Large deviation theory and extreme waves

Oliver Bühler
Center for Atmosphere Ocean Science, Courant Institute of Mathematical Sciences
New York University, New York, NY, USA.
June 20, 2007

Abstract.
The mathematical tools of large deviation theory for rare events are illustrated with some simple examples. These include discrete and continuous Gaussian processes, importance sampling, and evolution equations of the Langevin type. Some of these methods have been used in the study of rogue surface waves but it seems that large deviation theory could have much wider application in geophysical problems.

Introduction

Our knowledge of many small-scale processes in atmosphere and ocean dynamics is necessarily statistical in nature. For instance, this obviously applies to ocean surface and internal waves, which in large part must be treated as a random field whose statistics are described (or at least constrained) by observed spectra. When such processes show extreme behaviour (e.g., rogue waves, or large-amplitude internal waves that may lead to wave breaking) then a natural question to ask is whether 'new physics' is involved, or whether the extreme event is just that, i.e., an extreme form of the same physical dynamics that also governs the non-extreme, typical events. Statistical methods using observed 'normal' spectra can be used to study the second kind of extreme event but not the first kind, and this offers a test for detecting the presence of new physics.

In this context an important general tool from applied probability is large deviation theory (LDT). This theory is well-developed in the mathematical community (e.g., *Freidlin and Wentzell* [1998], *Varadhan* [1994]) and it has being applied successfully to diverse problems such as communication network behaviour, chemical reactions, conformation changes between meta-stable states, and phase transitions in general. There are two basic facts that make LDT relevant to these applications: first, it turns out that events with very little likelihood, when they occur, do so in the overwhelming majority of cases by following the path that is *least* unlikely. In other words, the probability distribution conditional on the occurrence of a rare event is tightly localized around the most likely way in which the rare event can be realized. In this sense, the shape of rare events of a stochastic process becomes nearly deterministic.

Second, in problems where multiple time scales are involved a rare event on one time scale is not rare when viewed from another timescale. For example, a chemical reaction (or the folding of a protein) are rare when compared to the natural time scale of molecular oscillations, but they are not rare when viewed from a macroscopic time scale. This second point obviously applies to geophysical systems in which fast small-scale processes (such as internal wave breaking and the concomitant ocean mixing) are known to contribute significantly to the slow large-scale evolution of the global system.

This indicates that LDT might be useful in geophysical problems involving extreme events. In fact, some aspects of LDT have been rediscovered independently in the surface oceanography community (e.g., *Boccotti* [1989], *Phillips et al.* [1993]), but it appears that LDT itself is not well known. Here the basic tools of LDT are illustrated with some simple examples. Perhaps the most important feature of LDT is that it is disarmingly easy to use in practice, because it converts the problem of finding the most likely shape of a rare event into a problem of constrained minimization of a certain action functional. This allows the use of calculus of variations and numerical optimization on these problems and it also allows a very flexible definition of the rare event itself, which can be a lot more complex than simply exceeding an amplitude threshold at some point, say.

Discrete Gaussian random variables

This is the simplest context in which to discuss large deviations because here an exact solution for simple large-amplitude events exists and can be compared with LDT. We consider a discrete random process that is a sequence of real-valued random variables X_i with $i \in Z$. For instance, the X_i could represent measure-

ments of sea surface height at a fixed position and at times $t = i\Delta t$ where Δt is the time resolution of the instrument. We assume that the first two moments of X_i are

$$\mathbb{E}[X_i] = 0 \quad \text{and} \quad \mathbb{E}[X_i X_j] = C_{ij} \tag{1}$$

where $\mathbb{E}[]$ denotes statistical expectation. Thus the variables have zero mean and covariances given by the positive definite covariance matrix $C_{ij} = C_{ji}$. In the special case of a stationary process $C_{ij} = C_{|i-j|}$. We will always assume that C_{ij} goes to zero as $|i-j|$ goes to infinity. The information in (1) is enough to define a normal distribution for the X_i as follows: if we select any N variables out of the X_i then the N-point probability distribution is given by the multivariate normal density

$$p(x_1, \ldots, x_N) = \frac{1}{Z} \exp\left(-\frac{1}{2}\sum_{i,j}^{N} x_i C_{ij}^{-1} x_j\right) \tag{2}$$

where $Z = (2\pi)^{N/2}\sqrt{\det C_{ij}}$. (Note that this definition of a Gaussian process far exceeds the requirement that X_i should be normally distributed for a single value i.) This holds[1] for any finite N and completely specifies the statistics of the process. For instance, the probability to find all of the selected X_i in some arbitrary intervals B_i can now be computed as

$$\mathbb{P}[X_1 \in B_1, \ldots, X_N \in B_N] = \int_{x_1 \in B_1} \ldots \int_{x_N \in B_N} p\, dx_1 \ldots dx_N. \tag{3}$$

The key component of the Gaussian density in (2) is the positive definite[2] quadratic form in the exponent. By inspection, it suggests that "likely configurations" of the random process will be associated with smaller values of this quadratic form. More precisely, we can define coarse-grained configurations by picking N real numbers ϕ_i and setting $B_i = [\phi_i - \delta, \phi_i + \delta]$ with small bin-size δ. This yields

$$\mathbb{P}\left[\max_i |X_i - \phi_i| \le \delta\right] \approx \frac{(2\delta)^N}{Z} \exp\left(-\frac{1}{2}\sum_{i,j}^{N} \phi_i C_{ij}^{-1} \phi_j\right) \tag{4}$$

as $\delta \to 0$, which exhibits the role of the quadratic form. Clearly, the most likely configuration is $\phi_i = 0$ for all i, which corresponds to a flat ocean surface at all times. This shows vividly that the most likely configuration (which for a Gaussian distribution always coincides with the mean configuration) need not dominate the statistics, because the aggregate probability of all other possible configurations may far exceed (4). This changes once we look at conditional probabilities for extreme waves, where the most likely configuration can indeed dominate the statistics.

