WOODHEAD PUBLISHING
IN MECHANICAL ENGINEERING

Estimation of Rare Event Probabilities in Complex Aerospace and Other Systems

A Practical Approach

Jérôme Morio and Mathieu Balesdent

WP
WOODHEAD
PUBLISHING

Estimation of Rare Event Probabilities in Complex Aerospace and Other Systems

A Practical Approach

Related title

Rare Event Simulation Using Monte Carlo Methods
(ISBN 978-0-47077-269-0)

Woodhead Publishing in Mechanical Engineering: Number 720

Estimation of Rare Event Probabilities in Complex Aerospace and Other Systems

A Practical Approach

Jérôme Morio and Mathieu Balesdent

AMSTERDAM • BOSTON • CAMBRIDGE • HEIDELBERG
LONDON • NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
Woodhead Publishing is an imprint of Elsevier

Woodhead Publishing Limited is an imprint of Elsevier
80 High Street, Sawston, Cambridge, CB22 3HJ, UK
225 Wyman Street, Waltham, MA 02451, USA
Langford Lane, Kidlington, OX5 1GB, UK

Notices
Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

ISBN: 978-0-08-100091-5 (print)
ISBN: 978-0-08-100111-0 (online)

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

Library of Congress Number: 2015942649

To my wife Anne, J.M.
To my father, M.B.

This page intentionally left blank

Contents

Preface

This book is an opportunity to share our practical experience on rare event probability estimation. We tried to write the book that we would have appreciated having when we started working in this research domain several years ago. The book gives a broad view of current research on rare event probability estimation, and we hope that it will satisfy the readers.

We thank the contributors to this book, namely M. Brevault, Dr. De Visscher, M. Dolado-Perez, Dr. Duponcheel, Dr. Jacquemart, M. Lacaze, Prof. Le Gland, Dr. Missoum, Dr. Pastel, Dr. Vergé, and Prof. Winckelmans, for their helpful collaboration and for the time they devoted to the project. We also thank Prof. Raphael T. Haftka who has done us the great honor of writing a foreword to this book. The works and the daily interactions with the current and former PhD students whom we have supervized at ONERA have also an important part in this book.

This book would not exist without the confidence of Elsevier-Woodhead Publishing, especially Dr. Glyn Jones and Ms. Harriet Clayton. We would also like to thank the direction of ONERA-The French Aerospace Lab, especially Prof. Philippe Bidaud and Dr. Thérèse Donath, for providing us the opportunity to write this book. We also thank M. Sébastien Aubry and M. Florent Muller for their support in the challenging and long-term objective of writing a scientific book. We also thank all our colleagues at ONERA.

We could not conclude this section without thanking our families for always being there for us and being incredibly supportive.

Jérôme Morio and Mathieu Balesdent

Jérôme Morio has been working at ONERA-The French Aerospace Lab since 2007 as research engineer in the System Design and Performance Evaluation Department. He obtained a Ph.D. in physics from the University Aix-Marseille Paul Cézanne (France) in 2007 and defended his habilitation to supervise research in 2013. He is also a lecturer in probability and statistics at ISAE and ENAC. His main research interests include rare event probability estimation, sensitivity analysis, and uncertainty management.

Mathieu Balesdent has been working at ONERA-The French Aerospace Lab since 2011 as research engineer in the System Design and Performance Evaluation

Department. He obtained a Ph.D. in mechanical engineering from the “Ecole Centrale de Nantes” (France) in 2011. His main research interests include rare event probability estimation, reliability based and multidisciplinary optimization, and aerospace vehicle design. He is also a regular referee for several international conferences and journals such as the American Institute of Aeronautics and Astronautics, Springer and Elsevier.

Foreword

It is a great pleasure to be invited to write the foreword to Morio and Balesdent's book *Estimation of Rare Event Probabilities in Complex Aerospace* (*and other*) *Systems—A Practical Approach*, because it is a very timely and needed book and it is done well.

Wikipedia defines rare events as events that occur with low frequency. It says that the term is conventionally applied for those events that have potentially widespread impact and which might destabilize society. Rare events encompass natural phenomena (major earthquakes, tsunamis, hurricanes, floods, asteroid impacts, solar flares, etc.). In aerospace engineering, the term is applied also to less catastrophic events that may happen at low probabilities, typically less than 10^{-6}. A typical example in this book is the probability that two aircraft will get dangerously close to each other in a given airspace.

Calculating accurately the probability of a rare event is usually a challenge, both in terms of the required data and the computational effort required to translate the data into a probability estimate. There are two important reasons why estimating such probabilities has become a hot topic in the past decade or so. First, safety has become much more important to the public, and so the demands from automotive, civil, and aerospace designers are much tougher than they used to be. The rising level of safety means that the causes for failure have become rarer. As one Boeing engineer told me a few years ago, "It used to be that airplane accidents were due to one unlikely thing gone wrong. Now they mostly happen due to two or three unlikely things going wrong at the same time."

The second reason is our increasing ability to estimate well the probabilities of rare events. Big data developments mean that we more often have the required information. Better education of engineers in statistics means that our engineering workforce is capable of applying properly the sophisticated statistical methods required for that purpose. Faster computers allow us to estimate accurately the probabilities of rarer and rarer events.

However, because we usually strive to calculate the probabilities of rarer events than the combination of computer power and present algorithms permit, there has been a strong burst of efficient algorithm development for that purpose in the past 20 years. This book is a welcome treatise on most of the currently available methods and algorithms.

The book is certainly comprehensive. Even though I have worked in this field for more than a decade, I have come across many useful techniques that I did not know about. This appears to be partly due to the contribution of several other contributors, but the two main authors integrated the contributions very well. So the book retains a unity of notation and style.

The book has the right balance of mathematical rigor and practical implementation of the different techniques. Its "French connection" may be partly responsible for the former, and the practical experience of the authors at ONERA, the French Aerospace Lab, for the latter.

On the one hand, the book has a set of toy problems to which each technique is applied, but then it has aerospace applications that show how many of the techniques are applied to real important engineering problems.

I am thus looking forward to using the book in the near future. On the one hand, it will be a valuable resource for my research group. On the other hand, I will be using it in courses where I teach uncertainty quantification and optimization under uncertainty. I may even be tempted to offer a new course on probability estimation of rare events.

Prof. Raphael T. Haftka
Distinguished Professor of Mechanical and
Aerospace Engineering at the University of Florida

Biography of the external contributors to this book

Loïc Brevault received both his engineering degree from "Arts et Métiers" ParisTech (France) and an M.Sc. in Aerospace at Georgia Institute of Technology (USA) in 2013. He is currently preparing a Ph.D. thesis on multidisciplinary design optimization under uncertainty with application for launch vehicle design. His thesis is funded by ONERA and CNES under the supervision of Dr. Le Riche, Dr. Balesdent, M. Bérend, and M. Oswald. His fields of research are multidisciplinary optimization under uncertainty, aerospace vehicle design, and rare event probability estimation.

Ivan De Visscher holds a mechanical engineering degree and a Ph.D. degree from the "Université Catholique de Louvain" (UCL in Belgium). He is one of the developers of the DVM/PVM/WAKE4D wake vortex simulation platform and the author of various papers on wake vortex behavior simulation and modeling.

Juan Carlos Dolado Pérez is an expert on space surveillance and tracking topics at CNES (French Space Agency). Since 2008 he has worked at the orbit determination office where his main research topics concern long- and middle-term re-entry prediction, long-term evolution of the space debris population, orbital collision risk assessment, and orbit determination from radar and optical measurements. He holds a B.Sc. in Aerospace Engineering from Madrid's Polytechnic University and an M.Sc. in Aerospace Engineering from the "Institut supérieur de l'aéronautique et de l'espace" (France).

Matthieu Duponcheel holds a mechanical engineer degree and a Ph.D. degree from UCL (Belgium) and is a research engineer at UCL. He has worked on several research projects devoted to aircraft wake vortices and participated in the development of the DVM/PVM/WAKE4D software.

Damien Jacquemart passed the agregation of mathematics in 2010 and graduated from Aix-Marseille University (France) with an M.Sc. in probability and statistics. In 2010, he joined ONERA to work on a Ph.D. in rare event probability on time-variant systems modeled with Markov processes. He defended his Ph.D. thesis the University of Rennes 1 (France) in 2014.

Sylvain Lacaze defended his Ph.D. in mechanical engineering at the University of Arizona (USA) in 2015. His current research focuses on fields such as reliability

assessment, reliability-based design optimization, meta-modeling, and adaptive sampling strategies.

François Le Gland has been a research director at INRIA Rennes (France) and the current head of the research unit applications of interacting particle systems to statistics (ASPI). His main research interests include particle filtering and rare event estimation.

Raphael T. Haftka has been a distinguished Professor of Mechanical and Aerospace Engineering at the University of Florida. Before coming to the University of Florida in 1995, he taught at the Israel Institute of Technology, the Illinois institute of Technology, and Virginia Tech. His areas of research include structural and multidisciplinary optimization, design under uncertainty, and surrogate-based global optimization.

Samy Missoum has been an Associate Professor in the Aerospace and Mechanical Engineering Department at the University of Arizona (USA) since 2005. He obtained his Ph.D. in Mechanical Engineering from the Institute of Applied Sciences in Toulouse (France) and was a postdoctoral researcher in the Aerospace and Ocean Engineering Department at Virginia Polytechnic Institute and State University (USA). His current research focuses on the development of optimization and probabilistic design methods for complex systems. He is a member of the Multidisciplinary Design Optimization and Non-Deterministic Approach technical committees of the American Institute of Aeronautics and Astronautics.

Rudy Pastel was born in Martinique where he studied science until he joined ENSTA-ParisTech (France) specializing in applied mathematics. In 2008, he joined ONERA to work on a Ph.D. dealing with rare event probability and extreme quantile estimation. He defended his Ph.D. thesis in 2012 at the University of Rennes 1 (France). He is currently a data scientist in the Volkswagen data lab in Munich (Germany).

Christelle Vergé passed the agregation of mathematics in 2011 and graduated from the Bordeaux University (France) with an M.Sc. in statistics and stochastic models. She defended in 2015 a Ph.D. in applied mathematics in applied mathematics at ONERA and the "École Polytechnique," under the supervision of Prof. Del Moral, Prof. Moulines, and Dr. Morio. She is interested in parallelization of sequential Monte Carlo methods.

Grégoire Winckelmans holds a mechanical engineering degree from UCL (Belgium), a postgraduate degree in aeronautics and aerospace from the von Karman Institute (Belgium), and M.Sc. and a Ph.D. degrees in aeronautics from California Institute of Technology (USA). His expertise lies in fluid mechanics, advanced numerical simulation, and the modeling of turbulent flows. He and his team have participated on numerous research projects devoted to aircraft wake vortices, have authored papers, and have developed the DVM/PVM/WAKE4D software.

