Lecture Notes on Stochastic System Analysis

Siu-Kui Au
School of Civil & Environmental Engineering
Nanyang Technological University, Singapore

1. Introduction

1.0. Notation

We use capital letters to denote an uncertain parameter (random variable) and use lower case letters to denote its parameter value. For integrals, when the integration domain is not specified, it is assumed that the integration is over the whole domain. Vector-valued quantities are indicated by an underline. We reserve $P(\cdot)$ to denote the probability of a proposition and $p(\cdot)$ to denote probability densities. Common abbreviations and symbols are as follows:

PDF	: Probability density function
CDF	: Cumulative distribution function
MCS	: Monte Carlo Simulation
i.i.d.	: independent and identically distributed
$E_f[\cdot]$	: Expectation under PDF f If f is not specified, it is understood to be the parameter PDF q
$\mathrm{var}_f[\cdot]$	: variance under PDF f
$\|\cdot\|$	: Euclidean norm of a vector. For $\underline{x} \in R^n$, $\|\underline{x}\| = \sqrt{\sum_{i=1}^{n} x_i^{\ 2}}$
$\langle \cdot,\cdot \rangle$	: Inner product between two vectors. For $\underline{x}, \underline{y} \in R^n$, $\langle \underline{x}, \underline{y} \rangle = \sum_{i=1}^{n} x_i y_i$

1.1. Probability integrals

Let $\underline{X} = [X_1, ..., X_n]^T \in R^n$ be a vector of uncertain parameters modeled as random variables with joint probability density function and let $h: R^n \to R^+ \cup \{0\}$ be some non-negative function of $\underline{X}$ (often related to the response quantity of interest). In probabilistic analysis, we are often interested in computing the expectation of $h(\underline{X})$, which can be expressed as a 'probability integral' of the form:

$$J = E[h(X)] = \int h(\underline{x}) q(\underline{x}) d\underline{x}$$

Example 1.1.
Let $y(\underline{X})$ be the displacement of a structure that depends on the loading specified by $\underline{X}$, and let $\mu = E[y(\underline{X})]$ denote its expectation. Then if $h(\underline{X}) = (y(X) - \mu)^2$, J will be the variance of $y(\underline{X})$.

Example 1.2.
Define the indicator function $I(A)$ which takes as its argument a set or a proposition A, and returns 1 if A is true and zero otherwise, i.e.,

$$I(A) = \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{if } A \text{ is false} \end{cases}$$

Then $E[I(A)] = 1 \times P(A) + 0 \times P(\sim A) = P(A)$. Thus, if we let $h(\underline{X}) = I(y(X) > b)$ for some fixed b, $J = P(y(X) > b)$. In reliability theory, we usually call this the failure probability of $y(X)$ (exceeding b).

1.2. Failure probability and complementary CDF

In engineering applications, we often need to deal with failure events that are expressed in terms of union or intersection of exceedance events, say, corresponding to system components connected (logically) in series or in parallel. It turns out that failure events of this kind can always be expressed in terms of the exceedance of a single variable.

Example 1.3.
Let $F = \bigcup_{i=1}^{m} \{Y_i > b_i\}$, then F can be expressed as $F = \{Y > 1\}$ where $Y = \max_{i=1,\dots,m} Y_i / b_i$. The proof is left as an exercise.

Example 1.4.
Let $F = \bigcap_{i=1}^{m} \{Y_i > b_i\}$, then F can be expressed as $F = \{Y > 1\}$ where $Y = \min_{i=1,\dots,m} Y_i / b_i$. The proof is left as an exercise.

In general, a failure event defined by stacking of union or intersection of exceedance events can be expressed as the exceedance event of a single variable defined by stacking of max. or min. in the same order.

Example 1.5.
Let $F = \bigcap_{i=1}^{m_1} \bigcup_{j=1}^{m_2} \bigcap_{k=1}^{m_3} \{Y_{ijk} > b_{ijk}\}$, then $F = \{Y > 1\}$ where $Y = \min_{i=1,\dots,m_1} \max_{j=1,\dots,m_2} \min_{k=1,\dots,m_3} Y_{ijk} / b_{ijk}$.

This generic representation of failure event builds a link to the exceedance event of a single variable. Further, if we consider $F_y = \{Y > y\}$ generally for some other values of y (in addition to $y = 1$), then this will induce a series of failure events that correspond to different threshold values. E.g., if $F = \bigcup_{i=1}^{m}\{Y_i > b_i\} = \{Y > 1\}$ where $Y = \max_{i=1,\ldots,m} Y_i / b_i$, then $F_y = \{Y > y\} = \bigcup_{i=1}^{m}\{Y_i > b_i y\}$. Thus, by considering $P(F_y) = P(Y > y)$ for different values of y, we can investigate the trend of the probability failure at different extents, which is more informative than just a point estimate of $P(F) = P(Y > 1)$.

Viewing $P(F_y) = P(Y > y)$ as a function of y, finding the failure probability in our context is equivalent to finding the complementary CDF (cumulative distribution function) of a single variable Y (the complementary CDF = 1 – CDF), especially at the tail when small failure probabilities are the main interest.

1.3. Problem context

Throughout our discussion, we assume we are dealing with probability integrals under the following context (unless otherwise specified):

1. q is assumed to be given and correspond to some 'known' distributions (e.g., consisting of Gaussian, exponential PDFs) so that
   a. we can evaluate $q(\underline{x})$ efficiently for any given $\underline{x}$
   b. we can generate random samples from q efficiently

   This assumption distinguish our problem from Bayesian updating problems, where in the latter case the PDF is often of some form for which the efficient generation of random samples is highly-nontrivial and the PDF is only known up to a scaling constant.

2. $X_1,\ldots,X_n$ are assumed to be mutually independent. This does not generate much loss of generality because in forward analysis problems dependent variables are generated by independent ones.

3. h is a non-negative function. The relationship between $\underline{x}$ and $h(\underline{x})$ may not be explicitly known, in the sense that, although we can evaluate the value of $h(\underline{x})$ for a given $\underline{x}$, we may not be able to obtain other information such as gradient or Hessian; the latter quantities may have to be computed numerically, e.g., using finite difference.

4. The computational effort for evaluating $h(\underline{x})$ for each value of $\underline{x}$ is significant, and we want to reduce the number of evaluations in our computational procedure for evaluating J.

5. We are interested in small failure probabilities, or equivalently, the tail of the complementary CDF of Y.

6. We will pay attention to whether a particular method will work for large n (theoretically infinite), i.e., high dimensions. Computational methods that are applicable in high dimensions are generally more robust and more sustainable.

1.4. Different perspectives

We will discuss different approaches available for evaluating probability integrals. Setting aside the technical issues in the particular algorithms, the limitation and the mechanism by which the algorithms derive their computational efficiency can be better appreciated when the perspective through which the integral is viewed by a particular method is recognized. A brief overview of different approaches to be discussed here is given below by surveying through different perspectives.

1) *Numerical integration*

If we view the probability integral as a sum of differential contributions $h(\underline{x})q(\underline{x})d\underline{x}$, then we will use numerical integration to evaluate J, e.g., dividing R^n into a number of disjoint hypercubes (intervals if $n=1$), evaluate $h(\underline{x})q(\underline{x})d\underline{x}$ at the center of each element, and then sum the contribution from all the hypercubes as an approximation for J. If we divide the each dimension into N intervals, then the total number of elements is N^n and this often leads to an error of $O(N^{-1})$. Put it the other way, if altogether we spend N evaluations, then the approximation error is $O(N^{-1/n})$. The dependence of the error on n has important implication on the applicability of numerical integration – it is not efficient in high dimensions and for $n \geq 3$ is inferior to Monte Carlo simulation, whose error is $O(N^{-1/2})$ regardless of n. We will not discuss numerical integration here.

2) *Asymptotic approximation*

In many applications, it is observed that the integrand $h(\underline{x})q(\underline{x})$ has one or more peaks in R^n, which can be reasoned from the fact that $q(\underline{x})$ is a parameter PDF that is often peaked at the mean value. If we recognize that the main contribution of J comes from the neighborhood of the peak(s), then we can first identify their neighborhood, e.g., by locating the peak(s), and then try to make use of information at the peak(s) to approximate J. This leads to a class of methods called 'asymptotic approximation' (or, it can be viewed simply as a Gaussian approximation of integrand). These methods provide estimates for J whose error, however, cannot be reduced by spending more computational effort. Further improvement of accuracy can be achieved by adopting a stochastic simulation method called 'importance sampling' which makes use of the information already obtained about the important region but whose stochastic character allows the estimate be improved and to converge to the exact value when more computational effort is spent.

3) *Direct Monte Carlo*

The probability integral J can be viewed as a mathematical expectation of $h(\underline{X})$ with $\underline{X}$ distributed as q, and this perspective leads to the direct Monte Carlo method, where J is estimated as a sample average of h over independent and identically distributed samples of $\underline{X}$ drawn from the PDF q. The error of this method is $O(N^{-1/2})$ regardless of n and so it is extremely robust with respect to applications. Of course, since now the estimate for J depends on the random numbers that are generated in each trial, we will get different answers in different trials, and the error is better measured in terms of the variance of the estimator rather than the error involved in a particular trial. Methods that involve generating random samples for statistical averaging are referred here as ‘stochastic simulation method’.

4) Importance sampling method

It turns out, however, that when we try to use direct Monte Carlo to estimate failure probabilities (or generally some h that takes on significant value at the tail of q), the relative error (relateive to the target failure probability we want to estimate) grows dramatically with decreasing failure probability. Essentially, if the failure probability is P_F, then for small P_F the relative error is approximately $1/\sqrt{P_F N}$ and so will be quite large for rare failure events (small P_F). There are stochastic simulation methods that try to reduce this relative error, collectively known as ‘variance reduction techniques’. One popular way to reduce the variance is to gain some information about the important region in the parameter space that gives significant contribution to J, and then try to use such information to construct a simulation method that can generate more samples in that important region, which hopefully can reduce the variance of estimator. We will discuss one method following this spirit, called ‘importance sampling simulation’.