Conditional distribution

We now make the assumption

$$A: \quad X_0 = a \tag{5}$$

for some positive amplitude $a > 0$ and then consider the conditional distribution of the process X_i under this assumption. The motivation is that we want to investigate the process under the assumption that an extreme amplitude a has been recorded at $t = 0$. This is particularly easy for a Gaussian process because the conditional distribution is again Gaussian with conditional means and covariances that are given by the standard linear regression formulas, which are exact for Gaussian variables. This yields

$$\mathbb{E}[X_i|A] = a\frac{C_{i0}}{C_{00}} \quad \text{and} \quad \mathbb{E}\left[X_i' X_j'|A\right] = C_{ij} - \frac{C_{i0}C_{0j}}{C_{00}} \tag{6}$$

where X_i' denotes the deviation of X_i from its expected value. The Gaussian distribution corresponding to (6) has a number of remarkable features:

- the conditional mean configuration is nonzero and equal to the scaled covariance vector C_{i0} (or autocorrelation vector) of the original process. For a Gaussian process this is also the most likely configuration.
- The conditional variance vector $\mathbb{E}\left[X_i'^2|A\right]$ is zero at $i = 0$, where the conditional process is equal to a by assumption; if the original covariance vector C_{i0} goes to zero for large values of i then the conditional variance matrix relaxes to C_{ij} if both i and j are large.
- Most importantly: the conditional covariance matrix is independent of the value a.

Crucially, the last point implies that the standard deviation of X_i divided by its mean value scales explicitly as $1/a$, which goes to zero for very large a. In other words, the conditional shape of the Gaussian process becomes deterministic as $a \to \infty$ in the sense that the relative size of statistical fluctuations goes to zero in

[1] In this example we picked N consecutive members of X_i but (2) holds with obvious modifications in the general case.

[2] This is easily extended to singular covariance matrices C_{ij}, which have a zero eigenvalue associated with a linear combination of the X_i that yields a deterministic variable. In this case the quadratic form in (2) is defined by its action on the orthogonal eigenvectors and is set to $+\infty$ when acting on the null eigenvector.

this limit. For instance, for a stationary process with $C_{ij} = C_{|i-j|}$ and $\sigma = \sqrt{C_0}$ the signal-to-noise ratio for the conditional process under (5) is

$$\frac{\mathbb{E}\left[X_i|A\right]}{\sqrt{\mathbb{E}\left[X_i'^2|A\right]}} = \frac{a}{\sigma}\frac{C_i}{\sqrt{\sigma^2 - C_i^2}}. \tag{7}$$

This shows that for fixed i the relative error becomes small if $a \gg \sigma$. Also, for fixed relative amplitude a/σ the relative error becomes large if i becomes large and C_i goes to zero. This shows that in the case of a very large amplitude a/σ at $i = 0$ the shape of the random process near $i = 0$ (where $C_i \approx \sigma^2$) is essentially deterministic. Far away from $i = 0$ (where $|C_i| \ll \sigma^2$) the influence of the extreme event has faded away. Based on this fact the easily computed autocorrelation function of a random process emerges as a simple candidate shape for large-amplitude waves.

Draupner rogue wave

As said before, the basic results in the last section are well known in surface wave oceanography as are their natural extensions to continuous functions, in which both the function value and a zero slope can be specified at one point. For moderate values of a/σ this helps discerning maxima of the wave field, although for large a/σ a zero slope is virtually implied by the large function value, which with high probability corresponds to a maximum of the wave field.

This approach has also been used to analyze data sets from rogue waves such as the Draupner wave, although there are far too few data sets to allow a comprehensive study. Of particular interest is the recent work on the Draupner wave by *Walker et al.* [2004], who adjusted the most likely shape in (6a) with nonlinear Stokes corrections up to fifth order. This heuristic procedure, in which the classical Stokes correction expansion for nearly monochromatic small-amplitude surface waves is applied to the Fourier components of the most likely large wave, narrows the peaks and widens the troughs of the shape. Figure 1 shows that this improves the fit with the Draupner wave. Importantly, this nonlinear procedure also breaks the explicit linear up–down symmetry in (6a), which is clearly unrealistic for surface waves because according to this theory the most likely shape of a wave with large surface *depression* $X_0 = -a$ would be given by the inverted shape in (6a).

Large deviation theory

The previous results are examples of large deviation theory (LDT), which deals quite generally with the structure and the probability of rare events in random

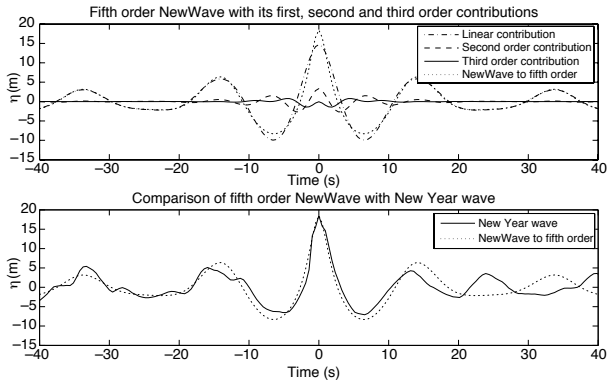

Figure 1. Draupner rogue wave data together with large deviation estimate (termed "NewWave") based on (6a) combined with Stokes corrections. The covariance function was estimated from 20 mins of storm data near the time of the rogue wave. The top panel shows the bare prediction (dot-dashed line) and its modification including Stokes corrections to fifth order (dotted line). The bottom panel shows the fit of actual wave (full line) by the prediction (dotted line). From *Walker et al.* [2004].

systems. LDT allows access to asymptotic results similar to (6a), but in a much wider range of settings. There are two key ingredients: first, that the set of configurations that contribute significantly to the probability of a rare event is tightly localized around the most likely configuration; and second, that the most likely configuration can be computed by constrained minimization of a suitable action functional. Both points are neatly illustrated by the present example of a discrete Gaussian process.