Abbreviations

^{2}SMART	support vector margin algorithm for reliability estimation
ADS	adaptive directional sampling
AMISE	asymptotic integrated square error
ANOVA	analysis of variance
as	almost surely
AST	adaptive splitting technique
AV	antithetic variates
cdf	cumulative distribution function
CE	cross entropy
CFD	computational fluid dynamics
CMC	crude Monte Carlo
CC	correlation coefficient
CV	control variate
DoE	design of experiment
DS	directional sampling
DVM	deterministic wake vortex model
DSS	directional stratified sampling
EGRA	efficient global reliability analysis
EVT	extreme value theory
FAST	Fourier amplitude sensitivity test
FORM	first-order reliability method
FOSPA	first-order saddle point approximation
GEV	generalized extreme value
GAISA	general adaptive importance splitting algorithm
GPD	generalized Pareto distribution
iid	independent and identically distributed
IS	importance sampling
kde	kernel density estimator
LDT	large deviation theory
LHS	Latin hypercube sampling
LS	line sampling
MCMC	Monte Carlo Markov chain
MLW	maximum landing weight
MISE	mean integrated square error
MSE	mean squared error
MT	mean translation
n/a	not applicable

NAIS	nonparametric adaptive importance sampling
NORAD	North American Aerospace Defense Command
OAT	one variable at a time
PCC	partial correlation coefficient
POT	peak over threshold
pdf	probability density function
PSO	particle swarm optimization
QMC	quasi-Monte Carlo
QP	quadratic programming
rv	random variable
SA	sensitivity analysis
SRC	standardized regression coefficient
SS	stratified sampling
SC	scaling
SORM	second-order reliability method
SUR	stepwise uncertainty reduction
SVM	support vector machine
SVR	support vector regression
TLE	two-line elements
WIR	weighted importance resampling

Introduction to rare event probability estimation

1

J. Morio, M. Balesdent

1.1 The book purposes

Rare event probability estimation has become a large area of research in the reliability engineering and system safety domains. A significant number of methods have been proposed to reduce the computational burden for the estimation of rare events from advanced sampling approaches to extreme value theory. However, it is often difficult in practice to determine which algorithm is the most adapted to a given problem. The purposes of this book are thus to provide a broad up-to-date view of the currently available techniques to estimate rare event probabilities described with unified notations, mathematical pseudocodes to ease their potential implementation, and finally, a large spectrum of simulation results on academic and realistic use cases. We detail the pros and cons of the different algorithms depending on the problem characteristics to answer this question: Which rare event probability estimation techniques are the most adapted to the reader's situation?

1.2 What are the events of interest considered in this book?

Let us first specify the main characteristics of the events that are considered in this book:

- If the event of interest happens, it has a highly severe affect on the performances or the safety of the system. If we want to analyze the risk of the event, a measure of damage must be defined (Kaplan & Garrick, 1981).
- The event probability is low; more precisely, it is difficult to observe realizations of this event of interest considering the available simulation budget. Thus, such rare event probability cannot be estimated easily with direct Monte Carlo simulations.
- The rare event probability must be assessed very accurately because an error on the rare event probability could have serious consequences. The underestimation of the rare event probability must be completely avoided.

The issue of estimating rare event probabilities corresponds to a very large number of applicative situations. In this book, we focus on rare event probability estimation in aerospace systems. For instance, Chapter 11 presents the estimation of the probability that a satellite and a spatial debris collide. This probability must be estimated with accuracy because a collision will end the mission of the satellite. Thus, it must be

Estimation of Rare Event Probabilities in Complex Aerospace and Other Systems
http://dx.doi.org/10.1016/B978-0-08-100091-5.00001-0

avoided at all costs because of the required financial budget to launch a satellite. Typical probabilities of collision between satellite and debris are lower than 10^{-4} and thus cannot be estimated with direct Monte Carlo simulations with a sufficient accuracy.

1.3 The book organization

The first part of the book provides the essential background in probability and statistics and describes the different complex systems considered for the estimation of rare event probabilities. In the second part, we review the main approaches to estimate rare event probabilities. For each technique, its principle and its main theoretical and experimental features are presented. We also propose a mathematical pseudocode and some generally efficient tuning parameters. Some application results on academic test cases are finally given to evaluate the potential of the described method on different situations (high dimension, multiple failure domain, etc.). The performances of the different algorithms to estimate rare event probabilities are then compared in a third part on realistic test cases from the aerospace field. This comparison enables us in the last part of this book to propose some guidelines to determine which methods are the most adapted to a given system, depending on its main characteristics (e.g., dimension, the available simulation budget).

Reference

Kaplan, S., & Garrick, B. J. (1981). On the quantitative definition of risk. *Risk Analysis*, *1*(1), 11–27.

Part One

Essential Background in Mathematics and System Analysis

This page intentionally left blank

Basics of probability and statistics

2

J. Morio, M. Balesdent

It is necessary to have a minimal knowledge of some fundamental definitions and theorems of probability and statistics to understand the principles of the different rare event probability estimation methods. The main goal of this chapter is thus to review the elementary notions on this subject, which will be continuously used in the remainder of the book. For a detailed introduction to probability, we can consult Jacod and Protter (2003) or Tijms (2004).

2.1 Probability theory operators

2.1.1 Elements of vocabulary

Definition 2.1.1. *A random experiment (also called* trial*) is an experiment that leads to different results depending on the randomness. The result of the experiment is called* outcome *and is generally denoted by ω. The set of all the possible outcomes (also called* sample space*) of a random experiment is denoted by Ω.*

The set Ω can be more or less complex, depending on the system to analyze or the phenomenon to observe. For instance, if the considered random experiment is a six-sided dice throw, the sample set Ω is equal to $\Omega = \{1, 2, 3, 4, 5, 6\}$. Figure 2.1 represents 100 random results of a dice throw experiment. The set Ω can also be infinite when the system is more complex. Indeed, the fallout position of a launcher stage (i.e., latitude and longitude) evolves in the two-dimensional sample space $\Omega = [-90°, +90°] \times [-180°, +180°]$ (where $\times$ stands for the Cartesian product between the longitude and the latitude spaces). One hundred trials of launcher stage fallout positions obtained with a simulator are represented in Figure 2.2.

Definition 2.1.2. *We denote a random event as a set of outcomes of a random experiment. A random event is thus a subset of Ω, the sample space.*

In the previous example, the set of outcomes corresponding to "the longitude of the launcher stage fallout is in $[-30°, -25°]$ and its latitude in $[-2°, +1°]$" describes a random event.

Definition 2.1.3. *Let $\mathcal{P}(\Omega)$ (also denoted 2^{Ω}) be the powerset of Ω (i.e., the set of all the subsets of Ω, including Ω and the empty set $\emptyset$). A nonempty subset $\mathcal{A} \in \mathcal{P}(\Omega)$ is called a* σ-algebra on Ω *if it verifies the three following properties:*

- $\Omega \in \mathcal{A}$
- *Stability by complementation: if $A \in \mathcal{A}$, then ${}^{c}A \in \mathcal{A}$ (with ${}^{c}A$ the complement of A in Ω)*
- *Stability by countable unions: if $A_1, A_2, \ldots A_n$ are in $\mathcal{A}$, then $A_1 \cup A_2 \cup \cdots \cup A_n \in \mathcal{A}$*

In the case of a six-sided dice throw, different σ-algebras can be defined such as $\{\emptyset, \{1, 2, 3, 4, 5, 6\}\}$, $\{\emptyset, \{1\}, \{2, 3, 4, 5, 6\}, \{1, 2, 3, 4, 5, 6\}\}$, or $\mathcal{P}(\Omega)$. The most known

Estimation of Rare Event Probabilities in Complex Aerospace and Other Systems
http://dx.doi.org/10.1016/B978-0-08-100091-5.00002-2

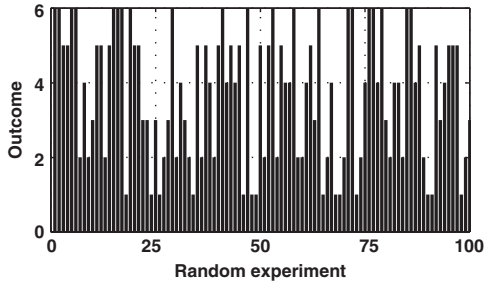

Figure 2.1 One hundred random results of a six-sided dice throw.

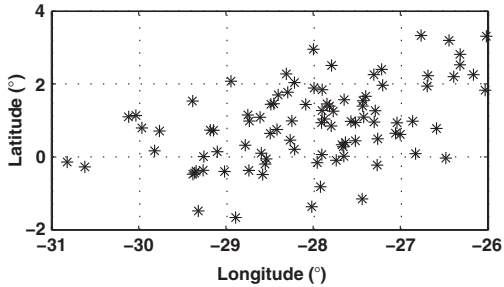

Figure 2.2 One hundred trials of launcher stage fallout position.

σ-algebra is the Borel algebra. A Borel algebra on Ω is the smallest σ-algebra containing all open sets. For example, if $\Omega = \mathbb{R}$, the Borel algebra $\mathcal{B}(\mathbb{R})$ is the smallest σ-algebra on $\mathbb{R}$ that contains all the intervals. The notion of σ-algebra allows us to define the measurable sets. Indeed, according to the theory of measure, the doublet $(\Omega, \mathcal{A})$ is a measurable set.

Definition 2.1.4 (Kolmogorov, 1950)**.** *A probability* $\mathbb{P}$ *on the measurable set* $(\Omega, \mathcal{A})$ *is a mapping from* $\mathcal{A}$ *to* $[0, 1]$*, such that*

- $\mathbb{P}(\Omega) = 1$
- *Let* $(A_i)_{i\in\mathbb{N}}$ *be a countable collection of disjoint events in* $\mathcal{A}$: $\mathbb{P}\left(\bigcup_{i=0}^{+\infty} A_i\right) = \sum_{i=0}^{+\infty} \mathbb{P}(A_i)$

The triple $(\Omega, \mathcal{A}, \mathbb{P})$ *is called* probability space.

To model a dice throw, it is common to choose $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{A} = \mathcal{P}(\Omega)$, and $\mathbb{P}(\{\omega\}) = \frac{1}{6}$ for all the singletons $\omega \in \Omega$.

