1

Introduction

Modern engineering has seen a booming demand for analyses of complex systems to unprecedented detail, paralleled with an increasing reliance on numerical models for performance predictions. Systems are designed with an increasing expectation of high performance reliability and robustness in functionality. Assessing the effects of uncertainties and their mitigation in the design decision-making process allows one to make risk-informed decisions even in a state of uncertainty. Uncertainties in engineering may arise from incomplete knowledge about the modeling of system behavior, model parameter values, measurement, environmental loading conditions, and so on. Probability theory allows a rational framework for plausible reasoning and decision-making in the presence of uncertainties. The analysis of the effects of uncertainty includes, but is by no means limited to, the following objectives:

1. Reliability (or risk) analysis – to assess the likelihood of violating specified system performance criteria. It involves assessing the probability distribution or performance margins of some critical system response. This can be used for examining whether the system is likely to pass specified performance criteria in the presence of modeled uncertainties.
2. Failure analysis – to assess the characteristics of failure scenarios, for example, the likely cause and consequence of failure. The former provides insights about system failures and helps devise effective measures for their mitigation. The latter reveals the likely scenarios when failure occurs and provides information for loss estimation, devising contingency measures, or trading-off cost–benefits in design.

Models for complex systems are characterized by a large number of governing state variables, time-varying and response-dependent nonlinear behavior. They are also increasingly governed by multi-physics laws. Although the advent of computer technology has allowed the analysis of complex systems for a given scenario to be performed with affordable computational time, the same is not true for analyzing the effects of uncertainty, since the latter involves information from multiple scenarios and hence repeated system analyses. Even if resources are available, they should be deployed in an effective manner that yields information on failure scenarios of concern with a consistent weight on their likelihood. This motivates the development of efficient yet robust computational algorithms for propagating uncertainties in complex systems.

Engineering Risk Assessment with Subset Simulation, First Edition. Siu-Kui Au and Yu Wang.

This book is primarily concerned with performing risk and failure analysis by means of an advanced Monte Carlo method called "Subset Simulation." The method is based on the simple idea that a small failure probability can be expressed as the product of a number of not-so-small conditional failure probabilities. This idea has led to algorithms that generate random samples gradually propagating towards the failure region in the uncertain parameter space. The samples provide information for estimating the whole distribution of the critical response quantity that governs failure, covering large (central) to small (tail) probability regimes. The method has been found to be efficient for investigating rare failure events, but still retains some robustness to problem complexity in different applications. It treats the system as a black box and hence does not explore any prior information one may have regarding the system behavior, which can possibly be incorporated into the solution process. Thus, for a particular application, it may not be the most efficient method. However, since it can be applied without much knowledge about the system (like Direct Monte Carlo) it may still be a competitive algorithm when robustness is taken into consideration. The possibility of using the generated samples for investigating failure scenarios also makes the method versatile for risk and failure analysis.

1.1 Formulation

Despite the wide variety of problems encountered in engineering applications, a failure event can often be represented as the exceedance of a critical scalar response variable Y over a specified threshold b. The response variable Y is assumed to be completely determined by a set of "input variables" $\mathbf{X} = [X_1, \dots, X_n]$. The relationship is generically represented as

$$Y = h(\mathbf{X}) \tag{1.1}$$

where $h : \mathcal{R}^n \mapsto \mathcal{R}$ is a known deterministic function that represents the computational process, for example, the analytical formula, empirical formula, finite element model, computational dynamics, and so on. Clearly, when $\mathbf{X}$ is uncertain, so is Y. Using a probabilistic approach, $X_1, \dots, X_n$ are modeled as random variables with prescribed joint probability distribution assigned based on the analyst's knowledge. Induced by the probabilistic modeling on $\mathbf{X}$, Y is also a random variable. However, its probability distribution is not arbitrary and is not up to the analyst to decide. Rather, it is completely determined by the probability distribution of $\mathbf{X}$ and the function h. This is depicted in Figure 1.1.

In order to make decisions related to Y, which is nevertheless uncertain, one needs to have information about its probability distribution. This is generally unknown, however. It must be determined in accordance with the function h and the probability distribution of $\mathbf{X}$. The effort required depends largely on which part of the distribution of Y is relevant. Statistical quantities related to the "frequent" or central part of the distribution, such as the mean or variance, are often easier to obtain than those related to the "rare" or tail part of the distribution, such as the exceedance probability $P(Y > b)$ when b is large. The latter is the primary interest in this book.