To this end it is convenient to consider the family of scaled processes ϵX_i with covariance matrix $\epsilon^2 C_{ij}$ where $\epsilon > 0$ is a small parameter. We then consider the event that $\epsilon X_0 \geq a$ for some fixed $a > 0$, which is clearly a rare event for small values of ϵ. It is intuitively obvious that for large $a/(\epsilon\sigma)$ the most important contributions to the probability of this event will come from configurations in which $\epsilon X_0 - a$ is small. We can check this because $\mathbb{P}\left[\epsilon X_0 \geq a\right]$ is easily computed from the one-point marginal distribution for X_0 as

$$\frac{1}{\epsilon\sigma\sqrt{2\pi}}\int_a^\infty \exp\left(-\frac{x^2}{2\epsilon^2\sigma^2}\right)dx =$$
$$\exp\left(-\frac{a^2}{2\epsilon^2\sigma^2}\right)\frac{1}{\epsilon\sigma\sqrt{2\pi}}\int_0^\infty \exp\left(-\frac{ay}{\epsilon^2\sigma^2} - \frac{y^2}{2\epsilon^2\sigma^2}\right)dy$$

after the substitution $y = x - a$. Here $\epsilon^2\sigma^2 = \epsilon^2 C_{00}$ is the variance of ϵX_0. By Laplace's method for exponential integrals, the final integral is very well approximated by $\epsilon^2\sigma^2/a$ if $\epsilon \ll 1$. This implies localization between $\epsilon X_0 = a$ and a plus a modest multiple of $\epsilon\sigma^2/a$.

We then have the simple explicit result

$$\mathbb{P}\left[\epsilon X_0 \geq a\right] = \frac{\epsilon}{\sqrt{2\pi}}\,\frac{\sigma}{a}\,\exp\left(-\frac{a^2}{2\epsilon^2\sigma^2}\right) \quad \text{as} \quad \epsilon \to 0. \tag{8}$$

The structure of this expression is typical for the probability of extreme events: an algebraic prefactor related to the size of the set of relevant configurations times an exponentially small term that dominates the decay of the probability for small ϵ. Estimating and computing the exponent in the dominant second factor is the central topic of large deviation theory.

Indeed, we can now determine the most likely coarse-grained configuration by minimizing the quadratic form for $\epsilon X_i \approx \phi_i$ in (4), which is

$$\frac{1}{\epsilon^2} I[\phi] = \frac{1}{2\epsilon^2} \sum_{i,j}^{N} \phi_i C_{ij}^{-1} \phi_j, \tag{9}$$

subject to the constraint $\phi_0 \geq a$ that defines the rare event. It is clear that the minimum is achieved at $\phi_0 = a$. Using a Lagrange multiplier λ it follows for any constraint $g(\phi_1, \ldots, \phi_N) = 0$ that I is extremal where

$$\sum_{j}^{N} C_{ij}^{-1} \phi_j = \lambda \frac{\partial g}{\partial \phi_i} \quad \Leftrightarrow \quad \phi_i = \lambda \sum_{j}^{N} C_{ij} \frac{\partial g}{\partial \phi_j} \tag{10}$$

holds for a value of λ determined from the constraint. The extremal value of I is

$$\frac{1}{2\epsilon^2} \sum_{i,j}^{N} \phi_i C_{ij}^{-1} \phi_j = \frac{\lambda^2}{2\epsilon^2} \sum_{i,j}^{N} \frac{\partial g}{\partial \phi_i} C_{ij} \frac{\partial g}{\partial \phi_j}. \tag{11}$$

In the present case $g = \phi_0 - a$ and therefore $\partial g / \partial \phi_j = \delta_{j0}$ and $\phi_i = \lambda C_{i0}$ follows. Substitution in the constraint yields $\lambda = a/C_{00}$ and this recovers the most likely shape as $\phi_i^* = a C_{i0}/C_{00}$, say, consistent with the conditional expectation in (6a). Now, the minimum of (9) is

$$\frac{1}{\epsilon^2} I^* = \frac{1}{\epsilon^2} I[\phi^*] = \frac{a^2}{2\epsilon^2 C_{00}^2} \sum_{i,j}^{N} \delta_{i0} C_{ij} \delta_{j0} = \frac{a^2}{2\epsilon^2 C_{00}}, \tag{12}$$

which is indeed equal to minus the exponent in (8) with $\sigma^2 = C_{00}$. So this exponent can be computed in LDT as the constrained minimum of the relevant action functional. This is a generic result for LDT in all applications. It can be stated in general form as (e.g., §4 in *Freidlin and Wentzell* [1998])

$$\lim_{\epsilon \to 0} \epsilon^2 \ln \mathbb{P}\left[\epsilon X \in G\right] = -\inf_{\phi \in G} I[\phi] \tag{13}$$

where G denotes the set of configurations that defines the rare event, X and ϕ refer to the corresponding vectors, and $I[\phi]$ is a scaled action functional. In our case $\phi \in G \iff \phi_0 \geq a$ and $I = 0.5 \phi^T C^{-1} \phi$. The formula makes explicit that the logarithm of the probability can be computed from LDT and the corresponding minimization procedure in the limit $\epsilon \to 0$.