Property 2.1.1. *Let us consider A and B, two elements of* $\mathcal{A}$ *:*

- $\mathbb{P}({}^{c}A) = 1 - \mathbb{P}(A)$
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
- *if* $A \subseteq B$*, then* $\mathbb{P}(A) \leq \mathbb{P}(B)$

- *Let* $(A_i)_{i\in\mathbb{N}}$ *be a finite or countable collection of disjoint elements such that* $\sum_{i\in\mathbb{N}} \mathbb{P}(A_i) = 1$ *(also called a* partition of Ω*), then the law of total probability holds:*
$$\mathbb{P}(B) = \sum_{i\in\mathbb{N}} \mathbb{P}(A_i \cap B)$$

- *Let* $(A_i)_{i\in\mathbb{N}}$ *be an increasing sequence of elements of* $\mathcal{A}$ *,* $(\forall j \geq i \geq 0, A_i \subseteq A_j)$*, then it holds:*
$$\mathbb{P}\left(\bigcup_{i\in\mathbb{N}} A_i\right) = \lim_{i\to\infty} \mathbb{P}(A_i)$$

Definition 2.1.5. *Let* $(\Omega, \mathcal{A}, \mathbb{P})$ *and* $(E, \mathcal{E})$ *be a probability space and a measurable space, respectively. A random variable (rv) is a measurable application* $X : \Omega \to E$ *from* $(\Omega, \mathcal{A})$ *to* $(E, \mathcal{E})$*. A random variable is thus characterized by*
$$\forall B \in \mathcal{E},\ X^{-1}(B) = \{\omega \in \Omega, X(\omega) \in B\} \in \mathcal{A}$$

A real-valued random variable is a function $X : \Omega \to \mathbb{R}$ such that $\{\omega \in \Omega, X(\omega) \leq c\} \in \mathcal{A}$ for all $c \in \mathbb{R}$. A real-valued random vector of dimension d is a function $X : \Omega \to \mathbb{R}^d$ such that $\{\omega \in \Omega, X(\omega) \in \mathbf{B}\} \in \mathcal{A}$ for all $\mathbf{B} \in \mathcal{B}(\mathbb{R}^d)$.

Definition 2.1.6. *Let* $(\Omega, \mathcal{A}, \mathbb{P})$ *and* $(E, \mathcal{E})$ *be a probability space and a measurable set respectively, and let* $X : \Omega \to E$ *be a random variable. For every* $B \in \mathcal{E}$*, we define*
$$\mathbb{P}_X(B) = \mathbb{P}\left(\{\omega \in \Omega, X(\omega) \in B\}\right) = \mathbb{P}(X^{-1}(B))$$

$\mathbb{P}_X : \mathcal{E} \to [0, 1]$ *is the probability law (also called* probability distribution*) of* X*.*

It is often written for conciseness $\mathbb{P}_X(B) = \mathbb{P}(X \in B)$.

2.1.2 Notion of dependence of random events and conditional probabilities

Definition 2.1.7. *Let* A *and* B *be two random events such that* $\mathbb{P}(B) > 0$*. The probability of* A *knowing* B *denoted* $\mathbb{P}(A|B)$ *is defined by*
$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \tag{2.1}$$

Theorem 2.1.1 (Bayes' theorem)**.** *Let* A *and* B *be two random events such that* $\mathbb{P}(B) > 0$*. The probability of* A *knowing* B*,* $\mathbb{P}(A|B)$ *is equal to*
$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

This theorem is obtained from Equation (2.1) because we have
$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A)$$

Definition 2.1.8. *Let A and B be two random events. A and B are said to be independent if*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$$

Consequently, we deduce from the definition of a conditional probability that $\mathbb{P}(A|B) = \mathbb{P}(A)$ when the events A and B are independent. The realizations of A do not depend on the realizations of B.

2.1.3 Continuous random variables

2.1.3.1 Definitions

For the sake of conciseness, the case of discrete random variables is not detailed in this book because the vast majority of industrial systems to be analyzed involve continuous probability distributions. Nevertheless, for more details about discrete random variables, consult Gordon (1997) and Grinstead and Snell (1998). In the following, we use only continuous random variables defined as follows:

Definition 2.1.9. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and X a variable that is a real-valued measurable mapping from $(\Omega, \mathcal{A})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. The variable X is a continuous random variable if there exists a function $f : \Omega \to \mathbb{R}$ such that:*

1. $\forall x \in \mathbb{R}, f(x) \geq 0$
2. *f is a continuous function almost everywhere on Ω*
3. $\int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x = 1$
4. $\mathbb{P}(X \in A) = \int_A f(x)\,\mathrm{d}x,\ \forall A \in \mathcal{B}(\mathbb{R})$

The function f is the probability density function (pdf) of the rv X.

Given a random variable X with pdf f, a random sample of length N is a set of N independent and identically distributed (iid) random variables with pdf f. In Figure 2.3, we present a simple example of pdf f_1 defined on $\mathbb{R}$ and some random samples generated with this pdf.

Definition 2.1.10. *Let X be a continuous random variable with a pdf f. The cumulative distribution function (cdf) F of X is denoted by*

$$\begin{aligned} F &: \mathbb{R} \to [0,1] \\ x &\to \mathbb{P}(X \leq x) \end{aligned}$$

The cdf can be defined from the pdf by

$$\forall x \in \mathbb{R},\ F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$$

It is important to notice that the cdf F uniquely defines the probability law of X (see Definition 2.1.6). The cumulative distribution function F_1 of the pdf f_1 in Figure 2.3 is represented in Figure 2.4.

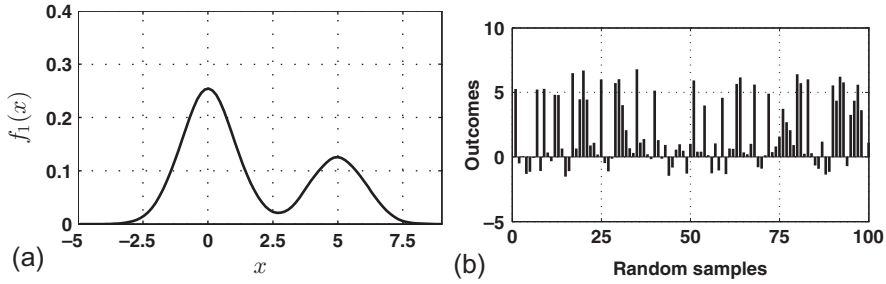

Figure 2.3 Description of the probability density function f_1. (a) pdf f_1. (b) Corresponding iid random samples generated with pdf f_1.

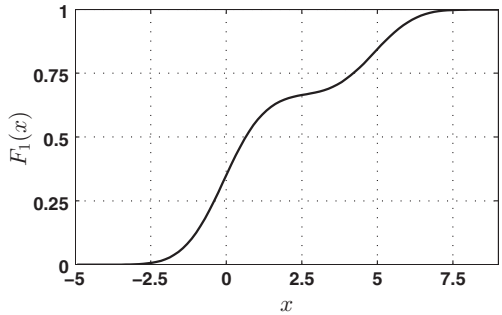

Figure 2.4 Cumulative distribution function F_1 of the pdf f_1.

2.1.3.2 Parameters of continuous random variables

Assuming that X is a continuous rv with a pdf f, then we define the following parameters, if the integral convergence is ensured:

- The mathematical expectation (first-order moment) $\mathbb{E}(X)$ of X with $\mu = \mathbb{E}(X) = \int_{\mathbb{R}} xf(x)\,\mathrm{d}x$
- The variance (centered second-order moment) $\mathbb{V}(X)$ of X with $\mathbb{V}(X) = \mathbb{E}\left[(X - \mathbb{E}(X))^2\right] = \int_{\mathbb{R}} (x-\mu)^2 f(x)\,\mathrm{d}x$
- The standard deviation of X with $\sigma(X) = \sqrt{\mathbb{V}(X)}$
- The p-order centered moment $m_p(X)$ with $m_p(X) = \mathbb{E}\left[(X - \mathbb{E}(X))^p\right] = \int_{\mathbb{R}} (x-\mu)^p f(x)\,\mathrm{d}x$ (for $p \geq 2$)
- The coefficient of variation (also called relative standard deviation or relative standard error) of X with $c_v(X) = \frac{\sigma(X)}{\mathbb{E}(X)}$
- The continuous entropy $H(X)$ of X with

$$H(X) = \int_{\mathbb{X}} f(x)\ln(f(x))\,\mathrm{d}x$$

where $\mathbb{X}$ is the support of f, that is, $\mathbb{X} = \{x \in \mathbb{R} | f(x) > 0\}$

Theorem 2.1.2 (König–Huygens)**.** *If the moments* $\mathbb{E}(X)$ *and* $\mathbb{E}(X^2)$ *of a random variable X are known, then the variance is given by*

$$\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

Theorem 2.1.3 (Law of total variance)**.** *If X and Y are random variables on the same probability space, and the variance of Y is finite, then*

$$\mathbb{V}(Y) = \mathbb{V}\left(\mathbb{E}\left(Y|X\right)\right) + \mathbb{E}\left(\mathbb{V}\left(Y|X\right)\right)$$

The conditional variance of Y given X is defined by

$$\mathbb{V}\left(Y|X\right) = \mathbb{E}\left(\left[Y - \mathbb{E}\left(Y|X\right)\right]^2 |X\right)$$

Definition 2.1.11. *If X is an rv with cdf F and* $\alpha \in [0,1]$*, the* α*-quantile is defined by the real* q_α *such that:*

$$q_\alpha = \inf_{v\in\mathbb{R}}\{\mathbb{P}(X \leq v) \geq \alpha\} = \inf_{v\in\mathbb{R}}\{F(v) \geq \alpha\}$$

α is equivalent to a probability, and thus an α-quantile is assimilated to a quantile of probability α. It is then possible to define the generalized inverse cdf F^{-1} with

$$F^{-1}(y) = \inf_{v\in\mathbb{R}}\{F(v) \geq y\}$$

The median corresponds, for instance, to the 0.5-quantile and thus to $F^{-1}(0.5)$.

Definition 2.1.12. *A mode* x_m *of a continuous rv X of pdf f corresponds to a value x at which f has a local maximum value.*

If a mode x_m is the global maximum of the pdf f, then x_m is a *major mode* of f. Otherwise, x_m is called a *minor mode* of f. If a pdf f has only one mode on all its support, then f is said to be *unimodal*. In the contrary case, f is said to be *multimodal*. Some pdfs such as the uniform distribution have no mode.

2.1.4 Continuous multivariate random variables

2.1.4.1 Definitions and theorems

The notions of pdf, cdf, mathematical expectation, and so on can be extended to multivariate random vectors composed of d random variables.