If we denote the failure event as $F = \{Y > b\}$, then we can write

$$\text{Failure probability} = P(F) = P(Y > b) \tag{1.2}$$

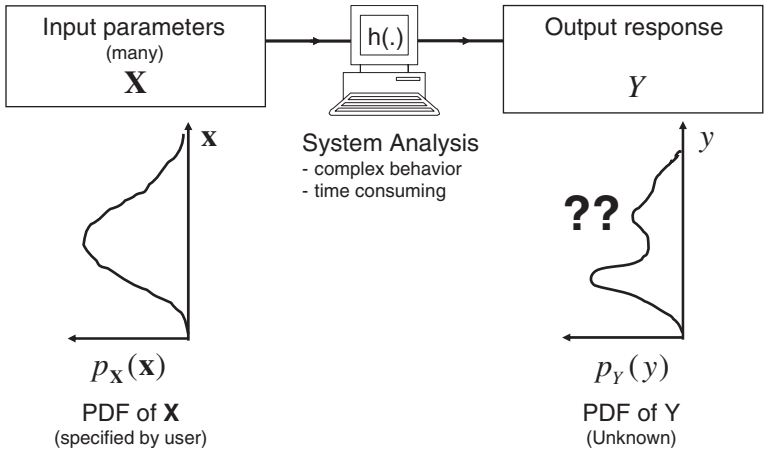

Figure 1.1 Input–output context.

Complementary to the failure probability is the "reliability":

$$\text{Reliability} = 1 - P(F) = P(Y \leq b) = 1 - P(Y > b) \tag{1.3}$$

Evaluating the failure probability and conditional expectation for failure analysis requires information about the system when failure occurs. Properly designed engineering systems are intended to have high reliability (close to 1) and hence small failure probability (close to zero). Target failure probabilities often needed to be estimated are in the order of $10^{-3} \sim 10^{-6}$, which nevertheless depends on the class of applications.

For complex problems, the relationship between $\mathbf{X}$ and Y is analytically intractable and is often only known implicitly. That is, the value of Y for a given $\mathbf{X}$ can be calculated but no other information (e.g., derivative) is available. The relationship is also difficult to visualize when $\mathbf{X}$ contains a large number of uncertain variables. Analytical or closed-form solutions for the required statistics of Y are rarely available.

The Direct Monte Carlo method provides a robust means for estimating the statistics by averaging over pseudo-random samples generated according to the distribution of $\mathbf{X}$. It has become increasingly popular due to the advent of modern computer technology. When the statistics are related to the tail of the distribution of Y, however, it is not efficient because most of the samples lie in the frequent region. Only those lying at the tail of the distribution of Y provide useful information for estimating the tail statistics, but their occurrence is rare.

The failure probability can be mathematically formulated in several ways that lead to different strategies for its computation. Without loss of generality (see Section 1.2), assume that $\mathbf{X} = [X_1, \ldots, X_n]$ is a set of continuous-valued random variables with probability density function (PDF) $q(\mathbf{x})$. The failure probability can be formulated as a "probability integral":

$$P(Y > b) = \int_F q(\mathbf{x}) d\mathbf{x} \tag{1.4}$$

where

$$F = \{\mathbf{x} : h(\mathbf{x}) > b\} \tag{1.5}$$

denotes the "failure region," that is, a subset in the parameter space of $\mathbf{X}$ that corresponds to failure. The failure probability can thus be viewed as a sum of the probability content within the failure region. Alternatively, the integral can be written as being over the whole parameter space:

$$P(Y > b) = \int I_F(\mathbf{x}) q(\mathbf{x}) d\mathbf{x} \tag{1.6}$$

where

$$I_F(\mathbf{x}) = \begin{cases} 1 & \text{if } h(\mathbf{x}) > b \\ 0 & \text{if } h(\mathbf{x}) \leq b \end{cases} \tag{1.7}$$

is the "indicator function" that reveals whether $\mathbf{x}$ lies in the failure region or not. This form is often used for mathematical derivations. Another useful perspective is via the expectation:

$$P(Y > b) = E[I_F(\mathbf{X})] \tag{1.8}$$

where $E[\cdot]$ denotes the mathematical expectation when $\mathbf{X}$ is distributed as q. This leads to the idea of "statistical averaging" and hence Monte Carlo simulation.