The key observation here is that LDT works for very general constraints. For instance, it is easy to add further thresholding constraints at different locations i by adding further Lagrange multiplier terms to (10), which again yields results consistent with the standard extension of the regression formula (6a) for multiple conditions. Nonlinear event constraints involving ϵX_i at multiple locations (e.g. $\phi \in G \iff \phi_1^2 + \phi_2^4 \geq a$) can be treated in precisely the same way in LDT even though there is no simple regression formula such as (6a) available in this case. This illustrates the great flexibility of LDT as a numerical tool.

Finally, the localization property of rare events near the most likely configuration can also be made precise in general form as

$$\lim_{\epsilon \to 0} \mathbb{P}\left[\max_i |\epsilon X_i - \phi_i^*| < \delta | \epsilon X \in G\right] = 1. \tag{14}$$

Here $\delta > 0$ is arbitrary and ϕ^* is the conditional minimizer of $I[\phi]$. This is an impressive mathematical theorem about uniform convergence although it flatters the physical reality a little bit because it is only the scaled process ϵX_i that fits arbitrarily closely (i.e., for any finite δ) to the minimizer ϕ^* at all locations i. In a physical application such as ocean waves it might be more natural to consider the process X_i as fixed and let the constraint scale up as $X_0 \geq a/\epsilon$. Then (14) still holds but with δ replaced by δ/ϵ, which is large in absolute value but again finite and of arbitrary size compared to the large wave amplitude a/ϵ. A more practically useful expectation for the shape of extreme waves is that near the maximum their shape is tightly localized around the most likely shape predicted by LDT whereas far away the process reverts to its unconstrained variability as described by (7).

Continuous Gaussian processes

Both the theory for Gaussian random variables and LDT are easily extended to continuous random functions $X(t)$ of continuous time t, which is the usual setting for presenting observational data and power spectra. In practice, the continuous case can always be reduced to the discrete case by considering a discrete sample $X_i = X(i\Delta t)$. Such an approach with $\Delta t \to 0$ is also necessary in order to construct a probabilistic measure in the space of continuous functions, which yields the analog of (2). However, the continuous version allows some useful analytical tools to be used, most notably Fourier series and the calculus of variations for

the action functional of LDT. We restrict ourselves to stationary random functions and thus we consider the real-valued continuous random function $X(t)$ such that

$$\mathbb{E}\,[X(t)] = 0 \quad \text{and} \quad \mathbb{E}\,[X(t)X(t+s)] = C(s) \tag{15}$$

with an even covariance function $C(s) = C(-s)$. We now restrict to periodic random functions such that $X(t+T) = X(t)$ for some period T. This allows the use of Fourier series and if we choose T much larger than the autocorrelation time of our process then the assumption of periodicity plays a very limited role. If we denote the domain by $L = [-T/2, +T/2]$ then we have the spectral representation

$$C(t) = \frac{1}{T}\sum_{\omega} e^{i\omega t}\hat{C}(\omega) \quad \text{and} \quad \hat{C}(\omega) = \int_L e^{-i\omega t}C(t)\,dt. \tag{16}$$

Here the frequency $\omega = n2\pi/T$ with $n \in Z$ and $\hat{C}(\omega)$ is the discrete energy spectrum[3] of our process. By the nature of $C(s)$, the energy spectrum $\hat{C}(\omega)$ is real, even, and non-negative. As usual, inner products are related by

$$(a,b) = \int_L a^* b\,dt = \frac{1}{T}\sum_{\omega} \hat{a}^*\hat{b} \tag{17}$$

where the star denotes complex conjugation and convolutions satisfy

$$c = a * b = \int_L a(t-s)b(s)\,ds \quad \Rightarrow \quad \hat{c} = \hat{a}\hat{b}. \tag{18}$$

Now, it can be shown (e.g., *Yaglom* [1962]) that $X(t)$ can be represented by the following random Fourier coefficients:

$$\begin{aligned} \omega > 0: &\qquad \hat{X}(\omega) = \sqrt{\frac{T\hat{C}(\omega)}{2}}(A_\omega + iB_\omega) \\ \omega = 0: &\qquad \hat{X}(0) = \sqrt{T\hat{C}(\omega)}A_0 \\ \omega < 0: &\qquad \hat{X}(\omega) = \hat{X}^*(-\omega). \end{aligned}$$

Here all the A_ω and B_ω are independent random numbers drawn from a normal distribution with zero mean and unit variance (note the special treatment for the $\omega = 0$ mode). This explicit formula allows the easy numerical generation of samples from the random process. Actually, there is a simpler and more efficient version, namely

$$\hat{X}_c(\omega) = \sqrt{T\hat{C}(\omega)}(A_\omega + iB_\omega) \tag{19}$$

with i.i.d. coefficients (A_ω, B_ω) for *all* values of ω, including the zero mode. The inverse Fourier transform of $\hat{X}_c$ is complex and its real and imaginary parts are *two* independent samples of the real process $X(t)$. Therefore, this simpler formula allows generating twice the number of samples with the same numerical effort.