Definition 2.1.13. *Let us define the sample space* $\Omega \subset \mathbb{R}^d$*,* $\mathcal{A}$ *a* σ*-algebra and a vector* $\mathbf{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$ *of dimension d that is a real-valued measurable mapping from* $(\Omega, \mathcal{A})$ *to* $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$*. The variable* $\mathbf{X}$ *is a continuous random vector or a continuous multivariate random variable if there exists a function f defined on* Ω *such that:*

1. $\forall\, \mathbf{x} = (x_1, \ldots, x_d)^{\mathrm{T}} \in \mathbb{R}^d, f(\mathbf{x}) \geq 0$
2. *f is a continuous function almost everywhere on* Ω
3. $\int_{\mathbb{R}^d} f(x_1, \ldots, x_d)\, \mathrm{d}x_1 \cdots \mathrm{d}x_d = 1$
4. $\forall \mathbf{A} \in \mathcal{B}(\mathbb{R}^d)$, $\mathbb{P}(\mathbf{X} \in \mathbf{A}) = \int_{\mathbf{A}} f(x_1, \ldots, x_d)\, \mathrm{d}x_1 \cdots \mathrm{d}x_d$

The function f is the pdf of the random vector $\mathbf{X}$*. The cdf of* $\mathbf{X}$ *is defined by the function F on* $\mathbb{R}^d$*:*

$$F(x_1, \ldots, x_d) = \mathbb{P}\left(X_1 \leq x_1, \ldots, X_d \leq x_d\right)$$

Definition 2.1.14. *If* $\mathbf{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$ *is a random vector on* $\mathbb{R}^d$*, then the ith component* X_i *of* $\mathbf{X}$ *is an rv. The marginal pdf* f_{X_i} *of* X_i *can be determined when f is known in the following way:*

$$f_{X_i}(x_i) = \int_{\mathbb{R}^{d-1}} f(x_1, \ldots, x_i, \ldots, x_d)\, \mathrm{d}x_1 \cdots \mathrm{d}x_{i-1}\, \mathrm{d}x_{i+1} \cdots \mathrm{d}x_d$$

The cdf of f_{X_i} *is denoted* F_{X_i}*.*

An example of a two-dimensional random vector pdf is given in Figure 2.5.

Remark 2.1.1. *In this book, vectors and matrices will be displayed with a bold letter. The components* X_i*,* $i = 1, \ldots, d$ *of a d-dimensional random vector* $\mathbf{X}$ *are scalar so that* $\mathbf{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$*. N iid random samples with the same distribution as* $\mathbf{X}$ *are denoted* $\mathbf{X}_i$*,* $i = 1, \ldots, N$*. A random sample* $\mathbf{X}_i$ *has d components such that* $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})^{\mathrm{T}}$*. If X is a one-dimensional rv, then N iid random samples with the same distribution as X are denoted* X_i*,* $i = 1, \ldots, N$*.*

Similarly to the univariate rv parameters described in Section 2.1.3.2, it is possible to define several characteristics for the random vectors:

- The mathematical expectation $\mathbb{E}(\mathbf{X})$ of $\mathbf{X}$ is the vector of the expectations of its components

$$\mathbb{E}(\mathbf{X}) = \mathbb{E}\begin{pmatrix} X_1 \\ \vdots \\ X_d \end{pmatrix} = \begin{pmatrix} \mathbb{E}(X_1) \\ \vdots \\ \mathbb{E}(X_d) \end{pmatrix}$$

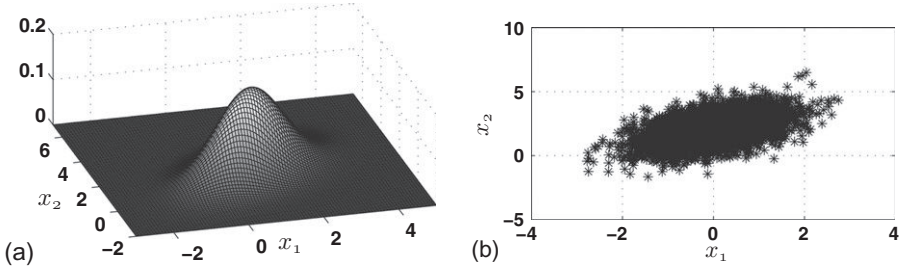

Figure 2.5 Example of a two-dimensional random vector. (a) Two-dimensional pdf. (b) Corresponding iid random samples.

- The covariance matrix $\mathbb{V}(\mathbf{X})$ of $\mathbf{X}$ with

$$\mathbb{V}(\mathbf{X}) = \mathbb{E}\left[(\mathbf{X} - \mathbb{E}(\mathbf{X}))\,(\mathbf{X} - \mathbb{E}(\mathbf{X}))^{\mathrm{T}}\right]$$
$$= \begin{pmatrix} \mathbb{V}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_d) \\ \mathrm{Cov}(X_2, X_1) & \mathbb{V}(X_2) & \cdots & \mathrm{Cov}(X_2, X_d) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_d, X_1) & \mathrm{Cov}(X_d, X_2) & \cdots & \mathbb{V}(X_d) \end{pmatrix}$$

  with $\mathrm{Cov}(X_i, X_j) = \mathbb{E}\left[(X_i - \mathbb{E}(X_i))\,(X_j - \mathbb{E}(X_j))\right] = \mathrm{Cov}(X_j, X_i)$
- The continuous entropy $H(\mathbf{X})$ of $\mathbf{X}$ with

$$H(\mathbf{X}) = \int_{\mathbb{X}} f(x_1, \ldots, x_d) \ln\left(f(x_1, \ldots, x_d)\right) \mathrm{d}x_1 \cdots \mathrm{d}x_d$$

  where $\mathbb{X}$ is the support of f

Theorem 2.1.4 (Transport theorem)**.** *Let* $\mathbf{X}$ *be a continuous d-dimensional random vector with a joint pdf* f*,* $\phi : \mathbb{R}^d \to \mathbb{R}$ *a measurable function,* $\mathbb{E}(\phi(\mathbf{X}))$*, is given by*

$$\mathbb{E}(\phi(\mathbf{X})) = \int_{\mathbb{R}^d} \phi(x_1, \ldots, x_d) f(x_1, \ldots, x_d)\, \mathrm{d}x_1 \cdots \mathrm{d}x_d$$

if the integral is absolutely convergent.

This theorem allows us to define the probabilities of interest of an industrial system, such as probability of failure. This theorem will be extensively used in the following chapters.

2.1.4.2 Dependence of multivariate random variables

Definition 2.1.15. *Let* $\mathbf{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$ *be a continuous d-dimensional random vector with joint pdf* f *and marginal pdfs* $f_{X_1}, \ldots, f_{X_d}$*. The random variables* $X_1, \ldots, X_d$ *are said mutually independent (or independent) if*

$$f(x_1, \ldots, x_d) = \prod_{i=1}^{d} f_{X_i}(x_i)$$

or

$$F(x_1, \ldots, x_d) = \prod_{i=1}^{d} F_{X_i}(x_i)$$

If the random variables $X_1, \ldots, X_d$ are independent, then the covariance matrix of the random vector $\mathbf{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$ is equal to a diagonal matrix with components $\mathbb{V}(X_i)$, $i = 1, \ldots, d$. The converse implication is not true.

Copulas are mathematical objects that capture the structure of dependence between different random variables. In this chapter, we give only a short introduction concerning the copula theory. For more details, consult Nelsen (2006), Jaworski, Durante, Haerdle, & Rychlik (2010) and Lebrun (2013).

Definition 2.1.16. *A cdf $C : [0,1]^d \to [0,1]$ is a copula if it is a joint cumulative distribution of a d-dimensional random vector defined on the unit hypercube $[0,1]^d$ with a uniform marginal pdf.*

The copula allows to separate from a joint probability distribution the contribution of the marginals and the contribution of the dependence structure between each component of the considered random vector.

Theorem 2.1.5 (Sklar's theorem (Sklar, 1959))**.** *Any multivariate cumulative distribution F of $\mathbf{X} = (X_1, \ldots, X_d)^{\mathrm{T}} \in \mathbb{R}^d$ can be expressed according to its marginal cumulative distributions $F_{X_i}, i \in \{1, \ldots, d\}$ and a copula C:*

$$F(x_1, \ldots, x_d) = C(F_{X_1}(x_1), \ldots, F_{X_d}(x_d))$$

If the marginal distributions $F_{X_1}, \ldots, F_{X_d}$ are continuous, the copula C is unique and we have,

$$\forall \mathbf{u} = (u_1, \ldots, u_d)^T \in [0,1]^d, C(\mathbf{u}) = F(F_{X_1}^{-1}(u_1), \ldots F_{X_d}^{-1}(u_d))$$

Samples of different bivariate distributions generated with Gaussian, Student, Clayton, and Gumbel copulas are provided in Figure 2.6. Unlike the marginals, the structure of dependency between different random variables is generally very complex to grasp. One has to choose the appropriate copula (e.g., Gaussian, Clayton, Frank, Gumbel) that reflects at best the structure of dependence of the random vector to characterize. In practice, the data at one's disposal in complex industrial applications are often too insufficient to have an accurate estimation of the input dependencies. In such cases, one generally consider the inputs of a system as independent. This assumption is not trivial and might lead to a misrepresentation of the studied phenomena.

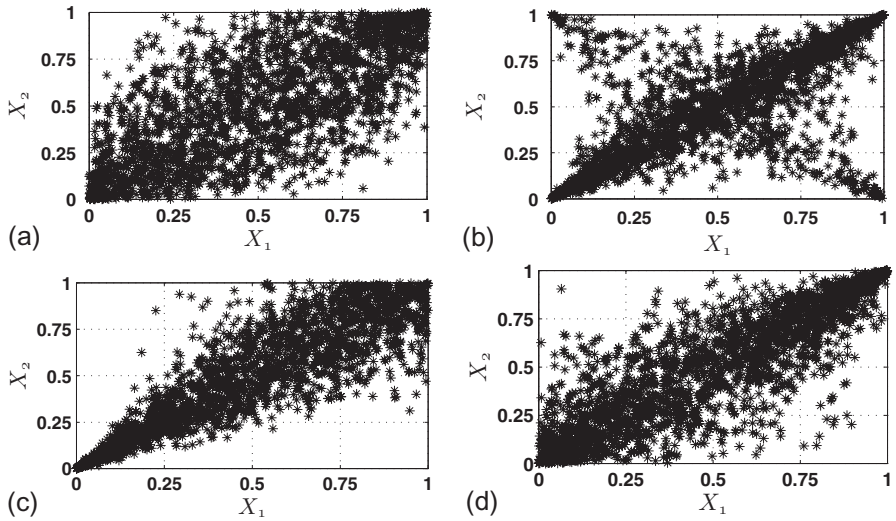

Figure 2.6 Random samples generated with different copulas. (a) Gaussian copula. (b) Student copula. (c) Clayton copula. (d) Gumbel copula.

2.1.5 Point estimation

Let us consider $X_1, \ldots, X_N$, N independent samples of the random variable X with density f_{θ_0}. The term θ_0 is an unknown *a priori* parameter such that $\theta_0 \in \Theta$ with Θ the parameter space. For example, if f_{θ_0} is a Gaussian pdf, θ_0 might be the mean or the variance. The objective of point estimation is to estimate the parameter θ_0 from the samples $X_1, \ldots, X_N$.