5) Subset Simulation

We will then find that there are many cases where getting information about the important region directly is indeed quite difficult, and we will sail along a some what different perspective, where we view a small failure probability as a product of large conditional failure probabilities. In terms of simulation, we effectively convert a rare simulation problem into a sequence of more frequent ones, and we progressively generate samples that populate towards the important region. Such method is called ‘Subset Simulation’, which is the result of insights about the probabilistic structure of rare events and the use of a powerful simulation method called Markov Chain Monte Carlo (MCMC) method (which can be a big topic by itself!).

2. ASYMPTOTIC APPROXIMATION

2.0. Preliminaries

We start with a brief introduction of asymptotic relationships. In Calculus, we say that as $x \to a$, $f(x) \to g(x)$ if $\lim_{x \to a} f(x) - g(x) = 0$. In asymptotic theory, we say that as $x \to a$, $f(x) \sim g(x)$ (read as 'f is asymptotic to g as $x \to a$' or 'f is asymptotically equivalent to g as $x \to a$') if $\lim_{x \to a} f(x)/g(x) = 1$. That is, asymptotics deals with limit of ratios. Note the followings:

1. As $x \to a$, let f, g be non-vanishing. If $f \to g$, then $f \sim g$ as well, i.e., convergence implies asymptotics. The proof is left as an exercise. Note that the reverse is not true.
2. The requirement that f, g are non-vanishing in the last comment is necessary, e.g., consider $f = x$ and $g = x^2$, then $f \to g \to 0$ as $x \to 0$ but $f/g = 1/x \to \infty$ as $x \to \infty$.

Example 2.1.

Let $f = x^2 + x$ and $g = x^2$, then as $x \to \infty$, $f/g = 1 + 1/x \to 1$ so $f \sim g$. However, $f - g = x \to \infty$ and so f does not converge to g as $x \to \infty$. This illustrates that asymptotic relation is generally weaker than convergence. In fact, f and g only need to have their 'dominant' term (x^2 in our example) to be equal in the limit, in order for them to be asymptotically equivalent.

2.1. Approximation based on Taylor series of log integrand

Consider again

$$J = \int h(\underline{x}) q(\underline{x}) d\underline{x} = \int k(\underline{x}) d\underline{x}$$

where for simplicity in notation we have let $k(\underline{x}) = h(\underline{x}) q(\underline{x})$. The method that we will discuss in this chapter depends on information about $k(\underline{x}) \geq 0$ and does not need to resolve into the details of h and q. Here we assume that $k(\underline{x})$ has only one a peak at, say, $\underline{x}^*$ (often called 'design point' or 'check point' in reliability literature) and the Hessian matrix of $k(\underline{x})$ at $\underline{x}^*$ is negative definite. We are going to derive an approximation of J based on approximating $\ln k(\underline{x})$ using up to second order derivatives at its peak (we will explain why we approximate $\ln k(\underline{x})$ instead of $k(\underline{x})$ later).

First, we approximate $\ln k(\underline{x})$ by a second order Taylor series about $\underline{x}^*$,

$$\ln k(\underline{x}) \approx \ln k(\underline{x}^*) + [\nabla \ln k(\underline{x}^*)](\underline{x} - \underline{x}^*) + (\underline{x} - \underline{x}^*)^T H_{\ln k}(\underline{x}^*)(\underline{x} - \underline{x}^*)$$

where $\nabla \ln k(\underline{x}^*)$ and $H_{\ln k}(\underline{x}^*)$ are the gradient and Hessian of $\ln k$ at $\underline{x}^*$. Note that it is sufficient for the above approximation to be good in the neighborhood of

$\underline{x}^*$ since we assume that elsewhere the contribution to the integral is insignificant.

Since $\underline{x}^*$ maximizes $k(\underline{x})$, $\nabla k(\underline{x}^*)=\underline{0}$, and so $\nabla \ln k(\underline{x}^*)=k(\underline{x}^*)^{-1}\nabla k(\underline{x}^*)=\underline{0}$. Using this fact and writing $H_{\ln k}(\underline{x}^*)=-H_{-\ln k}(\underline{x}^*)$, we have

$$\ln k(\underline{x}) \approx \ln k(\underline{x}^*) - \frac{1}{2}(\underline{x}-\underline{x}^*)^T H_{-\ln k}(\underline{x}^*)(\underline{x}-\underline{x}^*)$$

Substituting this approximation into the integral, we have

$$J \approx k(\underline{x}^*)\int \exp\left[-\frac{1}{2}(\underline{x}-\underline{x}^*)^T H_{-\ln k}(\underline{x}^*)(\underline{x}-\underline{x}^*)\right]d\underline{x} \qquad (2.1)$$

To proceed, we need to evaluate the integral in the above equation, and it turns out that it can be integrated analytically. For this, we note that, since $H_k(\underline{x}^*)$ is negative definite, so is $H_{\ln k}(\underline{x}^*)$, and hence $H_{-\ln k}(\underline{x}^*)=-H_{\ln k}(\underline{x}^*)$ is positive definite. Thus, if we let $\underline{C}=[H_{-\ln k}(\underline{x}^*)]^{-1}$, $\underline{C}$ will also be a positive definite matrix, since its eigenvalues are the reciprocal of those of $H_{-\ln k}(\underline{x}^*)$. Noting that any positive definite matrix is a legitimate covariance matrix, we consider the n-dimensional Gaussian PDF with mean $\underline{x}^*$ and covariance matrix $\underline{C}$:

$$\phi(\underline{x})=\frac{(2\pi)^{-n/2}}{\sqrt{\det \underline{C}}}\exp\left[-\frac{1}{2}(\underline{x}-\underline{x}^*)^T \underline{C}^{-1}(\underline{x}-\underline{x}^*)\right]$$

Now, by noting that the PDF integrates to 1 over R^n, we immediately get

$$\int \exp\left[-\frac{1}{2}(\underline{x}-\underline{x}^*)^T \underline{C}^{-1}(\underline{x}-\underline{x}^*)\right]d\underline{x}=(2\pi)^{n/2}\sqrt{\det \underline{C}}=\frac{(2\pi)^{n/2}}{\sqrt{|\det H_{\ln k}(\underline{x}^*)|}} \quad (2.2)$$

where the last equality has made use of the fact that the determinant is equal to the product of eigenvalues and that the eigenvalues of $\underline{C}$ are just the reciprocal of those of $H_{-\ln k}(\underline{x}^*)=-H_{\ln k}(\underline{x}^*)$. Substituting this integral into (2.1), we obtain

$$J \approx \frac{(2\pi)^{n/2}k(\underline{x}^*)}{\sqrt{|\det H_{\ln k}(\underline{x}^*)|}} \quad (2.3)$$

Some comments about this approximation are in order:

1. The formula depends on the value and Hessian of the integrand at the design point $\underline{x}^*$ (the gradient information does not appear because it is zero at $\underline{x}^*$).

2. The approximation involves approximating the integrand by an exponential function of a concave quadratic form, i.e., a Gaussian type function. Obviously, whether the formula can give a good approximation to the actual value of J depends on how close the integrand is to a Gaussian type function, and seems to have nothing to do with ‘asymptotics’. So why is it called ‘asymptotic approximation’? It is due to the fact that this approach is

motivated by the asymptotic results of Laplace integrals, which will be discussed in the next section.

3. One may wonder why we used a Taylor series to approximate $\ln k$ rather than k. It is because in applications the integrands that we are dealing with are non-negative and decay to zero as $\| \underline{x} - \underline{x}^* \| \to 0$, the latter due to the decaying property of $q(\underline{x})$. A Gaussian type function that resulted from approximating $\ln k$ as a second order Taylor series has this property. In contrast, a second order Taylor series approximation of k will produce a concave quadratic form that to tend to negative infinity as $\| \underline{x} - \underline{x}^* \| \to 0$, which violates the behavior of k. Finally, in many applications q is chosen to correspond to some exponential family (e.g., Gaussian PDF, Exponential PDF) and so a Gaussian type function seems a good form for approximation. Of course, in a particular application, if the form of the integrand is known, one may be able to come up with a better approximation by using a functional form that is closer to that of the integrand.

4. When there are more than one design points, a similar argument generalizes our result to include contributions from all design points:

$$J \approx \sum_{i=1}^{m} \frac{(2\pi)^{n/2} k(\underline{x}_i^*)}{\sqrt{|\det H_{\ln k}(\underline{x}_i^*)|}}$$

In this case, the approximation is equivalent to fitting the integrand with a sum of Gaussian type functions centered at each design point. For neighboring design points (especially when they are close), their fitting functions may overlap, and this may lead to a poor approximation. One may try to come up with a way to account for such overlap although it is somewhat non-trivial.

2.2. Asymptotics of Laplace integrals

The theorem and the proof below follows that of Brietung (1994), though in a simplified form to suit our context and avoid unnecessary mathematical technicalities. Further reference can also be found in Bleistein & Handelsman (1975).

Theorem: Let $f : R^n \to R$ be a twice continuously differentiable (i.e., all partial derivatives up to second order exist and are continuous) and $g : R^n \to R$ be continuous. Define the Laplace integral (with large parameter β) as

$$J(\beta) = \int g(\underline{x}) e^{\beta^2 f(\underline{x})} d\underline{x}$$

If f has a single maximum at $\underline{x}^*$ and its Hessian $H_f(\underline{x}^*)$ is negative definite, then

$$J \sim (2\pi)^{n/2} \frac{g(\underline{x}^*) e^{\beta^2 f(\underline{x}^*)}}{\sqrt{|\det H_f(\underline{x}^*)|}} \beta^{-n}$$, as $\beta \to \infty$

Before we give the proof, we note that following:

1. The above formula is identical to (2.3) if we take $g(\underline{x}) \equiv 1$ and $f(\underline{x}) = \ln k(\underline{x}) / \beta^2$. Of course, this does not mean that (2.3) was asymptotic in any sense, because in this theorem f *should not* depend on β.

2. Intuitively, the asymptotic relationship in the theorem holds because as β increases, the term $\exp[\beta^2 f(\underline{x})]$ will be very peaked around $\underline{x}^*$, leaving other regions unimportant in accounting for the value of the integral, and making only the value of $g(\underline{x})$ at $\underline{x}^*$ relevant. The Hessian of f is involved since f is attached to β^2 and so will affect the rate at which the integral is peaked as $\beta \to \infty$ and the effective region that gives significant contribution to the integral.