In order to compute the action functional for the continuous process we need to find the analog of the quadratic form in (4) and (9) of the discrete case. This involved the inverse covariance matrix C_{ij}^{-1} and it can be shown (e.g., §4 in *Freidlin and Wentzell* [1998]) that for continuous functions this involves the inverse of the self-adjoint non-negative covariance operator $\mathcal{C}$ defined by the convolution $\phi(t) = \mathcal{C}\psi = C*\psi$ such that $\hat{\phi} = \hat{C}\hat{\psi}$. Using this operator the LDT action functional is defined as

$$I[\phi] = \frac{1}{2}(\phi, \mathcal{C}^{-1}\phi). \tag{20}$$

This gives a well-defined answer if $\phi(t)$ is in the range of $\mathcal{C}$. If that is not the case then we set $I = +\infty$, which as before indicates that configurations in the neighbourhood of such $\phi(t)$ have zero probability. Formally, the most likely configuration subject to a functional constraint $g[\phi] = 0$ can be computed using a Lagrange multiplier just as in (10) and yields

$$\phi = \lambda\,\mathcal{C}\frac{\delta g}{\delta \phi} = \lambda\, C * \frac{\delta g}{\delta \phi} \tag{21}$$

for the minimizer $\phi(t)$ in terms of the functional derivative of g. For instance, if $g = \phi(t_0) - a$ we have $\delta g/\delta\phi = \delta(t - t_0)$ and therefore $\phi = \lambda C(t - t_0)$. This shows again that the most likely shape of the process conditioned on taking a certain value at some position is given by the autocorrelation function centred at this position. More generally, if a set G of admissible configurations is defined as before then we again have the LDT results (13-14), where in the second equation the maximum over the index i is replaced by the supremum over $t \in L$.

Now, using (17-18) the action can be written explicitly as

$$I[\phi] = \frac{1}{2T}\sum_{\omega} \hat{\phi}^*\widehat{(\mathcal{C}^{-1}\phi)} = \frac{1}{2T}\sum_{\omega}\frac{|\hat{\phi}|^2}{\hat{C}}. \tag{22}$$

This remarkable formula shows that computing the action density in spectral space reduces to division by the energy spectrum $\hat{C}(\omega)$. This allows easy numerical computation of the action and therefore nonlinear optimization procedures can easily be used to find the most likely shape of extreme waves using LDT. This expression also makes explicit that finite $I[\phi]$ implies that $\hat{\phi}$ is zero whenever $\hat{C} = 0$, otherwise ϕ is not in the range of $\mathcal{C}$ and the action is infinite.

Another advantage of (22) is that in some cases this action density can be converted into an explicit differ-

[3] In observations the energy spectrum is often given by a continuous function $S(\omega)$ such that $\int_0^\infty S\,d\omega = C(0)$. In this case $\pi S(\omega) = \hat{C}(\omega)$.

ential operator acting on the field $\phi(t)$. We give some examples in the next section.

Three examples and numerical LDT

The first example is the stationary Ornstein–Uhlenbeck process (see also (35) below), which is the prototypical example of red noise and also arises as the invariant distribution of the linear Langevin equation. It is characterized by

$$C_1(t) = \frac{1}{2}\exp(-|t|) \quad \text{and} \quad \hat{C}_1(\omega) = \frac{1}{1+\omega^2}. \tag{23}$$

The sample paths are almost surely everywhere continuous but nowhere differentiable. This non-smooth behaviour is typical for processes driven by white noise and its trademark is the slow ω^{-2} decay of the energy spectrum for large ω. The action functional is

$$I_1[\phi] = \frac{1}{2T}\sum_{\omega} |\hat{\phi}|^2(1+\omega^2) = \frac{1}{2}\int_L \left(\phi^2 + \phi_t^2\right)\, dt, \tag{24}$$

which follows from (17) and $\widehat{(\phi_t)} = i\omega\hat{\phi}$. This explicit differential form allows the use of calculus of variations. For instance, the Euler–Lagrange equation for (24) is the ODE $\phi_{tt} - \phi = 0$ and minimizers constrained to have fixed values at two points t_1 and t_2, say, will satisfy those boundary conditions as well as this ODE. The autocorrelation $C_1(t)$ satisfies this ODE together with a decay condition at infinity, which shows once more that $C_1(t)$ is the most likely shape conditional on an isolated fixed value.

Clearly, whenever the energy spectrum is the reciprocal of a polynomial in ω^2 then the action functional can be written as an integral over a quadratic form in the derivatives of the function. For instance, a smoother version of (23b) is $\hat{C}_2(\omega) = 1/(1+\omega^4)$ with covariance function

$$C_2(t) = \frac{1}{2}\exp\left(-\frac{|t|}{\sqrt{2}}\right)\cos\left(\frac{|t|}{\sqrt{2}} - \frac{\pi}{4}\right) \tag{25}$$

and action

$$I_2[\phi] = \frac{1}{2}\int_L \left(\phi^2 + \phi_{tt}^2\right)\, dt. \tag{26}$$

On the other hand, the spectrum $\hat{C}_3(\omega) = \omega^2/(1+\omega^4)$ with covariance function

$$C_3(t) = \frac{1}{2}\exp\left(-\frac{|t|}{\sqrt{2}}\right)\cos\left(\frac{|t|}{\sqrt{2}} + \frac{\pi}{4}\right) \tag{27}$$

does not lead to a simple expression for the action in terms of local derivatives. This third process is interesting because its spectrum has an interior maximum at $\omega^2 = 1$, which is typical for many energy spectra in geophysical fluid dynamics such as the frequency spectrum of surface waves, for instance. The three autocorrelation functions are plotted in figure 2.

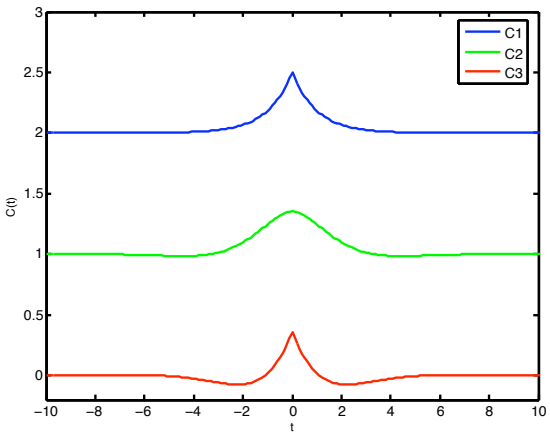

Figure 2. Staggered plot of the autocorrelation functions $C_1(t)$, $C_2(t)$, and $C_3(t)$. These are the most likely shapes of a Gaussian process under the constraint that the value at the origin exceeds a given positive threshold.