Viewing $P(Y > b)$ as a function of b, finding the failure probability is equivalent to finding the "complementary cumulative distribution function" (CCDF) of Y (CCDF = 1 – CDF), especially at the tail where small failure probabilities are the main interest. Of course, finding the whole CDF is much more difficult, or at least computationally more expensive, than finding just the failure probability at a single threshold level. Nevertheless, estimating small failure probabilities is intimately related to estimating the upper tail of the CCDF.

Example 1.1 Definition of response variable

Many system failure events can be expressed in terms of the union or intersection of exceedance events, say, corresponding to system components connected (logically) in series or in parallel. A failure event of this kind can be expressed in terms of the exceedance of a scalar response Y. Clearly Y should be defined such that $P(Y > b)$ corresponds to the failure probability of interest. It is also preferable to define Y in a non-dimensional manner.

Suppose $F = \{C < D\}$, where C and D are the "capacity" and "demand" of a system that can possibly depend on $\mathbf{X}$. Then Y may be defined in a dimensionless manner as $Y = D/C$ so that $P(F) = P(Y > 1)$.

Suppose now $F = \bigcap_{i=1}^{n_1} \{C_i < D_i\}$, where C_i and D_i $(i = 1, \ldots, n_1)$ can possibly depend on $\mathbf{X}$. This can be interpreted as the failure of a system of components connected in parallel where the system fails only when all the components have failed. In this case the critical response Y may be defined as $Y = \min_{i=1}^{n_1} D_i/C_i$ so that $P(F) = P(Y > 1)$.

On the other hand, if $F = \bigcup_{i=1}^{n_2} \{C_i < D_i\}$, then it can be interpreted as the failure of a system of components connected in series where the system fails if any one of the components fails. In this case Y may be defined as $Y = \max_{i=1}^{n_2} D_i/C_i$ so that $P(F) = P(Y > 1)$.

In general, if F is defined via $\cap$ and/or $\cup$, then Y can be defined using "min" and/or "max" appearing in the same order. For example, if $F = \bigcap_{i=1}^{n_1} \bigcup_{j=1}^{n_2} \{C_{ij} < D_{ij}\}$ then we can define $Y = \min_{i=1}^{n_1} \max_{j=1}^{n_2} D_{ij}/C_{ij}$ so that $P(F) = P(Y > 1)$. ■

1.2 Context

Unless otherwise stated, the problems that we deal with in this book have the following context:

1. The input random variables $X_1, \ldots, X_n$ are continuous-valued.
2. The input random variables $X_1, \ldots, X_n$ are mutually independent.
3. The (one-dimensional) PDF of each X_i, denoted by $q_i(x)$ corresponds to some known "standard distribution" (e.g., Gaussian, exponential) so that
   (a) the value of $q_i(x)$ can be evaluated efficiently for any given x
   (b) random samples distributed as q_i can be generated efficiently.
4. The relationship between $\mathbf{X}$ and Y is not explicitly known. That is, we can evaluate the value of $Y = h(\mathbf{x})$ for a given $\mathbf{x}$ but generally we are not able to obtain other information such as gradient or Hessian. The latter quantities if needed have to be computed numerically, for example, using finite difference.
5. The computational effort for evaluating $h(\mathbf{x})$ for a given $\mathbf{x}$ is significant. The total computational effort is dominated by the number of function evaluations of $h(\mathbf{x})$.
6. Interest is focused on small failure probabilities or, equivalently, the tail of the CCDF of $Y = h(\mathbf{X})$.
7. The number of random variables in $\mathbf{X}$ can be very large (possibly infinite).

Some comments are in order regarding the above context. Assumption 1 on continuous random variables is introduced primarily for the sake of discussion and elegance in the theory (e.g., integrals instead of sums). It does not introduce much loss of generality in practice, because discrete-valued random variables can be generated by a mapping of continuous-valued random variables. Assumption 2 on mutual independence of input random variables does not generate any loss of generality because, in reality, dependent variables are generated by independent ones. Assumption 3 on standard distributions distinguishes the problems discussed in this book from Bayesian inference problems, in which case the posterior distribution of random variables given data often do not correspond to any standard distribution (see Section 1.4).