Proof:
For convenience and without loss of generality, assume $\underline{x}^* = \underline{0}$ and $f(\underline{x}^*) = 0$. First,

$$\beta^n J(\beta) = \beta^n \int g(\underline{x}) e^{\beta^2 f(\underline{x})} d\underline{x}$$

Changing integration variable $\underline{x} = \underline{z}/\beta$, then $d\underline{x} = \beta^{-n} d\underline{z}$, and

$$\beta^n J(\beta) = \int_{R^n} g(\underline{z}/\beta) e^{\beta^2 f(\underline{z}/\beta)} d\underline{z} = \int_{R^n} k_\beta(\underline{z}) d\underline{z}$$

where

$$k_\beta(\underline{z}) = g(\underline{z}/\beta) e^{\beta^2 f(\underline{z}/\beta)}$$

The proof will be completed by showing that as $\beta \to \infty$,

$$\lim_{\beta\to\infty} \int k_\beta(\underline{z}) d\underline{z} = \int \lim_{\beta\to\infty} k_\beta(\underline{z}) d\underline{z} = \int g(\underline{0}) \exp\left[\frac{1}{2}\beta^2 \underline{z}^T H_f(\underline{0}) \underline{z}\right] d\underline{z} = \frac{g(\underline{0})(2\pi)^{n/2}}{\sqrt{|\det H_f(\underline{0})|}} \quad (2.4)$$

We will proceed to show the first two equalities, where the last equality follows from (2.2).

The first equality involves exchanging the order of limit and integration, which can be shown using the Lebesgue Dominated Convergence Theorem (see, e.g., Rudin 1966). It says that if a sequence of functions $\{f_n : n = 1,2,...\}$ converges to f and is bounded (in absolute value) by an integrable function g, i.e., for every n, $|f_n(\underline{x})| < g(\underline{x}) \quad \forall \underline{x} \in R^n$, then $\lim_{n\to\infty} \int f_n(\underline{x}) d\underline{x} = \int \lim_{n\to\infty} f_n(\underline{x}) d\underline{x} = \int f(\underline{x}) d\underline{x}$, i.e., order of limit and integration can be exchanged. To make use of the Dominated

Convergence Theorem, we need to show that $k_\beta(\underline{z})$ is bounded by an integrable function. This is through the following lemma whose proof is referred to Breitung (1994), p.56:

Lemma:
Let $f: R^n \to R$ be twice continuously differentiable, and

1) $f(\underline{0}) > f(\underline{x})$, $\forall \underline{x} \in R^n / \{\underline{0}\}$
2) $H_f(\underline{0})$ is negative definite

Then there exists $K > 0$ such that $\forall \underline{x} \in R^n$,

$$f(\underline{x}) \le f(\underline{0}) - K \underline{x}^T \underline{x}$$

That is, it is possible to find a concave quadratic form that bounds f.

Now, using the lemma,

$$k_\beta(\underline{z}) = g(\underline{z}/\beta) e^{\beta^2 f(\underline{z}/\beta)} \le g(\underline{z}/\beta) e^{-\beta^2 K \underline{z}^T \underline{z}}$$

and, since the function on the RHS is integrable, the Dominated Convergence Theorem holds for k_β and this validates the first equality in (2.4).

To show the second equality in (2.4), first note that

$$\lim_{\beta\to\infty} k_\beta(\underline{z}) = \lim_{\beta\to\infty} g(\underline{z}/\beta) e^{\beta^2 f(\underline{z}/\beta)} = g(\lim_{\beta\to\infty} \underline{z}/\beta) e^{\lim_{\beta\to\infty} \beta^2 f(\underline{z}/\beta)} = g(\underline{0}) e^{\lim_{\beta\to\infty} \beta^2 f(\underline{z}/\beta)}$$

since g and $\exp(\cdot)$ are continuous. Next, using L'Hospital's rule:

$$\lim_{\beta\to\infty} \beta^2 f(\underline{z}/\beta) = \lim_{\beta\to\infty} \frac{f(\underline{z}/\beta)}{\beta^{-2}} = \lim_{\beta\to\infty} \frac{\nabla f(\underline{z}/\beta)\underline{z}\ (-\beta^{-2})}{-2\beta^{-3}} = \lim_{\beta\to\infty} \frac{\nabla f(\underline{z}/\beta)\underline{z}}{2\beta^{-1}}$$

$$= \lim_{\beta\to\infty} \frac{\underline{z}^T H_f(\underline{z}/\beta)\underline{z}\ (-\beta^{-2})}{2(-\beta^{-2})} = \frac{1}{2}\underline{z}^T H_f(\lim_{\beta\to\infty} \underline{z}/\beta)\underline{z} = \frac{1}{2}\underline{z}^T H_f(\underline{0})\underline{z}$$

where the last but one equality holds because f is twice continuously differentiable. Thus,

$$\lim_{\beta\to\infty} k_\beta(\underline{z}) = g(\underline{z}) \exp\left[\frac{1}{2}\underline{z}^T H_f(\underline{0})\underline{z}\right]$$

which shows the second equality of (2.4). Thus, the proof is completed.

Note:

1. There is another case where the integration domain is a subset F of R^n and where the maximum point $\underline{x}^*$ appears at the boundary of F. This case is often encountered in reliability analysis, where $h(\underline{x}) = I(\underline{x} \in F)$ is an indicator function and $J = \int I(\underline{x} \in F) q(\underline{x}) d\underline{x} = \int_F q(\underline{x}) d\underline{x} = P(\underline{X} \in F)$ is the probability of failure. The resulting formula in this case will depend on the gradient of $\ln k$ at

$\underline{x}^*$ with respect to the normal to the boundary of F. We will present one asymptotic result in this case for q being a standard Gaussian PDF.

2. In the case of multiple design points, the asymptotic result for the Laplace integral still holds, with the asymptotic formula including contributions from each design point.

2.3. Reliability integrals (FORM/SORM)

Here, we focus our attention on the case of finding failure probabilities when q is a n-dimensional standard Gaussian joint PDF, i.e., for

$$q(\underline{x}) = \phi(\underline{x}) = (2\pi)^{-n/2}\exp\left(-\frac{1}{2}\underline{x}^T\underline{x}\right),\ h(\underline{x}) = I(\underline{x}\in F)$$

and

$$J = P(F) = \int I(\underline{x}\in F)q(\underline{x})d\underline{x} = \int_F \phi(\underline{x})d\underline{x}$$

The reason why we consider a standard Gaussian space is that, first of all, Gaussian distribution is used in many applications, and even if an uncertain parameter is not Gaussian, we may still be able to transform it to a Gaussian one (although transformation of dependent variables are technically more difficult). Secondly, and perhaps more importantly, the standard Gaussian space has some nice properties such as rotational symmetry, and it is possible to obtain some insightful results.

In the context of finding failure probabilities, we assume that the failure region F is some half space that is away from the orgin. This is usually encountered in engineering applications because the nominal state of an engineering system is situated near the origin in the standard Gaussian space and often correspond to a 'safe' rather than 'failure' state. The failure region is commonly defined through a limit state function g, in the form $F = \{\underline{x}\in R^n : g(\underline{x}) < 0\}$. Thus, the failure boundary is given by $\partial F = \{\underline{x}\in R^n : g(\underline{x}) = 0\}$.

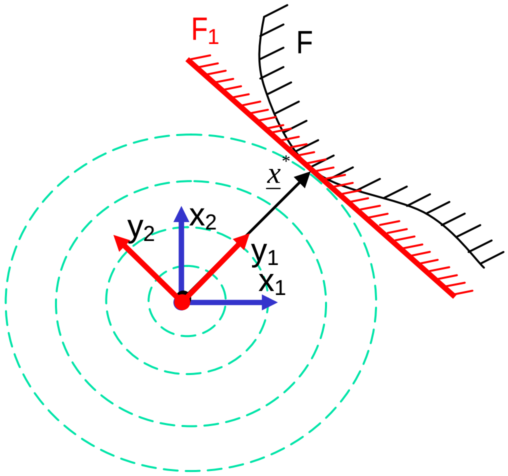

Fig. 2.1: Schematic diagram for FORM

First Order Reliability Method (FORM)

Consider the failure region F as shown in Fig.2.1. By noting that a standard Gaussian PDF decays with the distance of $\underline{x}$ from the origin, it can be reasoned that the point in F that has the highest PDF value among all points in F is the one that lies at the boundary and is closest to the origin, say, $\underline{x}^*$ in the figure. The neighborhood of $\underline{x}^*$ should give the major contribution to failure probability, and so we may approximate the failure probability by approximating the failure region in that neighborhood.

In FORM, we approximate the failure region by a linear half-space $F_1 = \{\underline{x} \in R^n : \beta^2 - \langle \underline{x}, \underline{x}^* \rangle < 0\}$, where $\beta = \| \underline{x}^* \|$ is often known as the 'Hasofer-Lind reliability index' (or 'reliability index' in short) in the structural reliability literature. The failure probability is then approximated as

$$P(F) = \int_F \phi(\underline{x}) d\underline{x} \approx \int_{F_1} \phi(\underline{x}) d\underline{x}$$

To evaluate the rightmost integral, we note that the standard Gaussian PDF is rotationally symmetric, in the sense that if we write our coordinates with respect to another orthonormal basis, i.e., from $\underline{x}$ to $\underline{y}$ in Fig.2.1, it does not change the value of the PDF nor the integral, thus,

$$P(F) \approx \int_{F_1} \phi(\underline{x}) d\underline{x} = \int_{F_1} \phi(\underline{y}) d\underline{y} = \int_{R^{n-1}} \phi(y_2, ..., y_n) dy_2 ... dy_n \int_\beta^\infty \phi(y_1) dy_1 = \Phi(-\beta)$$

where the second equality holds because, in the space of $\underline{y}$, $F_1 = \{y_1 > \beta\}$ and so does not depend on $y_2, ..., y_n$.