It is easy to minimize the action functional of LDT with a numerical method for very general kinds of constraints. A simple method combines a down-gradient flow together with a penalty term that enforces the constraint. Specifically, we introduce a marching time τ and define a smooth flow of $\phi(t,\tau)$ such that ϕ converges to the minimizer $\phi^*(t)$ as $\tau \to \infty$. The minimization procedure is then defined by

$$\frac{\partial\phi}{\partial\tau} = -\mathcal{C}^{-1}\phi + h \quad \Leftrightarrow \quad \frac{\partial\hat{\phi}}{\partial\tau} = -\frac{\hat{\phi}}{\hat{C}} + \hat{h}. \tag{28}$$

The first, down-gradient term is just minus the functional derivative $\delta I/\delta\phi$ and it is easily computed in spectral space. The second, penalty term is usually more easily computed in real space. For example, if the constraint is $\phi \geq a$ in some set $B \subset L$ then one can use

$$h = \alpha(1 - \tanh(\beta(\phi - a))) \quad \text{if} \quad t \in B \tag{29}$$

and $h = 0$ otherwise. The constants α and β are adjusted to make the scheme work; here $\beta = 100$ was always used and α varied between 20 and 100. The second equation in (28) is evolved in τ by freezing h at the beginning of the time step, which leads to

$$\hat{\phi}(\omega,\tau+\Delta\tau) = e^{-\frac{\Delta\tau}{\hat{C}}}\hat{\phi}(\omega,\tau) + \left(1 - e^{-\frac{\Delta\tau}{\hat{C}}}\right)\hat{C}\hat{h}(\tau). \tag{30}$$

It is the constraint that couples different Fourier modes. This scheme is applied for a finite range of ω and alternated with recomputing h from (29) until convergence

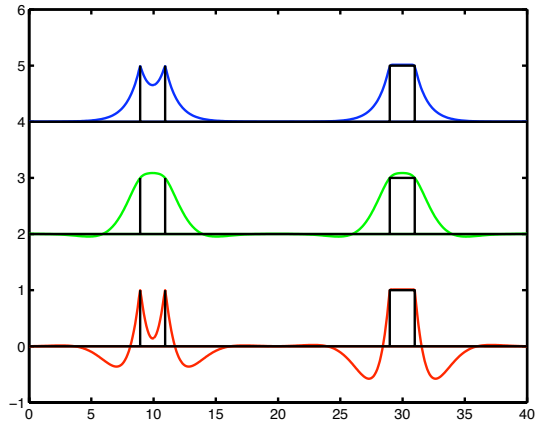

Figure 3. Most likely shapes subject to constraints indicated by the black lines. On the left $\phi \geq 1$ is enforced at two points and on the right the same constraint is enforced over the closed interval B between the points.

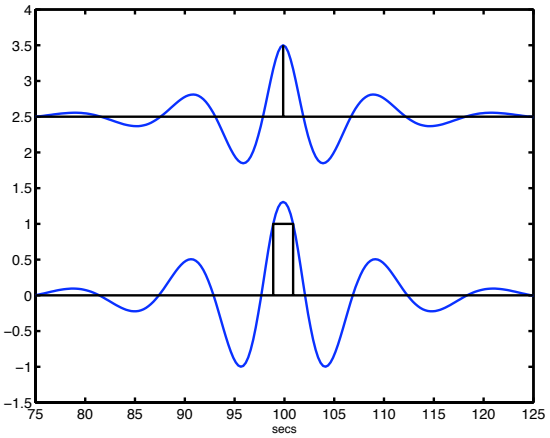

Figure 4. Autocorrelation function and most likely shape for standard JONSWAP spectrum dominated by 10 second surface waves. The constraint $\phi > 1$ is enforced first at the centre and then over 2 seconds as indicated by the black lines. There are notable depression precursors in either case.

is reached. Two examples are shown in figure 3. In the left column the constraint was $\phi \geq 1$ at *two* points t_0 and $t_0 + 2$ whilst in the right column this constraint was enforced throughout the interval $B = [t_0, t_0 + 2]$. There are several interesting features here. For C_1 the most likely shapes are non-negative, they are identical outside B (which could have been guessed from (24b)), and the first shape dips inside B to lower the action whilst the second shape hugs the constraint $\phi = 1$ for the same reason. In contrast, for C_3 the shapes are not identical outside B and there are also negative undershoots outside B; these are depression precursors before the high wave hits.

In the case of C_2, the two-point shape bulges upward in B and therefore *exceeds* the threshold $\phi = 1$ there. This implies that both shapes are completely identical, because the two-point constraint already implies the interval constraint. This would have been hard to guess and indeed the occurrence of this overshooting feature depends on the width of the interval. This can be understood from the exact solution for the two-point problem, which from (21) with both $t_0 = 0$ and $t_0 = d$ is the linear combination of covariance functions

$$\phi^*(t) = \frac{C(t) + C(t-d)}{C(0) + C(d)}. \tag{31}$$

Overshooting means that this function exceeds unity for some $t \in (0, d)$. This is impossible if $C(t)$ is convex for all $t > 0$ (which here rules out C_1) but for non-convex C this can be possible. For instance, under the assumption that the overshoot occurs at the midpoint $t = d/2$ (which must be true for very small d) overshooting occurs if $C(d/2) > (C(0) + C(d))/2$. By visual inspection of figure 2, this criterion rules out C_3 but not C_2, which is concave near the origin. Indeed, any smooth covariance function at $t = 0$ satisfies this criterion for small d and therefore any differentiable random function such that $-C''(0) = \mathbb{E}[\dot{X}^2] < \infty$ is capable of overshoots for small enough d.