1.3 Extreme Value Theory

As mentioned in the beginning of this chapter, Subset Simulation treats the input–output relationship of a system as a black box and so it (often) need not be the most efficient procedure for a particular application. When there is some knowledge about the relationship

between $\mathbf{X}$ and Y it may be possible to take advantage of it to derive useful statements about the distribution of Y. One classical example with profound results is when Y is defined as the maximum over a large number (theoretically infinite) of i.i.d. (independent and identically distributed) random variables in $\mathbf{X}$. This has been studied extensively, leading to "extreme value theory" (Gumbel, 1958; Galambos, 1978; David, 1981). When the problem context fits and the asymptotic distribution of the extreme exists, it is usually more efficient to apply the theory to determine the failure probability $P(Y > b)$. In this case the main task is to identify the type of limiting distribution and then to determine the distribution parameters accordingly. Standard statistical tools are available (Coles, 2001). Although one can still apply Subset Simulation to solve the same problem, it is less efficient because it does not take advantage of the special mathematical structure of the problem.

1.4 Exclusion

This book does not deal with the case when the distribution of $\mathbf{X}$ arises from Bayesian inference problems, which is nevertheless a very important problem with wide application (Cox, 1961; Jaynes, 2003). In this area, the interest is to determine the distribution of $\mathbf{X}$ and update response predictions based on some observed data D. According to Bayes' Theorem, the "posterior distribution" (i.e., given data) of $\mathbf{X}$ that incorporates the information from the data D is given by

$$p(\mathbf{x} \mid D) = p(D)^{-1} p(D \mid \mathbf{x}) p(\mathbf{x}) \tag{1.9}$$

The RHS of this equation should be viewed as a probability distribution of $\mathbf{X}$. The first term $p(D)^{-1}$ does not depend on $\mathbf{x}$ and so, as far as the distribution of $\mathbf{X}$ is concerned, it can be ignored. The middle term $p(D \mid \mathbf{x})$ is called the "likelihood function," which must be formulated based on modeling assumptions relating the observed data to $\mathbf{X}$ in a probabilistic manner. The last term $p(\mathbf{x})$ is called the "prior distribution" and it reflects one's knowledge about $\mathbf{X}$ in the absence of data.

Estimating the posterior statistics of $\mathbf{X}$ or updating system response prediction by means of Monte Carlo simulation requires efficient generation of samples according to the posterior distribution $p(\mathbf{x} \mid D)$. This is generally a highly non-trivial task, however. Although the prior distribution $p(\mathbf{x})$ is often chosen to follow a standard distribution (like those considered in Chapter 3), the resulting posterior distribution does not necessarily follow a standard distribution because the likelihood function $p(D \mid \mathbf{x})$ arises from system modeling and is problem-dependent. In many applications the likelihood function is only known implicitly and its dependence on $\mathbf{x}$ is rather complicated.

Conjugate prior distribution is one branch of research that examines the type of prior distribution that should be assumed for some type of likelihood function so that the resulting posterior distribution is also of a standard distribution. The use of conjugate prior distribution is convenient when applicable, but otherwise it limits the type of problem that can be solved. It has become less popular in modern applications due to the advent of computer technology and the development of advanced simulation methods that can efficiently handle arbitrary distributions. The "Markov chain Monte Carlo method" (MCMC) is one popular class of

methods that has been found useful. This method is discussed in Chapter 4 as it is used for generating failure samples in Subset Simulation.

1.5 Organization of this Book

This book is organized into seven chapters. After the introduction (this chapter), Chapter 2 gives an overview of relevant ideas that lead logically to Subset Simulation. These ideas differ in the way they view the failure probability and the way they gather and use information to account for the main contribution to the failure probability. Chapter 3 gives a basic introduction to the digital simulation of random samples according to standard distributions (e.g., Normal, Lognormal, exponential), which is indispensible for uncertainty modeling and performing Monte Carlo simulation. Chapter 4 gives a basic introduction to "Markov Chain Monte Carlo" (MCMC), which is a powerful method for generating random samples according to an arbitrarily given probability distribution. MCMC is not involved in uncertainty modeling in the context of this book, as the uncertain parameters are assumed to have standard distributions. Rather, it is involved in the efficient generation of failure samples in Subset Simulation, which is a highly non-trivial problem. Chapter 4 provides the necessary background where no pre-requisite in Markov Chain theory is needed.