Second Order Reliability Method (SORM)

In SORM, we approximate the failure boundary by a hyper-paraboloid with limit state function (see Fig.2.2)

$$g_2(\underline{y}) = (\beta - y_1) - \frac{1}{2} \sum_{i=2}^n \kappa_i {y_i}^2$$

where

$$\kappa_i = -\frac{1}{\| \nabla g(\underline{x}^*) \|} {\underline{v}_i}^T H_g(\underline{x}^*) \underline{v}_i \ , \ i = 1, ..., n$$

are the principal curvatures at $\underline{x}^*$ of the failure boundary $F = \{g(\underline{x}) = 0\}$ with respect to the orthonormal tangent space of F; $\{\underline{v}_i : i = 1, ..., n\}$ are the orthormal eigenvectors of $H_g(\underline{x}^*)$ with $\underline{v}_1$ parallel to $\underline{x}^*$. Note that the definition of the curvature is such that a sphere around the origin has a positive curvature.

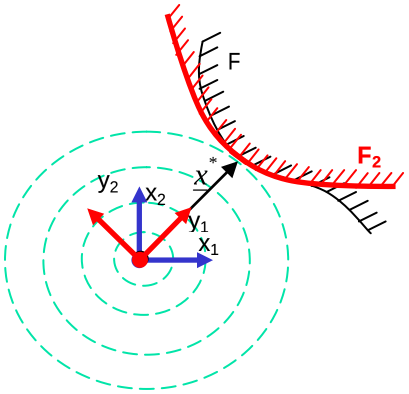

Fig. 2.2: Schematic diagram for SORM

Unfortunately, even after approximating F by $F_2 = \{g_2(\underline{x}) < 0\}$, the resulting failure probability integral still does not admit a close form solution. However, for $\beta \to \infty$, an asymptotic formula has been found (e.g., Breitung 1994):

If $\kappa_i < 1/\beta$ for $i = 2,...,n$, then

$$\int_{F_2} \phi(\underline{y}) d\underline{y} \sim \frac{\Phi(-\beta)}{\sqrt{\prod_{i=2}^{n} (1-\beta\kappa_i)}}, \text{ as } \beta \to \infty$$

The condition $\kappa_i < 1/\beta$ for $i = 2,...,n$ is to make sure that the design point $\underline{x}^*$ is indeed a point with minimum distance in its neighborhood on F.

[To better appreciate the condition $\kappa_i < 1/\beta$, it is a simple exercise to show that the curvature on any point on the spherical surface $g(\underline{x}) = \frac{1}{2}\beta^2 - \frac{1}{2}\underline{x}^T \underline{x}$ of radius β is equal to $1/\beta$. On this spherical surface we know that there is no point with minimum distance because all points are of the same distance β from the origin.]

2.4. Final Remarks

The discussions and results presented here work as long as the problem at hand is of similar context to what we discussed and the Gaussian approximation of integrand is reasonable. In applications, the major computational effort is usually spent on locating the design point, which can become prohibitive when the dimension n is large (of course, we should always investigate our problem first and see if we can get the design point analytically or see what advantage we can take based on our understanding about the problem). The potential existence of multiple design points, where usually it is difficult to find them all (if we ever bother to find them!), is another issue one needs to be mindful about. Later we will discuss a stochastic simulation method called importance sampling which may be able to correct the potential bias in the asymptotic approximation, of course by spending additional computational efforts.

Exercise

1. Let $Y=\sum_{i=1}^{n} c_i X_i$ where $\{X_i : i=1,...,n\}$ are i.i.d. standard Gaussian. Then from elementary probability we know that $\mathrm{var}[Y]=\sum_{i=1}^{n} {c_i}^2$. Suppose we don't know this and we try to approximate $\mathrm{var}[Y]$ using asymptotic approximation by viewing it as

   $$\mathrm{var}[Y]=\int k(\underline{x})d\,\underline{x}$$

   where $k(\underline{x})=(\sum_{i=1}^{n} c_i x_i)^2 \phi(\underline{x})$. Use the asymptotic approximation to estimate $\mathrm{var}[Y]$ and see if it is a good approximation.

3. Direct Monte Carlo Method

Consider

$$J = \int_{R^n} h(\underline{x})q(\underline{x})d\underline{x}$$

where now we write the integrand explicitly in terms of h and the parameter PDF q, because the method we are going to discuss explores this structure. Since q is a valid PDF, the integral can be viewed as an expectation of $h(\underline{X})$ when $\underline{X}$ is distributed as q. This implies J can be estimated as a sample average:

$$J \approx \tilde{J}_N = \frac{1}{N}\sum_{k=1}^{N} h(\underline{X}_k)$$

Where $\underline{X}_k, k = 1,...,N$, are i.i.d. samples drawn from q. This is the well-known (direct) Monte Carlo Simulation (MCS) method, which is applicable regardless of problem complexity, number of uncertain variables, etc. In the case of reliability analysis where $h(\underline{x}) = I(\underline{x} \in F)$,

$$J \approx \tilde{J}_N = \frac{1}{N}\sum_{k=1}^{N} I(\underline{X}_k \in F) = \frac{\text{no. of failed samples}}{\text{total no. of samples}}$$

which corresponds to our usual way of estimating probabilities by repeating experiments.

3.1. Statistical properties of estimators

Mean and variance

It can be easily shown that for every N

$$E[\tilde{J}_N] = J$$

and

$$\text{var}[\tilde{J}_N] = \frac{\text{var}[h]}{N} \qquad (3.1)$$

$$\sigma_{\tilde{J}_N} = \sqrt{\text{var}[\tilde{J}_N]} = \frac{\sigma_h}{\sqrt{N}}$$

where $\sigma_.$ denotes the standard deviation of the subscribed variable.

Thus, $\tilde{J}_N$ is an unbiased and convergent estimator for J. The variance of the estimator is often used for judging the efficiency of a stochastic algorithm. The $1/N$ decay of variance, or equivalently, $1/\sqrt{N}$ decay of standard deviation (which has the same unit as J), is a common phenomenon in stochastic estimation methods. It is $1/\sqrt{N}$ instead of $1/N$ (as in one-dimensional integration) due to the stochastic nature of samples. One important feature is that it does not depend on n, in contrast to $O(N^{-1/n})$ for numerical integration.

Stochastic convergence

So we know $\tilde{J}_N$ is unbiased and convergent, but does it mean that in any particular trial of run $\tilde{J}_N$ must converge to the target value if we keep on increasing N? Well, the expectation and variance do not say anything about this. For direct MCS, the answer is positive, due to a strong statement called the 'Strong Law of Large Numbers'. It says that if we take the sample average of N i.i.d. numbers with a finite variance, then with probability 1 it will converge to the expectation. 'With probability 1' here may be understood practically as 'for every trial of run'. Thus,

$$P(\lim_{N\to\infty} \tilde{J}_N = J) = 1$$

Note:

1. In probability theory where the argument inside $P(\cdot)$ is a set, the statement '$\lim_{N\to\infty} \tilde{J}_N = J$' needs to defined more technically. We omit this technicality here.

2. The Strong Law of Large Numbers refers to a convergence called 'convergence with probability 1' or 'almost sure convergence'. There are other kinds of convergence, such as 'mean-square convergence', 'convergence in probability', 'convergence in distribution', that differ in their strength of assertion. For example, if it converges with probability 1, then it can be shown to converge in probability, but the reverse is not true. These different notions of convergence are invented due to the mathematical difficulty in showing convergence in different problems. E.g., in some problems it is difficult to show convergence in probability 1 but it can be easy to show convergence in mean-square. For further details, one may refer to texts on probability, e.g., Papoulis (1996), Billingsley (1995).

Central Limit Theorem

It is generally difficult to obtain the distribution of $\tilde{J}_N$ for every N, but the Central Limit Theorem says that when N is 'large' (theoretically infinite), then $\tilde{J}_N$ is asymptotically Gaussian, i.e.,

$$\lim_{N\to\infty} P\left(\frac{\tilde{J}_N - J}{\sigma_{\tilde{J}_N}} < a \right) = \Phi(a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-z^2/2} dz$$

for every a. How 'large' an N is large? It depends on the distribution of $h(\underline{X})$. The closer the $h(\underline{X})$ is to a Gaussian distribution, the smaller the N to be considered large. In particular, if $h(\underline{X})$ has a Gaussian distribution, then $\tilde{J}_N$ is Gaussian for every N. In statistics, $N \geq 30$ is considered large, but bear in mind that it is just a rule of thumb.

3.2. Problem with rare event simulation

For estimating failure probabilities $h(\underline{x}) = I(\underline{x} \in F)$, and so in (3.1)

$$\text{var}[h] = E[I(\underline{X} \in F)^2] - E[I(\underline{X} \in F)]^2 = P(F) - P(F)^2$$

$$\text{var}[\tilde{J}_N] = \frac{\text{var}[h]}{N} = \frac{1}{N}[P(F) - P(F)^2]$$

Thus, for small failure probabilities (our usual interest), $\text{var}[\tilde{J}_N] \sim P(F)/N$, i.e., $\sigma_{\tilde{J}_N} \sim \sqrt{P(F)/N}$, which looks good since the smaller the $P(F)$, the smaller the $\sigma_{\tilde{J}_N}$. But, let's look closer. Suppose the failure event corresponds to $P(F) = 10^{-2}$ and we use $N = 100$, then $\sigma_{\tilde{J}_N} = 1$ - are we satisfied with this error? Probably not, since the error is so large compared to the estimate. Thus, it is often more useful to judge based on how big the error is relative to the target estimate, i.e., the 'coefficient of variation' (c.o.v.) defined as the ratio of $\sigma_{\tilde{J}_N}$ to $E[\tilde{J}_N]$:

$$\delta_{\tilde{J}_N} \sim \frac{\sigma_{\tilde{J}_N}}{E[\tilde{J}_N]} = \sqrt{\frac{1 - P(F)}{P(F)N}}$$

For small $P(F)$, $\delta_{\tilde{J}_N} \sim 1/\sqrt{P(F)N}$ so we immediate see that as having smaller $P(F)$ is not a good thing! It is due to this observation that direct MCS is commonly recognized as not efficient for estimating small failure probabilities, or more generally, not efficient for investigating rare events (e.g., tail of distributions). This can be understood intuitively, since when the failure probability is small, most of the samples do not fail, and with only a small number of failure samples (if any) it is difficult to get information about the likelihood of failure.

As a rule of thumb, to estimate a failure probability $P(F)$ with a coefficient of 30%, one requires on average $N = 10/P(F)$ samples, or '10 failure samples'.