Definition 2.1.17. *A point estimator $\hat{\theta}$ of θ_0 is a statistic $T : \mathbb{R}^N \to \Theta$, that is, any function from the data space to the parameter space.*

There are infinite numbers of candidate point estimators. It is nevertheless possible to characterize the quality of an estimator for the estimation of θ_0 by defining its bias and its variance.

Definition 2.1.18. *The bias $b_{\hat{\theta}}$ of an estimator $\hat{\theta}$ of θ_0 is defined by*

$$b_{\hat{\theta}} = \mathbb{E}(\hat{\theta}) - \theta_0$$

Its variance $v_{\hat{\theta}}$ is given by the following expression:

$$v_{\hat{\theta}} = \mathbb{E}\left((\hat{\theta} - \mathbb{E}(\hat{\theta}))^2\right)$$

The mean squared error $MSE(\hat{\theta})$ is then obtained with

$$MSE(\hat{\theta}) = \mathbb{E}\left((\hat{\theta} - \theta_0)^2\right) = v_{\hat{\theta}} + b_{\hat{\theta}}^2$$

The bias gives an indication of the distance between the estimator and θ_0 whereas the variance describes the estimator dispersion near $\mathbb{E}(\hat{\theta})$. In practice, valuable estimators of a parameter have a good trade-off between bias and variance.

Let us give a simple example of point estimation and consider the empirical mean $\hat{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$ as an estimator of the mathematical expectation $\mathbb{E}(X)$. Its bias can be easily determined. Indeed, we have

$$\mathbb{E}\left(\hat{X}_N\right) = \frac{1}{N}\mathbb{E}\left(\sum_{i=1}^{N} X_i\right) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left(X_i\right)$$

Because $\mathbb{E}\left(X_i\right) = \mathbb{E}\left(X\right)$, we obtain the following equality:

$$b_{\hat{X}_N} = \mathbb{E}\left(\hat{X}_N\right) - \mathbb{E}\left(X\right) = 0$$

The estimator $\hat{X}_N$ is thus said to be an unbiased estimator of $\mathbb{E}(X)$. The variance $v_{\hat{X}_N}$ of $\hat{X}_N$ for the estimation of $\mathbb{E}(X)$ can also be determined in the following way:

$$v_{\hat{X}_N} = \mathbb{E}\left((\hat{X}_N - \mathbb{E}(\hat{X}_N))^2\right)$$

and with the König–Huygens theorem, we have

$$v_{\hat{X}_N} = \mathbb{E}\left((\hat{X}_N)^2\right) - \mathbb{E}\left(\hat{X}_N\right)^2$$

Similar definitions for point estimation may be proposed for multivariate random variables.

2.2 Modeling and sample generation of random variable pdfs

Different statistical tools enable to define random variable pdfs adapted to a given problem. In simple cases, the pdf is chosen among common parametric probability laws and the generation of random samples is well known. A short description of the most common pdfs is presented in Section 2.2.1.1 for univariate rv and in Section 2.2.1.2 in the multivariate case. For more complex pdf designs, a nonparametric kernel density estimator can be used as described in Section 2.2.2. From a given set of samples, this estimator consists in approximating the pdfs of these samples with a mixture of smooth kernel densities. It is also sometimes difficult to sample from a pdf when, for instance, the pdf expression is known up to a constant. Generating random samples in that case requires more complex algorithms such as the Metropolis–Hastings algorithm (see Section 2.3.3).

2.2.1 Overview of common probability distributions

2.2.1.1 Univariate distributions

Uniform distribution

Definition 2.2.1. *The pdf of the univariate uniform law $U_{(a,b)}$ on the interval $[a, b]$ is defined by*

$$f(x) = \frac{1}{b-a}\mathbf{1}_{[a,b]}(x), \quad x \in \mathbb{R}$$

where $\mathbf{1}_{[a,b]}(x)$ is a function that is equal to 1 for $x \in [a, b]$ and 0 elsewhere. The cdf of $U_{(a,b)}$ is

$$F(x) = \begin{cases} 0, & \text{if } x < a \\ \frac{x-a}{b-a}, & \text{if } a \leq x < b \\ 1, & \text{if } b \leq x \end{cases}$$

The mathematical expectation and the variance depend on a and b in the following way: if $X \sim U_{(a,b)}$ ($\sim$ means *follows*), then $\mathbb{E}(X) = \frac{a+b}{2}$ and $\mathbb{V}(X) = \frac{(b-a)^2}{12}$. The uniform distribution is often chosen when it is known that an rv evolves between a and b but without any information on its distribution over this interval. Indeed, it is

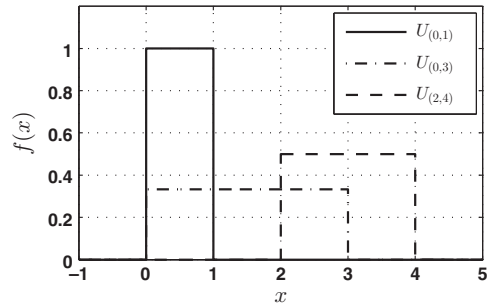

Figure 2.7 Examples of different uniform pdfs.

the maximum entropy distribution among all the continuous distributions with support $[a, b]$. Different examples of uniform distributions are given in Figure 2.7.

Exponential distribution

Definition 2.2.2. *The univariate exponential law $\mathcal{E}(\lambda)$ with $\lambda > 0$ is defined by*

$$f(x) = \lambda \mathrm{e}^{-\lambda x} \mathbf{1}_{x \geq 0}, \quad x \in \mathbb{R}$$

The cdf of $\mathcal{E}(\lambda)$ is

$$F(x) = \left(1 - \mathrm{e}^{-\lambda x}\right) \mathbf{1}_{x \geq 0}, \quad x \in \mathbb{R}$$

If $X \sim \mathcal{E}(\lambda)$, we also have $\mathbb{E}(X) = \frac{1}{\lambda}$ and $\mathbb{V}(X) = \frac{1}{\lambda^2}$. The exponential distribution can be used to model queue waiting time or component lifetime. It is also the maximum entropy distribution among all the continuous distributions with support $(0, +\infty)$ that have an expectation equal to $\frac{1}{\lambda}$. Different examples of exponential pdfs are presented in Figure 2.8.

Gaussian distribution

Definition 2.2.3. *The pdf of the univariate Gaussian law $\mathcal{N}(\mu, \sigma^2)$ is defined by*

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)^2}, \quad x \in \mathbb{R}$$

If $X \sim \mathcal{N}(\mu, \sigma^2)$, we have $\mathbb{E}(X) = \mu$ and $\mathbb{V}(X) = \sigma^2$. The cdf of $\mathcal{N}(\mu, \sigma^2)$ is

$$F(x) = \Phi_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} \mathrm{e}^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} \mathrm{d}t, \quad x \in \mathbb{R}$$

The cdf of the Gaussian distribution of parameters (μ, σ^2) is indeed often denoted by $\Phi_{\mu,\sigma^2}(\cdot)$. The Gaussian (or normal) law is one of the most employed probability

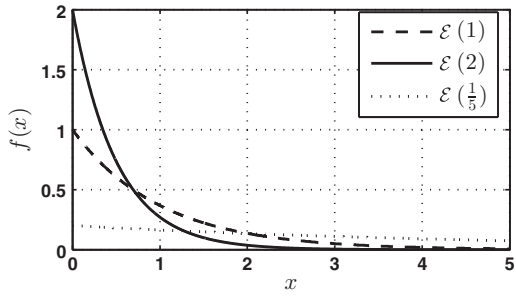

Figure 2.8 Examples of different exponential pdfs.

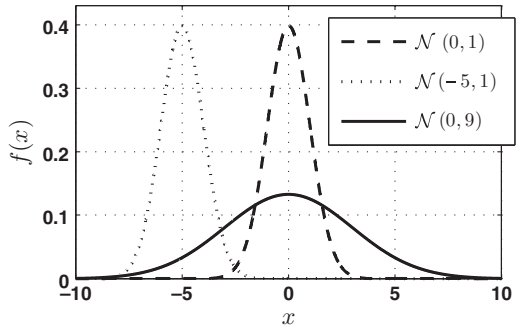

Figure 2.9 Examples of different Gaussian pdfs.

distributions to characterize the distribution of rv and will be extensively used in the application cases. The Gaussian law is the maximum entropy distribution among all real-valued distributions with specified expectation μ and variance σ^2. The central limit theorem (see Section 2.3.2) also justifies this pdf choice in several domains. Different examples of Gaussian pdfs are presented in Figure 2.9. The density $\mathcal{N}(0,1)$ is called the *standard Gaussian pdf*.

Truncated Gaussian distribution

Definition 2.2.4. *The truncated Gaussian univariate distribution $\mathcal{N}(\mu,\sigma^2,a,b)$ is defined by*

$$f(x)=\frac{\frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)^2}}{\Phi_{\mu,\sigma^2}(b)-\Phi_{\mu,\sigma^2}(a)}\mathbf{1}_{[a,b]}(x),\quad x\in\mathbb{R}$$

where $\Phi_{\mu,\sigma^2}(\cdot)$ is the cdf of $\mathcal{N}(\mu,\sigma^2)$ and $b>a$.

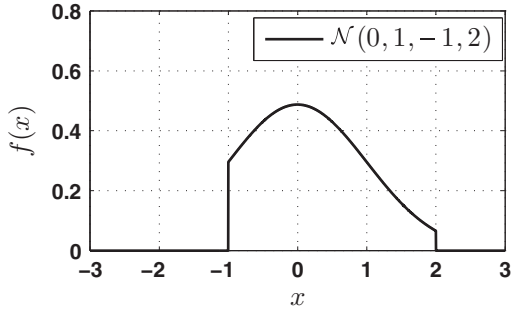

Figure 2.10 Example of a truncated Gaussian pdf.

The infinite support of a Gaussian pdf is not always realistic for some physical characteristics with Gaussian shape. In that case, to improve the modeling of the variable uncertainty, it can be interesting to consider a truncated Gaussian pdf. An example of truncated Gaussian pdf is given in Figure 2.10.