Finally, figure 4 shows the analogous results for a standard JONSWAP spectrum for surface waves with central frequency of 0.1 Hertz. Clearly, for a two-second threshold constraint the most likely shape overshoots the threshold, in this case by some 30%.

Importance sampling using LDT

The minimal action $I[\phi^*]$ corresponding to the most likely shape ϕ^* gives the exponentially small part of the probability of the rare event under consideration. However, it does not give the prefactor in front of the exponential (cf. (8)), which must be evaluated separately, usually by using numerical sampling. In fact, one very practical use of the maximal shapes predicted by LDT lies in importance sampling, which greatly improves convergence for the numerical estimation of the probability of rare events. Basically, by concentrating samples near the most likely shape the variance of the estimator can be sharply reduced because the exponentially small part of the probability is known explicitly and need not be estimated. This is a great practical advantage.

In general, we may wish to estimate $p_G = \mathbb{P}[X \in G]$ for some set G by drawing N independent samples of

the process X and computing the proportion of samples that fall in G. By the law of large numbers this proportion converges to p_G and the variance of this estimator is $p_G(1-p_G)/N$; this follows directly from $p_G = \mathbb{E}[\chi(X \in G)]$ where $\chi(\cdot)$ is the indicator function. The expected relative error after N samples is

$$\sqrt{\frac{1-p_G}{Np_G}} \approx \sqrt{\frac{1}{Np_G}} \tag{32}$$

for small p_G. This shows that $N \gg p_G$ samples are needed to estimate the probability of a rare event such that $p_G \ll 1$. If p_G is exponentially small, then this is a hopeless numerical task.

For example, (8) gave the probability for the event that a single zero-mean Gaussian variable ϵX with variance $\epsilon^2\sigma^2$ exceeds the threshold a. This probability is $\propto \exp(-I^*/\epsilon^2)$ and hence would be impossible to estimate directly for small ϵ. Now, the most likely configuration is $X = a/\epsilon$ and the equation preceding (8) can be viewed as giving the probability that $Y = X - a/\epsilon$ exceeds zero. The sought-after probability is now an explicit exponential prefactor (with the now familiar minimum action in the exponent) times an integral that can be viewed as the expected value of the function $\chi(Y > 0)\exp(-Ya/(\epsilon\sigma^2))$ relative to a normal distribution of Y with zero mean and variance σ^2. It is easy to show that both the expectation and the variance of this function are $O(\epsilon)$ and therefore the relative sampling error from (32) is now small if $N \gg 1/\epsilon$. This is a vast improvement over the condition $N \gg \exp(-I^*/\epsilon^2)$ without importance sampling.

Analogous results hold for multiple variables and discrete or continuous processes: the dominant part $\exp(-I^*/\epsilon^2)$ of the probability can be computed explicitly and the remaining prefactor can then be estimated with a well-conditioned numerical procedure by considering $\epsilon Y = \epsilon X - \phi^*$ where ϕ^* is the minimizer of the action relative to the constraint.

LDT for evolution equations

The general methods of LDT are not restricted to Gaussian processes and in particular can be applied to many evolution equations that contain some random component (e.g., *E et al.* [2004]). Examples include Markov chains, autoregressive processes, and dynamical equations under random forcing (see *Varadhan* [2003] for a discussion of such applications using entropy as a unifying principle). A simple example is the Langevin equation for the time-evolution of a continuous vector $X_t \in R^n$ such that

$$dX_t = b(X_t)dt + \epsilon\, dW_t \quad \text{and} \quad X_0 = x. \tag{33}$$

This is a stochastic differential equation in which dW_t is a the increment of the Wiener process, or standard Brownian motion. This is a vector of random numbers $(dW_t^1, dW_t^2, \ldots, dW_t^n)$ that are independent and identically distributed with a normal distribution that has zero mean and variance equal to dt (e.g., *Gardiner* [1997], *Oksendal* [2002]). Without this random term (33) would be a deterministic initial-value problem for X_t such that $\dot{X}_t = b$ with $t \in [0, T]$, say. For instance, this could describe truncated evolution equations for waves or other geophysical processes and the added random forcing term might represent unresolved degrees of freedom or other external influences and parametrizations that are not explicitly resolved.

For small ϵ the solution stays close to the deterministic trajectory, but it is now possible to deviate from this path and the probabilities to do so satisfy an action principle based on

$$I[\phi] = \frac{1}{2}\int_0^T \left|\dot{\phi} - b(\phi)\right|^2 dt \tag{34}$$

such that the probability of $\sup_t |X_t - \phi| < \delta$ is proportional to $\exp(-I[\phi]/\epsilon^2)$. Here $\phi(t) \in R^n$ is a function satisfying $\phi(0) = x$ and without further constraints the minimal action is achieved if ϕ satisfies the deterministic equation. However, constraints can again be added and then the optimal ϕ can be found by minimizing (34) using the calculus of variations. This yields most likely trajectories that are *not* trajectories of the deterministic system.