Facing with the problem of direct MCS with rare events, methodologists have been trying to come up with more advanced stochastic simulation methods that can investigate rare events more efficiently, by commonly trying to generate more failure samples that hopefully will yield more information about failure and hence providing a better estimate. How is it possible? E.g., by finding more pertinent information about the particular system we are dealing with and use such information to modify our stochastic sampling and estimation procedure, or even devising a method such that the information is found during the simulation process and used as soon as it is found. Yet we will find that as we try to beat direct MCS on the grounds of efficiency (smaller c.o.v.), we are bound to lose out on robustness (i.e., being applicable and efficiency for different kinds of problems) – this may be a law of the game.

3.3. Reduction of dimension

Let the set of uncertain parameters consist of two (sub-) set of uncertain parameters $\underline{Y}$ and $\underline{Z}$, i.e., $\underline{X} = [\underline{Y}, \underline{Z}]$. Suppose we have an efficient method (e.g., by analytical means) for computing the conditional expectation of $h(\underline{X}) = h(\underline{Y}, \underline{Z})$ given $\underline{Y}$, i.e., for every $\underline{y}$, we can compute easily

$$C(\underline{y}) = E[h(\underline{Y}, \underline{Z}) \mid \underline{Y} = \underline{y}] = \int h(\underline{y}, \underline{z}) q(\underline{z} \mid \underline{y}) d\underline{z}$$

where $q(\underline{z} \mid \underline{y})$ is the conditional PDF of $\underline{Z}$ given $\underline{Y}$. Note that

$$E[h(\underline{Y}, \underline{Z})] = E[E[h(\underline{Y}, \underline{Z}) \mid \underline{Y}]] = E[C(\underline{Y})]$$

where the inner expectation is taken over uncertainty in $\underline{Z}$ for given $\underline{Y}$ and the outer expectation is taken over uncertainty in $\underline{Y}$. This suggests that we can estimate $E[h(\underline{Y}, \underline{Z})]$ by averaging $C(\underline{Y})$ as

$$E[h(\underline{Y}, \underline{Z})] \approx \tilde{J}_N' = \frac{1}{N} \sum_{k=1}^{N} C(\underline{Y}_k)$$

where $\underline{Y}_k, k = 1, \ldots, N$ are i.i.d. drawn from $q(\underline{y})$ (the marginal PDF of $\underline{Y}$). Of course, we can also do direct MCS as

$$E[h(\underline{Y}, \underline{Z})] \approx \tilde{J}_N = \frac{1}{N} \sum_{k=1}^{N} h(\underline{Y}_k, \underline{Z}_k)$$

where $\underline{Y}_k, \underline{Z}_k, k = 1, \ldots, N$ are i.i.d. drawn from $q(\underline{y}, \underline{z})$ (the joint PDF of $\underline{Y}$ and $\underline{Z}$). Is it always more efficient (in the sense of c.o.v.) to use $\tilde{J}_N'$ rather than $\tilde{J}_N$? The answer is positive, as it can be shown that

$$\mathrm{var}[\tilde{J}_N'] \le \mathrm{var}[\tilde{J}_N]$$

where the equality holds if and only if $\mathrm{var}[h(\underline{Y}, \underline{Z}) \mid \underline{Y} = \underline{y}] = 0$ for every $\underline{y}$.

[Exercise to show that if $\underline{Z}$ is a dummy, i.e., h does not depend on $\underline{Z}$, then $\mathrm{var}[h(\underline{Y}, \underline{Z}) \mid \underline{Y} = \underline{y}] = 0$.]

The proof is a direct consequence of the 'conditional variance formula', which says that for any uncertain variable A and any proposition B,

$$\mathrm{var}[A] = \mathrm{var}[E[A \mid B]] + E[\mathrm{var}[A \mid B]]$$

The proof of the conditional variance formula is left as an exercise. Hint: Start with the definition of $\mathrm{var}[A \mid B]$ and take expectation. A proof for random variables can also be found in Ross (1971).

Applying the conditional variance formula with $A = h(\underline{Y}, \underline{Z})$ and $B = \underline{Y}$, we have

$$\mathrm{var}[h(\underline{Y}, \underline{Z})] = \mathrm{var}[E[h(\underline{Y}, \underline{Z}) \mid \underline{Y}]] + E[\mathrm{var}[h(\underline{Y}, \underline{Z}) \mid \underline{Y}]] = \mathrm{var}[C(\underline{Y})] + E[\mathrm{var}[h(\underline{Y}, \underline{Z}) \mid \underline{Y}]]$$

i.e.,

$$\mathrm{var}[\tilde{J}_N] = \mathrm{var}[\tilde{J}_N'] + E[\mathrm{var}[h(\underline{Y}, \underline{Z}) \mid \underline{Y}]] \ge \mathrm{var}[\tilde{J}_N']$$

since $\mathrm{var}[h(\underline{Y}, \underline{Z}) \mid \underline{Y}] \ge 0$.

This section is included here to remind that if we can integrate out certain uncertain variables in the problem, it is worth doing because it will surely reduce the estimation variance and hence computational effort.

4. Importance Sampling Method

Consider

$$J = \int_{R^n} h(\underline{x}) q(\underline{x}) d\underline{x}$$

where $h(\underline{x}) \geq 0$. The basic idea of importance sampling is to generate samples that lie more frequently in the 'important region' that gives major contribution to the integral, hopefully to reduce the variance of the estimator. In the context of reliability analysis where $h(\underline{x}) = I(\underline{x} \in F)$, this is equivalent to generate more samples in the failure region that has high PDF value. Of course, when we estimate using samples that are not generated from q, we cannot use the same estimator as in direct MCS and it needs to be modified to account for this fact.

Suppose we are able to come up with a PDF $f(\underline{x})$ (called 'importance sampling density', ISD) such that the random samples drawn from it will lie frequently in the important region, and such that

1. we can compute the value of $f(\underline{x})$ easily for any given $\underline{x}$

2. we can generate random samples from f easily

3. the support of f (i.e., the region in R^n over which $f > 0$) covers that of $I(\underline{x} \in F) q(\underline{x})$

Then, we can view the integral as

$$J = \int h(\underline{x}) q(\underline{x}) d\underline{x} = \int \frac{h(\underline{x}) q(\underline{x})}{f(\underline{x})} f(\underline{x}) d\underline{x} = E_f[R(\underline{X})]$$

where $R(\underline{x}) = h(\underline{x}) q(\underline{x}) / f(\underline{x})$ is called the 'importance sampling quotient'. The subscript in the expectation is to denote that $\underline{X}$ is now distributed as f rather than q. The above equation suggests that we can estimate J by

$$J \approx \tilde{J}_N^{IS} = \frac{1}{N} \sum_{k=1}^{N} R(\underline{X}_k)$$

where $\underline{X}_k : k = 1,...,N$ are i.i.d. as f. Since $\tilde{J}_N^{IS}$ is an average of i.i.d. samples of R, it has all the statistical properties as in direct MCS, i.e., it is unbiased, its variance is given by $\mathrm{var}_f[R]/N$, and it converges to J with probability 1. This, of course, assumes that f is appropriately chosen such that R has finite variance under f.

The most important task in applying importance sampling is the construction of the ISD f. If it is 'good' (we will discuss shortly what it means), then we can get tremendous improvement in efficiency compared to direct MCS. Otherwise, at the other extreme, we may even end up with a heavily biased estimate (and it is worst if we do not even know the estimate is biased).

The pre-requisite requirements of 1 and 2 for f often means that practically we need to construct f using conventional known distributions (e.g., Gaussian, exponential, etc., or a weighted sum of them). Thus, the usual procedure is to assume certain form of PDF (using known distributions) for f and then choose the parameters based on information about the important region so that the samples drawn from f can be expected to lie frequently in the important region.

4.1. Optimal Importance Sampling Density

The (theoretical) optimal ISD corresponds to the one that leads to the least variance of $\tilde{J}_N^{IS}$ (and hence of R). It turns out that the optimal choice of the ISD can be written explicitly and the least variance we can get is zero. The optimal ISD is given by

$$f_o(\underline{x}) = \frac{h(\underline{x})q(\underline{x})}{\int h(\underline{z})q(\underline{z})d\underline{z}}$$

i.e., it is just proportional to the integrand. The integral in the denominator serves as a normalizing constant so that f_o integrates to 1 (to be a valid PDF). So, with the optimal ISD, are we done?

Of course not, since in practice this choice of ISD is not feasible, due to the following two reasons:

1. We cannot evaluate f_o easily, since the denominator is an n-dimensional integral. In fact, had we know how to evaluate this integral efficiently, we don't need to do importance sampling, since this integral is just the answer we want!

2. We do not have an efficient method for generating random samples according to f_o (not to mention that we only know f_o up to a scaling constant, due to point 1).

Nevertheless, the result about the optimal ISD indicates that we should choose our ISD to have a shape as close to the integrand as possible, under our pre-requisite constraints on f.

For reliability problems with $h(\underline{x}) = I(\underline{x} \in F)$, the optimal ISD is just the conditional PDF (given F):

$$f_o(\underline{x}) = \frac{I(\underline{x} \in F)q(\underline{x})}{P(F)} = q(\underline{x} \mid F)$$

Again, we do not know how to evaluate f_o efficiently since we do not know $P(F)$ and we do not have an efficient method for generating samples according to the conditional PDF (acceptance-rejection method can be applied but is far from being efficient for small $P(F)$). But this suggests the capability of generating conditional samples play an important role in reliability analysis. Later we will discuss a powerful method called Markov Chain Monte Carlo (MCMC) that shows great promise for generating conditional samples, utilizing which we are able to come up with a method called Subset Simulation that also shows great promise for rare event simulation.

Next, we present one inequality that relates the estimation variance to the relative entropy between the optimal ISD and the actual ISD we use.