Log-normal distribution

Definition 2.2.5. *The univariate log-normal distribution* $\ln\mathcal{N}(\mu,\sigma^2)$ *is defined on* $\mathbb{R}_+^*$ *by*

$$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\left(\frac{\ln(x)-\mu}{\sqrt{2}\sigma}\right)^2}, \quad x \in \mathbb{R}_+^*$$

The cdf of $\ln\mathcal{N}(\mu,\sigma^2)$ *is*

$$F(x) = \Phi_{\mu,\sigma^2}(\ln(x)), \quad x \in \mathbb{R}_+^*$$

We also have $\mathbb{E}(X) = \mathrm{e}^{\left(\mu+\frac{\sigma^2}{2}\right)}$ and $\mathbb{V}(X) = \left(\mathrm{e}^{\sigma^2}-1\right)\ \mathrm{e}^{(2\mu+\sigma^2)}$ if $X \sim \ln\mathcal{N}(\mu,\sigma^2)$. The log-normal distribution can be used to model variables that stem from the multiplicative product of many independent positive random variables because of the central limit theorem (see Section 2.3.2). Different examples of log-normal pdfs are presented in Figure 2.11.

Cauchy distribution

Definition 2.2.6. *The Cauchy distribution* $\mathcal{C}(a,b)$*, with* $a > 0$ *is given by*

$$f(x) = \frac{a}{\pi}\frac{1}{(x-b)^2+a^2}, \quad x \in \mathbb{R}$$

The cdf of $\mathcal{C}(a,b)$ *is*

$$F(x) = \frac{1}{\pi}\arctan\left(\frac{x-b}{a}\right) + \frac{1}{2}$$

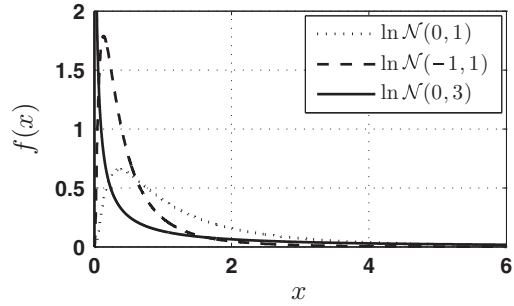

Figure 2.11 Examples of log-normal pdfs.

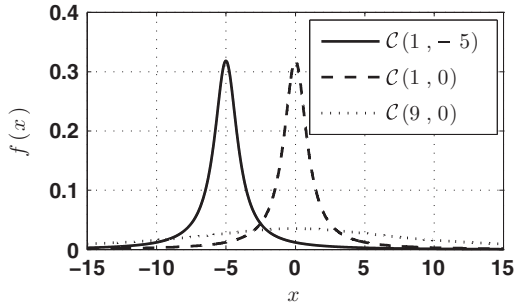

Figure 2.12 Examples of Cauchy pdfs.

The mathematical expectation and variance of a Cauchy distribution $\mathcal{C}(a,b)$ are not defined because their integral convergence is not ensured. The Cauchy distribution has several applications in physics such as in quantum mechanics or calibration problems. Different examples of Cauchy pdfs are presented in Figure 2.12.

Chi-squared distribution

Definition 2.2.7. *The pdf of the chi-squared distribution $\mathcal{X}^2_k$ with k degrees of freedom is defined by*

$$f(x) = \frac{x^{\frac{k}{2}-1}\mathrm{e}^{-\frac{x}{2}}}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})}\mathbf{1}_{x\geq 0}, \quad x \in \mathbb{R}$$

with $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}\mathrm{e}^{-x}\,\mathrm{d}x$.

If $X_1,\ldots,X_k$ are independent standard normal random variables, then the sum of their squares $Q = \sum_{i=1}^k X_i^2$ follows a chi-squared distribution with k degrees of freedom. If $Q \sim \mathcal{X}^2_k$, the mean and the variance of Q are given by $\mathbb{E}(Q) = k$ and $\mathbb{V}(Q) = 2k$, respectively. The chi-squared distribution is notably used to test the

goodness of fit of an observed distribution to a theoretical one or the independence between two random variables in the frame of chi-squared tests. Three examples of chi-squared pdfs with different degrees of freedom are presented in Figure 2.13.

Gamma and beta distributions

Definition 2.2.8. *The gamma law* $\Gamma(\lambda, \alpha)$ *with* $\lambda > 0$ *and* $\alpha > 0$ *is given by*

$$f(x) = \frac{1}{\Gamma(\alpha)} \lambda^{\alpha} x^{\alpha-1} \mathrm{e}^{-\lambda x}, \quad x \in \mathbb{R}_+^*$$

with $\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} \mathrm{e}^{-x} \, \mathrm{d}x$.

The expectation and the variance have the following expressions: if $X \sim \Gamma(\lambda, \alpha)$, then $\mathbb{E}(X) = \frac{\alpha}{\lambda}$ and $\mathbb{V}(X) = \frac{\alpha}{\lambda^2}$. The gamma distribution has been used in a wide range of fields including queuing models or financial services. Different examples of gamma pdfs are presented in Figure 2.14.

Definition 2.2.9. *The beta distribution* $\beta(a, b)$ *with* $a > 0$ *and* $b > 0$ *is given by*

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1}, \quad x \in [0, 1]$$

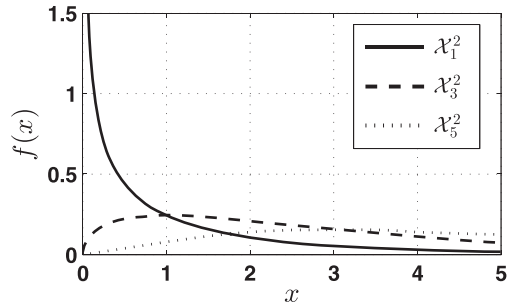

Figure 2.13 Examples of different chi-squared pdfs.

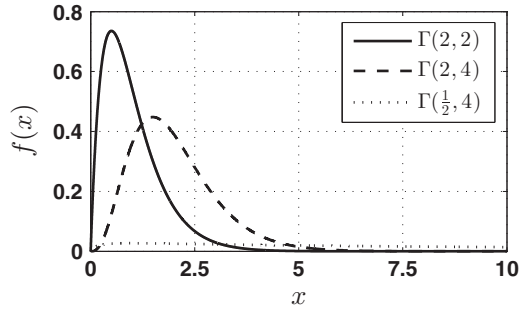

Figure 2.14 Examples of gamma pdfs.

Laplace distribution

Definition 2.2.10. *The univariate Laplace distribution $\mathcal{L}(\mu, b)$, with $b > 0$ and $\mu \in \mathbb{R}$, is given by*

$$f(x) = \frac{1}{2b} \mathrm{e}^{-|\frac{x-\mu}{b}|}, \quad x \in \mathbb{R}$$

The cdf of $\mathcal{L}(\mu, b)$ is

$$F(x) = \begin{cases} \frac{1}{2}\mathrm{e}^{\frac{x-\mu}{b}}, & \text{if } x < \mu \\ 1 - \frac{1}{2}\mathrm{e}^{-\frac{x-\mu}{b}}, & \text{if } x \geq \mu \end{cases}$$

Its expectation and variance are defined with $\mathbb{E}(X) = \mu$ and $\mathbb{V}(X) = 2b^2$ if $X \sim \mathcal{L}(\mu, b)$. The use of Laplace distribution has notably been proposed in signal processing, biological processes, or finance for modeling phenomena with a less smooth behavior than the Gaussian law. Different examples of Laplace pdfs are presented in Figure 2.15.

Some properties of univariate distributions

Definition 2.2.11. *The distribution of a random variable X with pdf f is said to have a (right) heavy tail if and only if*

$$\int_{\mathbb{R}} \exp(\lambda x) f(x) dx = \infty, \ \forall \lambda > 0$$

Heavy-tailed distributions are probability distributions that have heavier tails than exponential distributions. The Cauchy and the log-normal distributions are notable common heavy-tailed distributions. As a counterpart, a distribution is light tailed if and only if

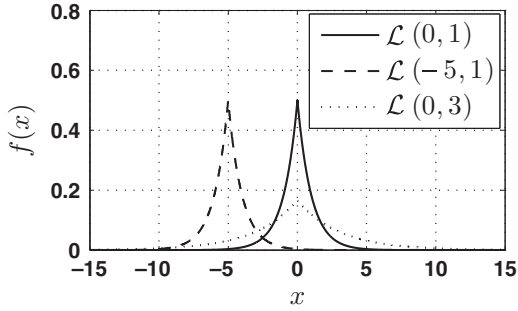

Figure 2.15 Examples of Laplace pdfs.

$$\int_{\mathbb{R}} \exp(\lambda x) f(x) dx < \infty, \ \forall \lambda > 0$$

Definition 2.2.12. *A univariate exponential family is a set of probability distributions whose pdf f with parameter θ has the following expression*

$$f(x|\theta) = h(x)b(\theta)\exp(\mu(\theta)T(x))$$

where the functions h, b, μ, and T are known functions.

The normal, gamma, exponential, log-normal, or chi-squared distributions are all exponential families for instance.

2.2.1.2 Multivariate distributions

If the different components of a random vector are independent, the joint probability density function can be obtained with the product of the univariate pdfs described previously. Moreover, several generalizations of the univariate distributions presented in the previous sections can be expressed. In most cases, only the multivariate normal distribution is practically used for modeling uncertainty in complex systems.

Multivariate normal distribution

Definition 2.2.13. *The multivariate normal (Gaussian) distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, defined on $\mathbb{R}^d$, can be expressed by its pdf*

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}, \mathbf{x} \in \mathbb{R}^d$$

with $\boldsymbol{\mu}$ the expectation and $\boldsymbol{\Sigma}$ the covariance matrix, and where $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$.

An example of multivariate normal pdf is given in Figure 2.16 for $\boldsymbol{\mu} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ and $\boldsymbol{\Sigma} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$.

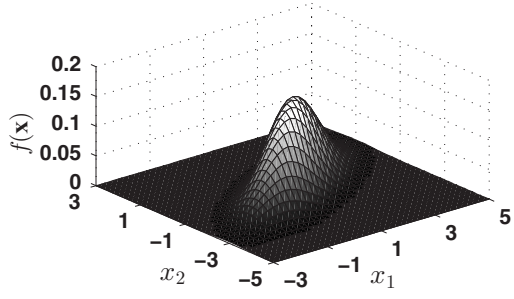

Figure 2.16 Example of a two-dimensional normal pdf.