Chapter 5 gives a comprehensive coverage of Subset Simulation for estimating failure probabilities through the CCDF of the critical response governing failure. It covers the basic algorithm, error estimation, choice of parameters, theoretical properties of estimators, and potential problems. Chapter 6 introduces the investigation of failure scenarios using the failure samples in Direct Monte Carlo and Subset Simulation. Chapter 7 presents an Excel spreadsheet package for performing risk assessment by Direct Monte Carlo and Subset Simulation. It contains step-by-step procedures that allow the reader to gain hands-on experience with Monte Carlo simulation. This will hopefully help the reader develop a correct perspective for interpreting and using simulation results. Mathematical tools are contained in the Appendix for reference.

1.6 Remarks on the Use of Risk Analysis

Reliability analysis or probabilistic failure analysis, or any kind of analysis in general, does not itself prevent failure from happening or provide warranty over losses. Nor does it necessarily provide information close to reality, because the underlying assumptions need not do so. These issues should not undermine the value of risk analysis because it is not meant to do so. Risk analysis is only meant to provide the decision-maker with information regarding the effects of uncertainty on the attributes that may affect a decision. The decision-maker is still required to make his or her own judgment on the use of the results. It is just a scientific way of producing relevant information consistent with the assumptions adopted regarding the modeling of uncertainty and system behavior. Having advanced computational tools hopefully allows one to focus on the problem itself, especially the decision-making part. Making assumptions is inevitable and this should be kept in mind. In many cases, an order of magnitude answer on the probability suffices for making decisions, which may also be consistent with the variability of such an answer in view of the assumptions made. Making

assumptions and placing the right confidence into the results is a human art. Practically, it is better to be "approximately right" rather than "precisely wrong."

1.7 Conventions

Before we leave this chapter, we cover some notations and conventions used in this book. We use $f(x)$ to denote a function of the argument x. When this may be confused with the value of a function at a specific x, we use f or $f(\cdot)$ to denote the function. The notation $f : A \to B$ is used to denote a function that takes an element in the set A to give a value in the set B. For example, $f : \mathcal{R}^n \to \mathcal{R}$ denotes a real scalar valued multi-variable function on the n-dimensional Euclidean space.

We reserve $P(\cdot)$ for the probability of the statement in the argument. The notation $p_X(x)$ refers to the PDF of the random variable X evaluated at the value x. When the random variable X is understood in the context it may be omitted for simplicity. Random variables are usually denoted in capital letters and their parameter value in small letters. For example, X is the random variable and $\{X = x\}$ is the event that it is equal to the given parameter value x. Vector-valued quantities are often denoted in bold, for example, $\mathbf{X} = [X_1, \ldots, X_n]$ is a vector of random variables. When the limits of summation or domain of integration are understood, they may be omitted for simplicity. An integral sign without the domain indicated is over the whole parameter space on which the integrand is defined. A sequence of quantities may be denoted in an abbreviated manner in curly braces with a running index. For example, $\{X_1, \ldots, X_N\}$ may be written as $\{X_k : k = 1, \ldots, N\}$ or abbreviated as $\{X_k\}$ when the limits the index runs through are clear. The terms "Gaussian distribution" and "Normal distribution" refer to the same distribution and are used interchangeably. Other notations and abbreviations are contained in the Nomenclature.

References

Coles, S. (2001) *An Introduction to Statistical Modeling of Extreme Values*, Springer-Verlag, Singapore.
Cox, R.T. (1961) *The Algebra of Probable Inference*, Johns Hopkins Press, Baltimore.
David, H.A. (1981) *Order Statistics*, John Wiley & Sons, Inc., New York.
Galambos, J. (1978) *The Asymptotic Theory of Extreme Order Statistics*, John Wiley & Sons, Inc., New York.
Gumbel, E.J. (1958) *Statistics of Extremes*, Columbia University Press, New York.
Jaynes, E.T. (2003) *Probability Theory: The Logic of Science*, Cambridge University Press, UK.