4.2. Variance of estimator and relative entropy

Definition: The relative entropy of a PDF p_2 (relative) to p_1 is defined as

$$H(p_2, p_1) = \int p_2(\underline{x}) \ln \frac{p_2(\underline{x})}{p_1(\underline{x})}\, d\underline{x}$$

The relative entropy is commonly used as a measure for the difference between two PDFs. It is non-negative, as shown by using Jensen's inequality and noting that $\ln(\cdot)$ is concave:

$$-H(p_2, p_1) = \int p_2(\underline{x}) \ln \frac{p_1(\underline{x})}{p_2(\underline{x})}\, d\underline{x} \le \ln \int p_2(\underline{x}) \frac{p_1(\underline{x})}{p_2(\underline{x})}\, d\underline{x} = \ln 1 = 0$$

Note, however, that $H(\cdot,\cdot)$ is not a valid metric, because it is not symmetric, i.e., $H(p_2, p_1) \ne H(p_1, p_2)$. The relative entropy $H(p_2, p_1)$ can be viewed as the amount of information gained when we update (hypothetically) our uncertainty about $\underline{X}$ from p_2 (prior) to p_1 (posterior). That is, suppose originally we think that $\underline{X}$ is distributed as p_1, but then after some process of gaining information we find out that it is actually distributed as p_2. Then $H(p_2, p_1)$ represents the amount of information we gained from such process.

Back to importance sampling, note that the c.o.v. of the importance sampling estimator J_N^{IS} can be written as

$$\delta_{IS} = \frac{\Delta_{IS}}{\sqrt{N}}$$

where ${\Delta_{IS}}^2 = \mathrm{var}_f[R]/E_f[R]^2 = \mathrm{var}_f[R]/J^2$ is called the 'unit c.o.v.', i.e., the c.o.v. when $N = 1$. Then it can be shown that

$${\Delta_{IS}}^2 + 1 \ge e^{H(f_o, f)}$$

Proof:
For simplicity we will omit dependence on $\underline{x}$ in integrals or on $\underline{X}$ in expectations.

$$\Delta_{IS}{}^2 = \frac{1}{J^2}\mathrm{var}_f[\frac{h\,q}{f}] = \frac{1}{J^2}\left\{E_f[\frac{h^2q^2}{f^2}] - E_f[\frac{h\,q}{f}]^2\right\} = \frac{1}{J^2}\left\{E_f[\frac{h^2q^2}{f^2}] - J^2\right\}$$

so

$$\Delta_{IS}{}^2 + 1 = \frac{1}{J^2}\int \frac{h^2q^2}{f^2} f\, d\underline{x} = \int \frac{h\,q}{J}\times\frac{1}{f}\times\frac{h\,q}{J} d\underline{x} = \int \frac{f_o}{f} f_o d\underline{x} = E_{f_o}[\frac{f_o}{f}]$$

By writing $E_{f_o}[f_o / f] = E_{f_o}[\exp(\ln f_o / f)]$ and using Jensen's inequality, noting that $\exp(\cdot)$ is convex, we get our desired result:

$$\Delta_{IS}{}^2 + 1 \ge e^{E_{f_o}[\ln(f_o / f)]} = e^{H(f_o, f)}$$

The inequality relating the unit c.o.v. and the relative entropy has important use for investigating whether importance sampling is still applicable in high dimension n for reliability problems. Essentially, as n increases, it can happen that the relative entropy between q (to which $f_o = q(\underline{x}\,|\,F)$ is proportional when restricted to F) and f grows without bound when the form of the latter is not chosen appropriately. Consequently, the importance sampling quotient R degenerates into a zero-infinity law that has infinite variance as $n \to \infty$.

4.3. Importance Sampling using design points

Refer to Au, Papadimitriou & Beck (1999).

4.4. High-dimensional problems

Refer to Au (2001) or Au & Beck (2003).

4.5. Importance Sampling for linear systems

Refer to Au (2001) or Au & Beck (2001).

5. Markov Chain Monte Carlo Method

Suppose we want to generate random samples from some distribution f_π, but

1. We can only evaluate f_π up to a scaling constant. E.g., we know f_π is proportional to some non-negative function f (for which we can calculate its value) but f does not integrate to 1. Of course, $f_\pi(\underline{x})$ can be written as

   $$f_\pi(\underline{x}) = \frac{f(\underline{x})}{\int f(\underline{z})d\underline{z}}$$

   but still it is not easy to evaluate the normalizing constant in the denominator since it is a multi-dimensional integral.

2. f_π does not correspond to any 'known' distribution and so we do not know how to efficiently generate samples from it.

This situation is frequently encountered in two important class of problems, namely, Bayesian updating and reliability/probabilistic failure analysis, as illustrated in the following two examples.

Example 5.1. *Bayesian model updating*
Suppose we have a model that makes prediction of a response $\{y_i(\underline{X}) : i = 1,...,m\}$ for a given set of uncertain parameters $\underline{X} \in R^n$ and we also have the corresponding actual data $D = \{\hat{y}_1,..., \hat{y}_m\}$. Assume the prediction model

$$\hat{y}_i = y_i(\underline{X}) + \varepsilon_i \qquad (5.1)$$

where $\{\varepsilon_i : i = 1,...,m\}$ are i.i.d. standard Gaussian (in the actual applications the variance of ε_i should also be an uncertain parameter to be updated, but for the sake of illustration let's fix it to 1 here). We want to update our probability model about $\underline{X}$.

Let M denote the proposition containing prediction error model (5.1) and the prior PDF for $\underline{X}$, which is assumed to be standard Gaussian centered at $\underline{x}_0$. Then the posterior PDF of $\underline{X}$ given the data D and M is given by, using Bayes' Theorem,

$$p(\underline{x} | D, M) \quad = \frac{p(D | \underline{x}, M)p(\underline{x} | M)}{P(D | M)}$$

$$= \frac{1}{P(D | M)} \times (2\pi)^{-m/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{m}\left(\hat{y}_i - y_i(\underline{x})\right)^2 \right] \times (2\pi)^{-n/2} \exp\left[-\frac{1}{2} \| \underline{x} - \underline{x}_0 \|^2 \right]$$

Note that we generally do not know the value of the normalizing constant $P(D | M)$ before we have solved the updating problem. Thus, in this example, we can evaluate efficiently $f(\underline{x}) = p(D | \underline{x}, M)p(\underline{x} | M)$ but not the normalizing constant $P(D \mid M)$, which is equal to the integral of f (since $p(\underline{x} \mid D, M)$ must integrate to 1).

On the other hand, although we know the form of f, it is not trivial to generate samples of $\underline{X}$ according to it because the relationship between $\underline{x}$ and $y_i(\underline{x})$ is often only known implicitly and could be quite complicated, rendering f a complicated function of $\underline{x}$.

Example 5.2. *Conditional distribution given failure*
Probabilistic failure analysis is concerned with the likely scenarios that may occur when systems with uncertainties fail. Mathematically it corresponds to finding expectation of quantities of interest given that failure occurs, which necessitates the efficient generation of samples conditional on failure (or 'conditional samples' in short). On the other hand, in reliability analysis, the capability of efficiently computing failure probabilities is intimately related to the capability of efficiently generating samples conditional on failure. The conditional PDF is given by

$$q(\underline{x} \mid F) = \frac{I(\underline{x} \in F) q(\underline{x})}{P(F)}$$

Here, we can evaluate $f(\underline{x}) = I(\underline{x} \in F) q(\underline{x})$ efficiently for a given $\underline{x}$, but not the normalizing constant $P(F)$ (which in fact is the answer we need!). Regarding the generation of samples, although we can generate samples easily from $q(\underline{x})$ when it is chosen from some known class of distributions, the same is not true for $q(\underline{x} \mid F)$, due to the conditioning on F. An acceptance-rejection algorithm can be applied but is far from being efficient, especially when $P(F)$ is small, because on average it requires $1/P(F)$ samples drawn from q to get one sample lie in F.

We will discuss a class of methods called Markov Chain Monte Carlo (MCMC) method that allows us to generate samples according to the target PDF under the context we mentioned. One important message brought from this method is that, although it is difficult to efficiently generate independent samples according to the target PDF, it is possible, using a specially designed Markov Chain, to efficiently generate *dependent* samples that are at least asymptotically distributed as the target PDF (as the number of Markov steps increases).

We will start with the simplest and original form of MCMC, called Metropolis algorithm.

5.1. Metropolis Algorithm

In this algorithm, we need to choose a 'proposal PDF' $p^*(\underline{x} \mid \underline{y})$ such that

1. For a given $\underline{y}$, $p^*(\cdot \mid \underline{y})$ is a valid PDF
2. p^* is symmetric, i.e., $p^*(\underline{x} \mid \underline{y}) = p^*(\underline{y} \mid \underline{x}) \quad \forall \underline{x}, \underline{y} \in R^n$

3. we can efficiently generate samples from $p^*(\cdot\,|\,\underline{y})$ for every $\underline{y}$

We are going to describe the algorithm for generating samples $\underline{X}_k : k=1,2,...$ that will be seen forming a first order Markov Chain. To start the chain, first choose $\underline{X}_1$ by 'some means' (e.g, from some PDF that is close to the target PDF, or simply a fixed vector; we will discuss this later). Then, for $k=1,2,...$, to generate the next sample $\underline{X}_{k+1}$ from the current sample $\underline{X}_k$,

Step 1. Generate a 'candidate' $\underline{\tilde{X}}$ from $p^*(\cdot\,|\,\underline{X}_k)$

Step 2. Calculate $r = \dfrac{f(\underline{\tilde{X}})}{f(\underline{X}_k)}$, then

Set $\underline{X}_{k+1} = \begin{cases} \underline{\tilde{X}} & \text{with probability } \min\{1,r\} \\ \underline{X}_{k+1} & \text{with probablity } 1-\min\{1,r\} \end{cases}$

Note:

1. the algorithm only involves ratio of the target PDF, and so does not require information about the normalizing constant

2. In computer code, Step 2 can be implemented simply as 1) generate U from a uniform PDF on [0,1]; 2) if $U<r$, set $\underline{X}_{k+1} = \underline{\tilde{X}}$, otherwise set $\underline{X}_{k+1} = \underline{X}_k$

3. It is possible that the next sample is equal to the current sample, in the case when $\underline{\tilde{X}}$ is rejected.

4. The proposal PDF p^* governs the distribution of the candidate and affects the transition of the chain from one state to another. A Gaussian PDF or a uniform PDF centered at the current sample is a common choice. The spread of $p^*(\underline{x}\,|\,\underline{y})$ around $\underline{y}$ affects how fast the chain can explore the parameter space and the dependence among samples:
    a. If the spread is too small, the acceptance rate is high, but the next sample (which is often taking the candidate) will be near the current one, increasing dependence
    b. If the spread is too large, the acceptance rate can be low, making the next sample identical to the current one, increasing correlation

    Thus, the choice of the spread of p^* should play a balance between acceptance and spatial correlation. It is commonly suggested to choose the spread of p^* to be at least a large as the spread of f_π.

