existence of the rate-function also tells us that it is convex.

Note that Chernoff's formula is a special case of Varadhan's theorem: if the X's are independent, then

$$\mathbb{E}e^{ntM_n} = \mathbb{E}e^{t(X_1 + \ldots + X_n)} = \left(\mathbb{E}e^{tX_1}\right)^n,$$

the last step because the X's are independent and identically distributed. Thus

$$\begin{aligned} \lambda(t) &= \lim_{n\to\infty} \frac{1}{n} \log \left(\mathbb{E}e^{tX_1}\right)^n \\ &= \lim_{n\to\infty} \frac{n}{n} \log \mathbb{E}e^{tX_1} \\ &= \log \mathbb{E}e^{tX_1}; \end{aligned}$$

in this case, the scaled cumulant generating function is the same as the cumulant generating function and Varadhan's theorem yields Chernoff's formula.

New rate-functions from old: the Contraction Principle

In many applications of probability theory , we model a process with a sequence $\{Y_n\}$ of random variables and we are able to show that some property of the process is described by a related process $\{f(Y_n)\}$. Does the Large Deviation behaviour of the first process tell us anything about the Large Deviation behaviour of the second process? The answer is given by the
Contraction principle: *If $\{Y_n\}$ has a rate-function I and f is continuous, then $\{f(Y_n)\}$ has a rate-function J and J is given by*

$$J(z) = \min\{I(y) : f(y) = z.\}$$

To see the Contraction Principle in action, let us return to the experiment in which n letters are drawn at random from a finite alphabet. We saw that the relative frequency vector ν_n satisfies a Large Deviation Principle in that

$$\mathbb{P}[\nu_{\mathbf{n}}(\omega) \text{ is close to } \mathbf{m}] \approx e^{-nH(\mathbf{m}|\mathbf{p})}.$$

Suppose we take the letters $a_1, \ldots a_r$ to be real numbers and that, instead of investigating the distribution of the relative frequency vector, we decide to investigate the distribution of the mean $M_n(\omega) = a_1\frac{n_1(\omega)}{n} + \ldots + a_r\frac{n_r(\omega)}{n}$. Do we have to go and work out the Large Deviation Principle for M_n from scratch? No, because M_n is a function of the relative frequency vector ν_n. It is a very simple function – just the inner product $f(\mathbf{m}) =< \mathbf{a}, \mathbf{m} >$, where $\mathbf{a}$ is the vector whose components are the letters $a_1, \ldots a_r$. It is obviously continuous; hence the contraction principle applies, allowing us to calculate the rate-function $I(x)$ for M_n in terms of the rate-function $H(\mathbf{m}|\mathbf{p})$ for ν_n. We have that

$$I(x) = \min_{\mathbf{m}} H(\mathbf{m}|\mathbf{p}) \textit{ subject to } \quad < \mathbf{a}, \mathbf{m} >= x.$$

This is a simple optimisation problem with one constraint: we can solve it using a Lagrange multiplier.
Exercise: Show that the value of $\mathbf{m}$ which achieves the minimum is given by

$$m_k = \frac{e^{\beta a_k} p_k}{e^{\beta a_1} p_1 + \ldots + e^{\beta a_r} p_r},$$

where β is the Lagrange multiplier whose value can be determined from the constraint.

Large Deviations in Queuing Networks: effective bandwidths

Consider a single-server queue Q with constant service rate s fed by a stream of arrivals which, at time t, brings an amount of work X_t to be served. For each t, the arrivals process A_t is defined to be the total amount of work which has arrived since time $-t$ in the past. It follows from a standard result in queuing theory that the current queue-length Q is determined by the arrivals process:

$$Q = \max_t \{A_t - st\}.$$

An argument based on the Contraction Principle shows that if the arrivals process satisfies a Large Deviation principle with rate-function I, so that

$$\mathbb{P}[A_t/t > x] \approx e^{-tI(x)},$$

then the tail of the queue-length distribution satisfies, for large q,

$$\mathbb{P}[Q > q] \approx e^{-q\delta},$$

where δ is determined by I. Of course, δ can also be calculated from the scaled cumulant generating function λ of the arrivals process:

$$\delta = \max\{\, \theta : \lambda(\theta) \leq s\theta \,\}, \qquad \lambda(\theta) := \lim_{t\to\infty} \frac{1}{t} \log \mathbb{E} e^{\theta A_t}.$$

If the queue has only a finite waiting space, then δ gives us an estimate of what that buffer-size must be in order to achieve a given probability of overflow. If we know what λ is, we can calculate δ for each value of the service rate, and so we can Estimate what size buffer is needed. Alternatively, we can turn the equation $\lambda(\delta) = s\delta$ around to answer the question: if we have a fixed buffer-size, what service rate is needed to make the probability of overflow acceptably small? We can specify δ and calculate the *effective bandwidth* s of the traffic from

$$s = \lambda(\delta)/\delta$$

.

Bypassing modelling: estimating the scaled CGF

All we need for the SCGF (scaled cumulant generating function) of the arrivals to exist is for the arrivals to be stationary and mixing. If these two conditions are satisfied, we can use the SCGF to make predictions about the behaviour of the queue.

One way to get the SCGF for the arrivals is to make a suitable statistical model, and then calculate it using techniques from Large Deviation theory. There are a number of problems with this approach. Firstly, real traffic streams cannot be accurately represented by simple models; any realistic model would have to be quite complex, with many parameters to be fitted to the data. Secondly, the calculation of the SCGF for any but the simplest model is a difficult problem. Thirdly, even if you could find a realistic model, fit it to your data and calculate the SCGF, this would be a wasteful exercise: the SCGF is a Large Deviation object, and it does not depend on the details of the model, only on its "bulk properties". Hence all the effort you put into fitting your sophisticated model to the data is, to a large extent, lost.

Our approach is to ask "Why not measure what you are looking for directly?" There are many good precedents for this approach. When engineers design a steam turbine they need to know the thermodynamic properties of steam. To find this out, they do not make a sophisticated statistical mechanical model of water and calculate the entropy from that; instead, they measure the entropy directly in a calorimetric experiment, or (more likely) they use steam tables – based on somebody else's measurements of the entropy. Now entropy is nothing but a rate-function, so how do we measure the rate-function – or, equivalently, the SCGF – of an arrivals stream? Well, assuming that the stream is mixing, we can approximate the SCGF by a finite-time cumulant generating function:

$$\lambda(\theta) \approx \lambda_T(\theta) = \frac{1}{T} \log \mathbb{E} e^{\theta A_T},$$

for T sufficiently large. We can now estimate the value of the expectation by breaking our data into blocks of length T and averaging over them:

$$\hat{\lambda}(\theta) := \frac{1}{T} \log \frac{1}{K} \sum_{k=1}^{k=K} e^{\theta \tilde{X}_k},$$

where the $\tilde{X}$'s are the block sums

$$\tilde{X}_1 = X_1 + \ldots + X_T, \quad \tilde{X}_1 = X_{T+1} + \ldots + X_{2T}, \quad \text{etc.}$$

This simple estimator has been used in preliminary investigations; it is likely that more sophisticated estimators will make buffer-dimensioning "on the fly" a practicable proposition.