It is notable that the random function X_t is not a Gaussian random function in general. For instance, if $b(X_t) = -\nabla H(X_t)$ then the invariant probability distribution for (33) can easily be shown to be $A\exp(-2H/\epsilon^2)$ for some constant A, which incidentally is the canonical distribution of statistical mechanics with Hamiltonian H. However, this is not Gaussian unless H is quadratic in X_t. Notably, in the special one-dimensional case $H = X_t^2/2$ we have the linear Langevin equation

$$dX_t = -X_t dt + \epsilon\, dW_t \tag{35}$$

whose solution is the Ornstein–Uhlenbeck process (with stationary covariance function $(\epsilon^2/2)\exp(-|t|)$) discussed as the first example in (23). This special process is both Gaussian and Markovian. The action functional is

$$\frac{1}{2}\int_0^T \left(\dot{\phi} + \phi\right)^2 dt = \frac{1}{2}\int_0^T (\dot{\phi}^2 + \phi^2)\, dt + \frac{1}{2}\,\phi^2\Big|_0^T. \tag{36}$$

For constraints with fixed end points this reduces to the functional (24b) for the periodic case.

The Langevin equation (33) is easily generalized to non-uniform noise terms such that $dW \in R^m$ and there

is a constant matrix $\sigma \in R^{n\times m}$ such that

$$dX_t = b(X_t)dt + \epsilon\,\sigma dW_t. \tag{37}$$

This allows for correlations between noise terms in different components of this equation. The corresponding action functional is

$$I[\phi] = \frac{1}{2}\int_0^T \left(\dot\phi - b\right)^T A^{-1} \left(\dot\phi - b\right) dt \tag{38}$$

where the quadratic form is governed by the inverse of the matrix $A = \sigma\sigma^T$. This works easily provided $A \in R^{n\times n}$ is invertible, as usual. Actually, (38) also applies to the case of 'multiplicative' noise, in which σ is a function of X_t. Notably, in this case the solution of the stochastic differential equation (37) depends on the precise definition of the noise term and care needs to be taken to use the right definition for modelling the physical situation at hand. This modelling problem brings in the well-known differences between the Itô and Stratonovich versions of the stochastic integral (e.g., *Gardiner* [1997], *Itô* [1974]). However, for fixed T these differences play no role in the leading-order expression for the action functional as $\epsilon \to 0$.

Finally, importance sampling using the most likely path of LDT can be applied here as well. In the context of (37) this involves the use of Girsanov's formula to transform the probabilistic measure between X_t and the centred path $Y_t = X_t - \phi^*$ (e.g., *Liu and Vanden-Eijnden* [2007]). This is analogous to the Gaussian process discussed before.

Concluding comments

LDT applies in essentially unchanged form to multi-dimensional processes, although in practice additional assumptions about the interpretation of observed spectra are needed in this case. For instance, for ocean surface waves the space–time spectra might be isotropic in the horizontal directions, but only because they represent an average over many wave realizations that individually had a strong preference in the wind direction, say. This preference is averaged out in the observed spectrum if the wind directions is random itself. Interpreting the observed spectra in this light affects the shape of the rare surface waves.

At the other end of the ocean there is the generation (and eventual dissipation) of internal tides by undulating topography (e.g., *Garrett and Kunze* [2007]). The topography can be viewed as a mixture of large-scale resolved and small-scale unresolved statistical features and LDT can be used to predict the shape and probability of random waves that exceed the breaking amplitude in the ocean interior, say, by three-dimensional focusing effects (e.g., *Bühler and Muller* [2007]).

Finally, a completely different application of LDT in geophysics could be in data assimilation. The probability and spatial structure of large errors in the assimilated fields should again be governed by a suitable version of LDT and so this could be a useful tool to apply in this area.

Acknowledgments. It is a pleasure to acknowledge many interesting discussions about this topic with E. Tabak and E. Vanden-Eijnden as well as a tutorial by S.R.S. Varadhan. This work is supported partly by the National Science Foundation grants OCE-0324934 and DMS-0604519. Finally, I would like to thank the organizers for the opportunity to attend this unique workshop.

References

Boccotti, P., Quasi- determinism of sea wave groups, *Meccanica*, *24*, 3–14, 1989.

Bühler, O., and C. M. Muller, Instability and focusing of internal tides in the deep ocean, *J. Fluid Mech.*, p. submitted, 2007.

E, W., W. Ren, and E. Vanden-Eijnden, Minimum action method for the study of rare events, *Comm. Pure Appl. Math.*, *57*, 637–656, 2004.

Freidlin, M., and A. Wentzell, *Random Perturbations of Dynamical Systems*, 2 ed., Springer, 1998.

Gardiner, C., *Handbook of Stochastic Methods*, 2 ed., Springer, 1997.

Garrett, C., and E. Kunze, Internal tide generation in the deep ocean, *Ann. Rev. Fluid Mech.*, *39*, 57–87, 2007.

Itô, K., Stochastic differentials, *Applied Math. and Optimization*, *1*, 374–381, 1974.

Liu, D., and E. Vanden-Eijnden, Variance reduction techniques based on measure transformation, 2007, in preparation.

Oksendal, B., *Stochastic Differential Equations*, 5 ed., Springer, 2002.

Phillips, O., D. Gu, and M. Donelan, Expected structure of extreme waves in a Gaussian sea. part i: theory and SWADE buoy measurements, *J. Phys. Ocean.*, *23*, 992–1000, 1993.

Varadhan, S. R. S., *Large Deviations and Applications*, SIAM, Philadelphia, 1994.

Varadhan, S. R. S., Large deviations and entropy, in *Entropy*, pp. 197–214, Princeton University Press, 2003.

Walker, D. A. G., P. H. Taylor, and R. Eatock Taylor, The shape of large surface waves on the open sea and the Draupner New Year wave, *Applied Ocean Research*, *26*, 73–83, 2004.

Yaglom, A. M., *An Introduction to the Theory of Stationary Random Functions*, Dover, 1962.