5. A Markov chain is a sequence of random variables $\underline{X}_1, \underline{X}_2...$ that satisfies

$$p(\underline{X}_{k+1}=\underline{x}_{k+1}\,|\,\underline{X}_k=\underline{x}_k, \underline{X}_{k-1}=\underline{x}_{k-1},...,\underline{X}_1=\underline{x}_1) = p(\underline{X}_{k+1}=\underline{x}_{k+1}\,|\,\underline{X}_k=\underline{x}_k)$$

i.e., given the present state, the future state is independent of the past states. See Ross (19??) or Doobs (19??) for Markov chain theory.

5.1.1. Properties of the Metropolis chain

For the Metropolis algorithm, Markov Chain Theory says that if the Markov chain $\{\underline{X}_k : k = 1,2,...\}$ is 'ergodic' (we will discuss this later), then

1. If $\underline{X}_1$ is distributed as the target PDF f_π, then so are $\underline{X}_2, \underline{X}_3,...$

2. Even if $\underline{X}_1$ is not distributed as f_π, the samples will still be asymptotically distributed as f_π, in the sense that
   $$\lim_{n\to\infty} p_{\underline{X}_N}(\underline{x}) = f_\pi(\underline{x})$$
   Here, f_π is called the limiting stationary PDF for the Markov chain.

Stationary chain & detailed balance

We will prove the first claim, which is easier and will illustrate one important mechanism of MCMC. The second claim involves issues about ergodicity, which will be discussed from an implementation point of view (there are of course theoretical discussions available, e.g., Roberts & Casella, 1999).

So, suppose $\underline{X}_k \sim f_\pi$ ('~' as a shorthand for 'is distributed as'), we will show that $\underline{X}_{k+1} \sim f_\pi$ as well. The proof is a direct consequence of a condition that the Metropolis chain satisfies, called 'detailed balance':

$$p_{X_{k+1}|X_k}(\underline{x}\,|\,\underline{y}) f_\pi(\underline{y}) = p_{X_{k+1}|X_k}(\underline{y}\,|\,\underline{x}) f_\pi(\underline{x}),\ \forall \underline{x}, \underline{y} \in R^n$$

where $p_{X_{k+1}|X_k}(\underline{x}\,|\,\underline{y})$ denote the 'transition PDF' of the Markov chain that gives the conditional PDF for $\underline{X}_{k+1}$ at $\underline{x}$ given that $\underline{X}_k = \underline{y}$. When the detailed balance holds, f_π is called the 'stationary distribution' of the Markov chain.

To shown the detailed balance, consider the transition PDF $p_{X_{k+1}|X_k}(\underline{x}\,|\,\underline{y})$ for $\underline{x} \neq \underline{y}$. According to the Metropolis algorithm, for $\underline{x} \neq \underline{y}$, $p_{X_{k+1}|X_k}(\underline{x}\,|\,\underline{y})$ is proportional to $p^*(\underline{x}\,|\,\underline{y})$ and $r = \min\{1, f(\underline{x})/f(\underline{y})\}$, and so

$$p_{X_{k+1}|X_k}(\underline{x}\,|\,\underline{y}) = c\ p^*(\underline{x}\,|\,\underline{y}) \min\{1, \frac{f(\underline{x})}{f(\underline{y})}\}$$

where c is a normalizing constant. Multiplying both sides by $f_\pi = f / \int f\, d\underline{z}$,

$$p_{X_{k+1}|X_k}(\underline{x}\,|\,\underline{y}) f_\pi(\underline{y}) = c\ p^*(\underline{x}\,|\,\underline{y}) \min\{1, \frac{f(\underline{x})}{f(\underline{y})}\} \frac{f(\underline{y})}{\int f\, d\underline{z}} \qquad (5.2)$$

To proceed, we note that for any $a, b > 0$, $\min\{1, \frac{a}{b}\} b \equiv \min\{1, \frac{b}{a}\} a$ (this identity is where the magic lies), which means that on the RHS of (5.2),

$$\min\{1,\frac{f(\underline{x})}{f(\underline{y})}\}f(\underline{y})=\min\{1,\frac{f(\underline{y})}{f(\underline{x})}\}f(\underline{x})$$

and so, together with the symmetry $p^*(\underline{x}|\underline{y})=p^*(\underline{y}|\underline{x})$, detailed balance follows:

$$p_{X_{k+1}|X_k}(\underline{x}|\underline{y})f_\pi(\underline{y})=c\;p^*(\underline{y}|\underline{x})\min\{1,\frac{f(\underline{y})}{f(\underline{x})}\}\frac{f(\underline{x})}{\int f\,d\underline{z}}=p_{X_{k+1}|X_k}(\underline{y}|\underline{x})f_\pi(\underline{x})$$

for $\underline{x}\neq\underline{y}$. For the case when $\underline{x}=\underline{y}$, detailed balance holds trivially, and therefore the Metropolis chain satisfies detailed balance.

We are now ready to obtain the PDF for $\underline{X}_{k+1}$. Using the theorem of total probability,

$$\begin{aligned}
p_{\underline{X}_{k+1}}(\underline{x}) &= \int p_{\underline{X}_{k+1}|\underline{X}_k}(\underline{x}|\underline{y})p_{\underline{X}_k}(\underline{y})d\underline{y}\\
&= \int p_{\underline{X}_{k+1}|\underline{X}_k}(\underline{x}|\underline{y})f_\pi(\underline{y})d\underline{y} \text{ since } \underline{X}_k\sim f_\pi \text{ for stationary chain}\\
&= \int p_{\underline{X}_{k+1}|\underline{X}_k}(\underline{y}|\underline{x})f_\pi(\underline{x})d\underline{y} \text{ due to detailed balance}\\
&= f_\pi(\underline{x})\int p_{\underline{X}_{k+1}|\underline{X}_k}(\underline{y}|\underline{x})d\underline{y}\\
&= f_\pi(\underline{x})\text{, since } \int p_{\underline{X}_{k+1}|\underline{X}_k}(\underline{y}|\underline{x})d\underline{y}=1
\end{aligned}$$

and so $\underline{X}_{k+1}\sim f_\pi$ when $\underline{X}_k\sim f_\pi$.

Note:

1. Detailed balance says that, in a stationary state (i.e., $\underline{X}_k\sim f_\pi$), the transition rate from $\underline{x}$ to $\underline{y}$ is the same as the transition rate from $\underline{y}$ to $\underline{x}$ for all $\underline{x}$ and $\underline{y}$ as the Metropolis chain steps forward.

2. If we step backward along the Metropolis chain, i.e., following the sequence $\{\underline{X}_N,\underline{X}_{N-1},\underline{X}_{N-2},...,\underline{X}_1\}$, then the transition probability is $p_{X_k|X_{k+1}}(\underline{y}|\underline{x})$. This backward transition probability is quite non-trivial to get from first principle according to the algorithm. However, if the detailed balance holds and the Metropolis chain is in a stationary state, using Bayes' Theorem,

   $$p_{X_k|X_{k+1}}(\underline{y}|\underline{x})=\frac{p_{X_{k+1}|X_k}(\underline{x}|\underline{y})p_{X_k}(\underline{y})}{p_{X_{k+1}}(\underline{x})}=\frac{p_{X_{k+1}|X_k}(\underline{x}|\underline{y})f_\pi(\underline{y})}{f_\pi(\underline{x})}=\frac{p_{X_{k+1}|X_k}(\underline{y}|\underline{x})f_\pi(\underline{x})}{f_\pi(\underline{x})}$$

   i.e., $p_{X_k|X_{k+1}}(\underline{y}|\underline{x})=p_{X_{k+1}|X_k}(\underline{y}|\underline{x})$

   which means that the transition PDF of the backward chain is identical to that of the forward chain. For this reason, detailed balance is also known as the 'reversibility condition', because in this case under a stationary state the probabilistic property of the forward chain is identical to the backward chain, and such chain is called 'reversible'.

Transient chain & Ergodicity

We mentioned that, in the case when the chain is not started with $\underline{X}_1 \sim f_\pi$, the samlples will still be asymptotically distributed as f_π, provided the Metropolis chain is 'ergodic' (in fact, even when $\underline{X}_1 \sim f_\pi$, ergodicity is still needed for correct statistical estimation using a *single chain*, see later). In the study of stochastic processes, ergodicity deals with the question of whether a sample average (i.e., averaging along a single ensemble) will tend to the ensemble average (i.e., averaging across different ensembles, or trials of run) as the number of samples increases. Ergodicity for Markov chains means that starting from any state, the chain will be able to visit the neighborhood of the starting state again infinitely many times (recurrence property) as the number of Markov steps $N \to \infty$ and there is a non-zero probability that it can visit any given state within a finite number of Markov steps (irreducibility property). Practically, ergodicity is concerned with whether the generated samples can populate sufficiently the regions in the parameter space over which f_π has significant probability content.

Example 5.3. *Reducible chain*

Suppose the support of f_π consists of two disconnected regions, say, D1 and D2; and the support of $p^*(\underline{x} \mid \underline{y})$ around $\underline{y}$ is small compared to the separation distance between D1 and D2. Then, when the chain is started from, say, D1, it is impossible to transit to a state in D2, since this necessitates a candidate $\underline{\tilde{X}}$ generated in D2, which is impossible. In this case, the chain will 'get stuck' in D1, and so the distribution of the samples can only represent f_π conditional on D1. In fact, the chain started in D1 will (at best) have a limiting stationary PDF equal to $f_\pi(\underline{x} \mid D_1)$. This corresponds to a 'reducible' situation, which is often the major concern when dealing with ergodicity (recurrence is often not a problem). Consequently, the chain is not ergodic and statistical estimation based on the samples along a single chain will be biased, because the information in D2 is not contained in the samples.