2.2.2 Kernel-based laws

When no parametric classic density is suitable for the modeling of random variable samples, nonparametric approaches can be efficient alternatives. Let $\mathbf{X}_1, \ldots, \mathbf{X}_N$ be a set of iid random samples with unknown pdf f and dimension d. Kernel density estimator (kde) enables to approximate the pdf f in the following way:

$$\hat{f}_{\mathbf{H}}(\mathbf{x}) = \frac{1}{N|\mathbf{H}|^{\frac{1}{2}}} \sum_{i=1}^{N} K\left(\mathbf{H}^{-\frac{1}{2}}(\mathbf{x} - \mathbf{X}_i)\right)$$

where K is a kernel (a non-negative symmetric function that integrates to 1) and $\mathbf{H}$ is a $d \times d$ symmetric positive definite bandwidth matrix. There are large numbers of potentially efficient kernels, but in practice, the most used kernel is the Gaussian kernel, defined by

$$K(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}} \mathrm{e}^{-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{x}}$$

and the Epanechnikov kernel, defined by

$$K(\mathbf{x}) = \frac{d+2}{2c_d}(1 - \mathbf{x}^{\mathrm{T}}\mathbf{x})\mathbf{1}_{(\mathbf{x}^{\mathrm{T}}\mathbf{x} \leq 1)}$$

where c_d is the volume of the unit sphere in $\mathbb{R}^d$, that is, $c_d = \int_{\mathbb{R}^d} \mathbf{1}_{(\mathbf{x}^{\mathrm{T}}\mathbf{x} \leq 1)}\, \mathrm{d}\mathbf{x}$. The choice of the kernel also depends on the assumptions one makes on the pdf of the samples $\mathbf{X}_1, \ldots, \mathbf{X}_N$ as their distribution tail, and so on. For instance, as noticed in Wand, Marron, & Ruppert (1991), the Gaussian kernel is efficient for estimating densities that are not so far from a Gaussian shape, but it might perform very poorly when this is not the case, especially near the boundaries. The bandwidth values correspond to a trade-off between the bias and the variance of $\hat{f}_{\mathbf{H}}$. In general, a small bandwidth in a given dimension implies a small bias but a large variance. Different approaches suggest an estimate of the optimal bandwidth matrix for a given criterion. In most cases, one can choose an adapted bandwidth $\mathbf{H}_{opt}$ that minimizes the mean integrated square error (MISE) criterion defined by

$$\mathrm{MISE}(\mathbf{H}) = \int_{\mathbb{R}^d} \mathbb{E}\left((f(\mathbf{x}) - \hat{f}_{\mathbf{H}}(\mathbf{x}))^2\right) \mathrm{d}\mathbf{x}$$

Asymptotic versions of MISE (AMISE) can also be considered (Wand & Jones, 1995). Some empirical choices of adapted bandwidths can be found in Silverman (1986). As soon as the kernel and the bandwidth have been determined, it is possible to generate random samples with $\hat{f}_{\mathbf{H}}$. An example of application of kde is presented in Figure 2.17. From the samples generated with an unknown pdf, it is possible to estimate the sample density with kde. The influence of the bandwidth is very significant in the shape of the kde and must be carefully tuned.

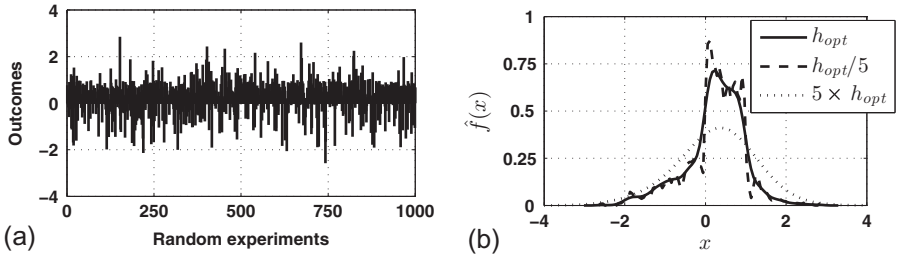

Figure 2.17 Application of kde on samples generated with an unknown pdf. (a) iid samples generated with an unknown pdf. (b) Corresponding kde for different bandwidths.

2.3 Convergence theorems and general statistical methods

2.3.1 Strong law of large numbers

Theorem 2.3.1 (Strong law of large numbers)**.** *Let $X_1, \ldots, X_N$ be a set of iid random variables with the same distribution as X and finite mathematical expectation $\mathbb{E}(X)$. Then the empirical mean $\hat{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$ converges almost surely (as) to the mathematical expectation $\mathbb{E}(X)$,*

$$\widehat{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i \xrightarrow[N\to+\infty]{as} \mathbb{E}(X), \quad \textit{that is } \mathbb{P}\left(\lim_{N\to+\infty} \widehat{X}_N = \mathbb{E}(X)\right) = 1.$$

The strong law of large numbers is very important in statistics because it ensures that the empirical mean converges to the mathematical expectation. A simple application of this theorem is proposed in Figure 2.18 with random variables X_i, $i = 1, \ldots, N$ following a standard univariate Gaussian pdf $\mathcal{N}(0, 1)$. The empirical mean $\hat{X}_N$ converges to 0, that is, the mathematical expectation of $\mathcal{N}(0, 1)$ when N increases. The law of large numbers can be easily extended to the case of multivariate random variables in the following way.

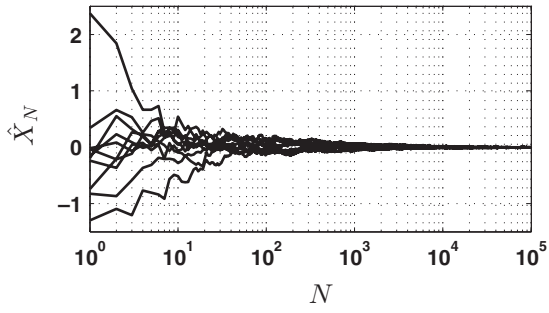

Figure 2.18 Application of the law of large numbers (X_i are sampled from $N(0, 1)$).

Theorem 2.3.2. *Let $\mathbf{X}_1, \ldots, \mathbf{X}_N$ be a set of iid d-dimensional random vectors with the same distribution as $\mathbf{X}$ and finite mathematical expectation $\mathbb{E}(\mathbf{X})$. Then the empirical mean $\widehat{\mathbf{X}}_N = \frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i$ converges almost surely to the mathematical expectation $\mathbb{E}(\mathbf{X})$,*

$$\widehat{\mathbf{X}}_N = \frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i \underset{N\to+\infty}{\overset{as}{\longrightarrow}} \mathbb{E}(\mathbf{X}), \quad \textit{that is } \mathbb{P}\left(\lim_{N\to+\infty} \widehat{\mathbf{X}}_N = \mathbb{E}(\mathbf{X})\right) = 1.$$

2.3.2 Central limit theorem

Theorem 2.3.3 (Central limit theorem (CLT))**.** *Let $X_1, \ldots, X_N$ be a set of iid random variables with the same distribution as X, $\mathbb{E}(X) = \mu$, and $\mathbb{V}(X) < +\infty$. The empirical mean is denoted $\hat{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$. Then the variable $\sqrt{N}\left(\hat{X}_N - \mu\right)$ converges in law $(\mathcal{L})$ to a Gaussian rv with zero mean and variance $\mathbb{V}(X)$,*

$$\sqrt{N}\left(\hat{X}_N - \mu\right) \underset{N\to+\infty}{\overset{\mathcal{L}}{\longrightarrow}} \mathcal{N}(0, \mathbb{V}(X)),$$

that is, for any set $A \subset \mathbb{R}$, which is a continuity set of Y,

$$\lim_{N\to+\infty} \mathbb{P}\left(\sqrt{N}\left(\hat{X}_N - \mu\right) \in A\right) = \mathbb{P}(Y \in A)$$

with Y a Gaussian rv with zero mean and variance $\mathbb{V}(X)$.

The intensive use of the Gaussian law in modeling uncertainty is notably justified by this theorem. An example of application of the CLT is proposed in Figure 2.19 with X_i, $i = 1, \ldots, N$, N rv sampled from the Gamma distribution. The rv $\sqrt{N}\left(\hat{X}_N - \mu\right)$ converges in law to a Gaussian rv when N increases.

Theorem 2.3.4 (Multivariate central limit theorem)**.** *Let $\mathbf{X}_1, \ldots, \mathbf{X}_N$ be a set of iid d-dimensional random vectors with the same distribution as $\mathbf{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$ with a mathematical expectation $\mathbb{E}(\mathbf{X}) = \boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. It is also assumed that $\mathbb{E}\left((X_{ij})^2\right) < +\infty$ for $i = 1, \ldots, N$ and $j = 1, \ldots, d$. The empirical mean is denoted $\widehat{\mathbf{X}}_N = \frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i$. Then the variable $\sqrt{N}\left(\widehat{\mathbf{X}}_N - \boldsymbol{\mu}\right)$ converges in law to a Gaussian rv with zero mean and covariance matrix $\boldsymbol{\Sigma}$,*

$$\sqrt{N}\left(\hat{\mathbf{X}}_N - \boldsymbol{\mu}\right) \underset{N\to+\infty}{\overset{\mathcal{L}}{\longrightarrow}} \mathcal{N}(0, \boldsymbol{\Sigma}),$$

that is, for any set $A \subset \mathbb{R}^d$, which is a continuity set of $\mathbf{Y}$,

$$\lim_{N\to+\infty} \mathbb{P}\left(\sqrt{N}\left(\widehat{\mathbf{X}}_N - \boldsymbol{\mu}\right) \in A\right) = \mathbb{P}(\mathbf{Y} \in A)$$

with $\mathbf{Y}$ a Gaussian rv with zero mean vector and covariance matrix $\boldsymbol{\Sigma}$.

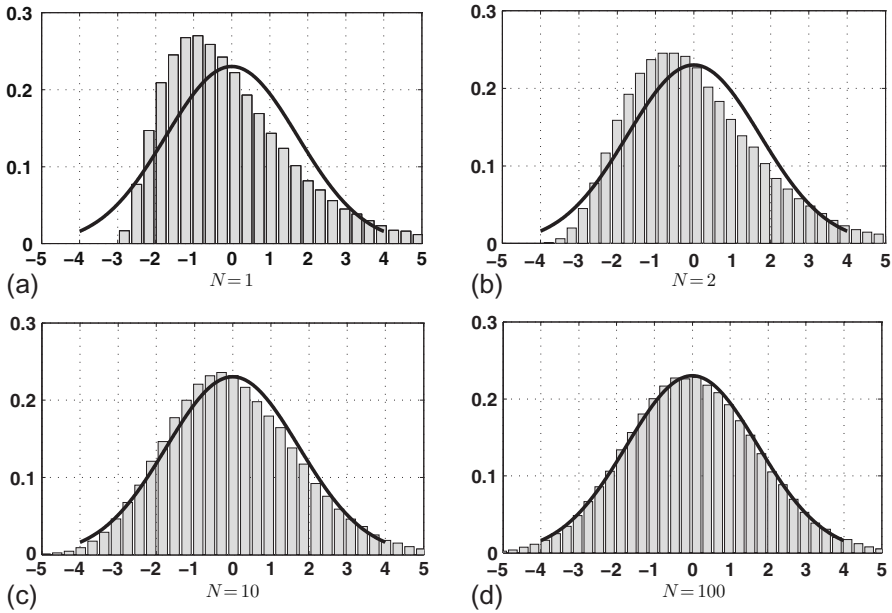

Figure 2.19 Drawing of trial histograms of random variable $\sqrt{N}\left(\hat{X}_N - \mu\right)$ and the pdf of $N\ (0, \mathbb{V}(X))$ to illustrate the CLT when X is sampled from a Gamma distribution. (a) $N = 1$. (b) $N = 2$. (c) $N = 10$. (d) $N = 100$.