In this example, even if $\underline{X}_1 \sim f_\pi$ (that will mean that $\underline{X}_1$ sometimes is generated in D1 and sometimes in D2), the chain is not ergodic, either. This can be reasoned intuitively because the chain can only develop in the region where it is initiated. But why were we able to show that if $\underline{X}_k \sim f_\pi$ then so is $\underline{X}_{k+1} \sim f_\pi$? Anything wrong in the proof? In this case, we can still check to see that both the detailed balance and $p_{\underline{X}_{k+1}}(\underline{x}) = f_\pi(\underline{x})$ are still valid (an exercise to check). And this *does not contradict* with the fact that the chain is *not ergodic*. It is because $p_{\underline{X}_{k+1}}(\underline{x}) = f_\pi(\underline{x})$ only means that $\underline{X}_{k+1}$ is distributed as f_π in an 'ensemble sense'. To understand this, suppose we generate $\underline{X}_1 \sim f_\pi$, and then we generate $\underline{X}_2$ from $\underline{X}_1$ using the Metropolis algorithm. If we repeat different trials of this process (i.e., different ensembles), we get different samples of $\underline{X}_2$, say, $\underline{X}_2^{(1)}$,

$\underline{X}_2^{(2)}$, … where the index in the superscript denotes the trial run number. The samples $\{\underline{X}_2^{(1)}, \underline{X}_2^{(2)}, \underline{X}_2^{(3)}, \ldots\}$ will have a histogram that approximate f_π, as can be guaranteed by a proof in the stationary case. In contrast, if we follow a particular chain (ensemble), say, the first chain, and observe $\{\underline{X}_1^{(1)}, \underline{X}_2^{(1)}, \underline{X}_3^{(1)}, \ldots\}$, then they will only have a histogram that is confined to either D1 or D2, depending on whether $\underline{X}_1^{(1)}$ is in D1 or D2, respectively. This highlights that the detailed balance does not guarantee ergodicity even when the chain is stationary, because ergodicity is concerned about the behavior along any given chain, whereas detailed balance and $p_{\underline{X}_{k+1}}(\underline{x})$ only involve ensemble average concept. The theoretical treatment of ergodicity can be quite involved, although the problem can often be solved by proper choice of proposal PDF (in not-so-tough problems).

5.1.2. Short historical notes

The Metropolis algorithm presented here was due to the celebrated paper: Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953), "Equation of state calculation by fast computing machine". Journal of Chemical Physics, 21:1087-91. The original algorithm was for discrete state-space, where Metropolis et. al. dealt with calculating the properties of chemical substances based on statistical mechanics. There were of course many follow-up papers, but Hastings (1970, working on Bayesian statistics) later generalized the Metropolis algorithm to allow the proposal PDF to be non-symmetric, leading to the so called 'Metropolis-Hastings algorithm': Hastings, W.K. (1970), "Monte Carlo sampling methods using Markov chains and their applications". Biometrika, 57:97-109.

There have been many variants of the algorithm with applications to different disciplines (e.g., image processing, biostatistics, Bayesian statistics, econometrics). Their common feature is that statistical averaging is done over Markov chain samples with the limiting stationary distribution equal to the specified target one. These algorithms are collectively known as Markov Chain Monte Carlo (MCMC) method.

5.2. Metropolis-Hastings Algorithm

The Metropolis-Hastings (M-H) algorithm allows the use of a non-symmetric proposal PDF, and is as follows:

To generate $\underline{X}_{k+1}$ from $\underline{X}_k$:

Step 1. Generate a candidate sample $\underline{\tilde{X}}$ from $p^*(\cdot \mid \underline{X}_k)$

Step 2. Calculate $r = \dfrac{f(\underline{\tilde{X}})}{f(\underline{X}_k)} \times \dfrac{p^*(\underline{X}_k \mid \underline{\tilde{X}})}{p^*(\underline{\tilde{X}} \mid \underline{X}_k)}$, then

Set $\underline{X}_{k+1} = \begin{cases} \underline{\tilde{X}} & \text{with probability} \min\{1, r\} \\ \underline{X}_{k+1} & \text{with probablity} \, 1-\min\{1, r\} \end{cases}$

The proof of detailed balance for the Metropolis-Hastings chain is left as an exercise.

Note:

1. The only modification is in the calculation of r which now also involves the ratio of the proposal PDF.

2. When the proposal PDF is symmetric, the Metropolis-Hastings algorithm reduces to the Metropolis algorithm.

3. If $p^*(\underline{x}|\underline{y})$ does not depend on $\underline{y}$, then the chain is not 'adaptive'. In this case, the choice of p^* is similar to the choice of importance sampling density, although it is generally more difficult to have a good choice compared to a symmetric p^*. [Exercise to show that if $p^*(\underline{x}|\underline{y}) = f_\pi(\underline{x})$, then $r \equiv 1$, the acceptance rate is 100% and all samples $\underline{X}_k$ are independent. But is it feasible to use such p^*?]

5.3. Statistical Estimation

The samples $\{\underline{X}_1, ..., \underline{X}_N\}$ generated according to the M-H algorithm can be used for statistical averaging similar to the case of direct Monte Carlo:

$$S = \int h(\underline{x}) f_\pi(\underline{x}) d\underline{x} \approx S_N = \frac{1}{N}\sum_{k=1}^{N} h(\underline{X}_k)$$

5.3.1. Statistical properties of estimator

In the following discussion, we assume that the chain is ergodic and $E_{f_\pi}[h^2] < \infty$.

Expectation

1. If $\underline{X}_1 \sim f_\pi$, i.e., the chain is stationary, then S_N is unbiased for every N [Exercise to show]

2. If $\underline{X}_1$ is not distributed as f_π, then S_N is biased for every N, but is asymptotically unbiased as $N \to \infty$.

   Proof:

   $$E[S_N] = \frac{1}{N}\sum_{k=1}^{N} E[h(\underline{X}_k)]$$

We know that $E[h(\underline{X}_N)] \to S$ because $\underline{X}_N$ is asymptotically distributed as f_π. But we need to show that the same is also true for S_N. The proof is a direct application of the following proposition:

Proposition: If the sequence $\{a_k \in R\}$ converges to $a \in R$ and $b_k = \frac{1}{k}\sum_{i=1}^{k} a_i$, then $b_k \to a$. The term b_k is often called a Cesàro average (Billingsley, 1995, Section A30). The proof is left as an exercise. Hint: Use the fundamental definition of limit.

Variance

If $\underline{X}_1 \sim f_\pi$, i.e., the chain is stationary, then

$$\mathrm{var}[S_N] = \frac{\mathrm{var}_{f_\pi}[h]}{N}(1+\gamma)$$

where $\gamma = 2\sum_{k=1}^{N-1}(1-\frac{k}{N})\rho(k)$ is a correlation factor,

and $\rho(k) = \dfrac{\mathrm{cov}(h(X_1), h(X_{1+k}))}{\mathrm{var}_{f_\pi}[h]}$ is the correlation coefficient of h at k samples apart.

Proof:

$$\mathrm{var}[S_N] = E[(S_N - S)^2] = \frac{1}{N^2}\sum_{i,j=1}^{N} E[(h(\underline{X}_i) - S)(h(\underline{X}_j) - S)] = \frac{1}{N^2}\sum_{i,j=1}^{N}\mathrm{cov}(h(\underline{X}_i), h(\underline{X}_j))$$

For the last summand, instead of summing with respect to i, j, we can sum along its diagonal, i.e., we sum for $i = j, i = j+1, \dots, i = j+k, \dots, i = j+(N-1)$, and for each k, i goes from 1 to $N-k$:

$$\mathrm{var}[S_N] = \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{var}_{f_\pi}[h] + \frac{1}{N^2}\sum_{k=1}^{N-1}\sum_{i=1}^{N-k}\mathrm{cov}(h(\underline{X}_i), h(\underline{X}_{i+k}))$$

Since the chain is stationary, $\mathrm{cov}(h(\underline{X}_i), h(\underline{X}_{i+k})) = \mathrm{cov}(h(\underline{X}_1), h(\underline{X}_{1+k}))$, for $i = 1, \dots, N-k$, and so

$$\mathrm{var}[S_N] = \frac{\mathrm{var}_{f_\pi}[h]}{N} + \frac{1}{N^2}\sum_{k=1}^{N-1}(N-k)\,\mathrm{cov}(h(\underline{X}_1), h(\underline{X}_{1+k}))$$

and the proof is completed after algebra.

Note: A sufficient condition for $\mathrm{var}[S_N] \to 0$ as $N \to \infty$ is $\mathrm{cov}(h(\underline{X}_1), h(\underline{X}_{1+N}) \to 0$ as $N \to \infty$[Exercise to prove. Hint: relate γ / N to a Cesàro average].

5.4. High-dimensional problems

The original M-H algorithm may have 100% in high-dimensional problems, especially in reliability applications, but which can be solved by using a component-updating algorithm. See Au (2001), Au & Beck (2001).

6. Subset Simulation

6.1. Raw idea
6.2. Algorithm
6.3. Statistical properties of estimator

Refer to Au (2001), Au & Beck (2001), Au & Beck (2003).

7. First Passage Probability and Rice's outcrossing theory

7.1. First Passage Probability
7.2. Up-crossing rate
7.3. Rice's formula
7.4. Poisson approximation
7.5. Random vibration of linear systems

References

Au, S. K., Papadimitriou, C. and Beck, J. L. (1999), “Reliability of uncertain dynamical systems with multiple design points”, *Structural Safety*, 21:113-133.

Au, S. K. (2001). *On the solution of First Excursion Problems by Efficient Simulation with Applications to probabilistic seismic performance assessment*. PhD Thesis in Civil Engineering, EERL Report No. 2001-02, California Institute of Technology.

Au, S. K. and Beck, J. L. (2001). “First excursion probabilities of linear dynamical systems by very efficient importance sampling”, *Probabilistic Engineering Mechanics*, 16: 193-207.

Au, S. K. and Beck, J. L. (2003). “Importance sampling in high dimensions”, *Structural Safety*, 25:139-163.

Bleistein, N. and Handelsman, R. A. (1986). Asymptotic Expansions of Integrals. Dover Publications.

Breitung, K. W. (1991). Asymptotic Approximations for Probability Integrals. Lecture Notes in Mathematics, Springer-Verlag.

Doobs

Rudin, W. (1966). Real and Complex Analysis, McGraw-Hill.

Billingsley (1995). Probability and Measure. John Wiley.

Roberts, C. P. and Casella, G. (1999). Monte Carlo Statistical Methods. Springer Texts in Statistics, Springer.

Ross, S. (19??). Stochastic Processes. Wiley.