2.3.3 Simulation of complex laws using the Metropolis–Hastings algorithm

2.3.3.1 Markov chain

Definition 2.3.1. *A Markov chain in discrete time is a sequence* $\mathbf{X}_0, \mathbf{X}_1, \mathbf{X}_2, \ldots$ *of continuous d-dimensional random vectors that satisfy the following Markov property for all* $n \geq 1$,

$$\mathbb{P}\left(\mathbf{X}_n \in \mathbf{A}_n | \mathbf{X}_0 \in \mathbf{A}_0, \ldots, \mathbf{X}_{n-1} \in \mathbf{A}_{n-1}\right) = \mathbb{P}\left(\mathbf{X}_n \in \mathbf{A}_n | \mathbf{X}_{n-1} \in \mathbf{A}_{n-1}\right)$$

for any measurable subsets $\mathbf{A}_0, \ldots, \mathbf{A}_n$ *of* $\mathbb{R}^d$.

The quantity $\mathbb{P}\left(\mathbf{X}_j \in B(\mathbf{x}', \mathbf{dx}') | \mathbf{X}_{j-1} = \mathbf{x}\right)$ where $B(\mathbf{x}', \mathbf{dx}')$ is a ball of infinitesimal length $\mathbf{dx}'$ centered on $\mathbf{x}'$ is called a *transition kernel*. The different transition kernels considered in this book have a pdf denoted $\pi_j\left(\mathbf{x}_j | \mathbf{x}_{j-1}\right)$ for $j \geq 1$ such that

$$\mathbb{P}\left((\mathbf{X}_0, \ldots, \mathbf{X}_j) \in \mathbf{A}_0 \times \cdots \times \mathbf{A}_j\right)$$

$$= \int_{\mathbf{A}_0 \times \cdots \times \mathbf{A}_j} \pi_0(\mathbf{x}_0) \pi_1\left(\mathbf{x}_1 | \mathbf{x}_0\right) \cdots \pi_j\left(\mathbf{x}_j | \mathbf{x}_{j-1}\right) \mathrm{d}\mathbf{x}_0 d\mathbf{x}_1 \cdots \mathrm{d}\mathbf{x}_j$$

A simple example of Markov chain is a Gaussian random walk in dimension 1. Let us define $n+1$ standard univariate Gaussian random variables $X_0, \ldots, X_n$. The sum $S_n = \sum_{i=0}^{n} X_i$ is a Markov chain on $\mathbb{R}$. The initial density $\pi_0(x_0)$ is given by the law of X_0. The transition kernels of the Markov chain S_n have the following expression

$$\pi_j\left(x_j|x_{j-1}\right) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\left(\frac{x_j - x_{j-1}}{\sqrt{2}}\right)^2}$$

Ten trials of the random walk S_n are proposed in Figure 2.20.

A Markov process $(\mathbf{X}_t)_{t\geq 0}$ in continuous time can also be defined with an initial law $\pi_0(\mathbf{x}_0)$ and transition kernels $\pi_{s,t}(\mathbf{x}'|\mathbf{x}) = \mathbb{P}\left(\mathbf{X}_t \in B(\mathbf{x}', \mathbf{dx}')|\mathbf{X}_s = \mathbf{x}\right)$. In practice, it is in fact not possible to simulate directly time-continuous Markov processes. We approximate processes with a Markov chain in discrete time such that $t = j\Delta_t$ with $j \in \mathbb{N}$ and Δ_t a positive constant as low as possible. More details about Markov processes and Markov chains can be found in Stroock (2014).

2.3.3.2 Some properties of transition kernels

Definition 2.3.2. *A transition kernel π is said to be symmetric if* $\pi(\mathbf{x}'|\mathbf{x}) = \pi(\mathbf{x}|\mathbf{x}') \quad \forall \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$.

The multivariate standard Gaussian distribution is a symmetric kernel. In that case, we have

$$\pi(\mathbf{x}'|\mathbf{x}) = \frac{1}{(2\pi)^{d/2}} \mathrm{e}^{-\frac{1}{2}(\mathbf{x}'-\mathbf{x})^{\mathrm{T}}(\mathbf{x}'-\mathbf{x})}$$

but also

$$\pi(\mathbf{x}|\mathbf{x}') = \frac{1}{(2\pi)^{d/2}} \mathrm{e}^{-\frac{1}{2}(\mathbf{x}-\mathbf{x}')^{\mathrm{T}}(\mathbf{x}-\mathbf{x}')}$$

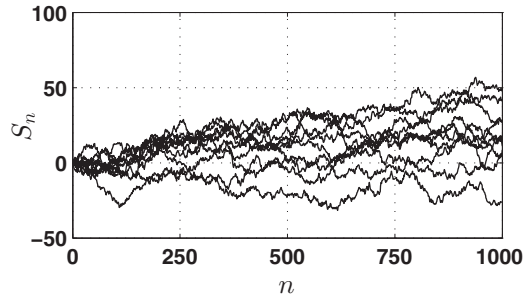

Figure 2.20 Ten trials of a Gaussian random walk.

Definition 2.3.3. *A transition kernel π is reversible relatively to a distribution f if*

$$f(\mathbf{x})\pi(\mathbf{x}'|\mathbf{x}) = f(\mathbf{x}')\pi(\mathbf{x}|\mathbf{x}')$$

Such kernel π is said to be f reversible.

If $\mathbf{X}$ follows a multivariate standard Gaussian distribution, a reversible kernel π relative to $\mathbf{X}$ is given by

$$\pi(\mathbf{x}'|\mathbf{x}) = \sqrt{1-a}\mathbf{x} + \sqrt{a}\mathbf{w}$$

where $\mathbf{W}$ follows a multivariate standard Gaussian distribution and $a \in (0,1)$. Indeed, if we apply π to $\mathbf{X}$, then $\mathbf{X}'$ has a multivariate standard Gaussian pdf and, conversely, if we apply π to $\mathbf{X}'$.

2.3.3.3 The Metropolis–Hastings algorithm

One sometimes might want to generate samples following complex densities that do not belong to the classic law families described in Section 2.2.1. For instance, consider the two-dimensional random vector $\mathbf{X} = (X_1, X_2)^{\mathrm{T}}$ with the following pdf:

$$f(\mathbf{x}) = \frac{|\sin(x_2 - x_1) \times \exp\left(-(x_2^2 + x_1^2)\right)|}{\int_{\mathbb{R}^2} |\sin(x_2 - x_1) \times \exp\left(-(x_2^2 + x_1^2)\right)| \mathrm{d}x_1\, \mathrm{d}x_2}$$

It is difficult to generate directly samples of $\mathbf{X}$, notably because of the integral computation. We can then consider that the pdf f is known up to a constant

$$f(\mathbf{x}) \propto |\sin(x_2 - x_1) \times \exp\left(-(x_2^2 + x_1^2)\right)| \tag{2.2}$$

In fact, a possible solution to generate samples according to the distribution f is to consider the Metropolis–Hastings (MH) algorithm (Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953), a very well-known Monte Carlo Markov chain (MCMC) method. The principle of MH is to build a Markov process $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ that will be approximately distributed, over a long term, that is, when $n \to \infty$ with pdf f. This Markov process is defined with a proposal/refusal method. Indeed, a sample $\mathbf{Z}'$ is proposed from $\mathbf{Z}_n$ with transition kernel π, and then $\mathbf{Z}_{n+1}$ is obtained with

$$\mathbf{z}_{n+1} = \begin{cases} \mathbf{z}', & \text{with probability } \min\left(1, \frac{f(\mathbf{z}')}{f(\mathbf{z}_n)} \times \frac{\pi(\mathbf{z}_n|\mathbf{z}')}{\pi(\mathbf{z}'|\mathbf{z}_n)}\right) \\ \mathbf{z}_n, & \text{otherwise} \end{cases}$$

The MH algorithm consists then of the different iterative stages described in Algorithm 1.

A transition kernel $\pi(\mathbf{z}'|\mathbf{z}_n)$ has to be defined in the MH algorithm. A usual choice of $\pi(\mathbf{z}'|\mathbf{z}_n)$ is the multivariate standard Gaussian distribution. We obtain $\frac{\pi(\mathbf{z}_n|\mathbf{z}')}{\pi(\mathbf{z}'|\mathbf{z}_n)} = 1$,

ALGORITHM 1 The Metropolis–Hastings algorithm

Input: The pdf f, the number of iterations N, the transition kernel π, the starting sample $\mathbf{z}_{start}$, and the burn in period n_b.
Output: The samples $(\mathbf{z}_n)_{n_b \leq n \leq N}$.

Set $\mathbf{z}_0 = \mathbf{z}_{start}$ and $n = 0$.
while $n \leq N$ **do**
  Sample $\mathbf{z}'$ with density $\pi(\mathbf{z}'|\mathbf{z}_n)$.
  Compute the acceptation rate $\alpha = \min\left(1, \frac{f(\mathbf{z}')}{f(\mathbf{z}_n)} \times \frac{\pi(\mathbf{z}_n|\mathbf{z}')}{\pi(\mathbf{z}'|\mathbf{z}_n)}\right)$.
  Sample u with a uniform random variable, $u \sim U_{(0,1)}$.
  if $u \leq \alpha$ **then**
    $\mathbf{z}_{n+1} = \mathbf{z}'$.
  else
    $\mathbf{z}_{n+1} = \mathbf{z}_n$.
  Set $n = n + 1$.
return $(\mathbf{z}_n)_{n_b \leq n \leq N}$.

which corresponds to the special case of the MH algorithm called the *Metropolis algorithm*. It arises when the kernel π is symmetric. If π is an f-reversible kernel, then the proposal $\mathbf{Z}'$ is always accepted.

Even if f is known up to a constant, the ratio $\frac{f(\mathbf{z}')}{f(\mathbf{z}_n)}$ is computable in practice. It is also important to notice that the samples generated with the MH algorithm are not independent. It could be necessary to subsample the population $\mathbf{z}_1, \ldots, \mathbf{z}_n$ to decrease their correlation. Moreover, if the MH starting sample $\mathbf{z}_0$ is not adapted, for instance if $f(\mathbf{z}_0)$ has a low value, it is necessary to define a burn in period n_b because the first samples obtained with MH will not be distributed according to the pdf f.

The MH algorithm has been applied to generate samples according to the pdf f given in Equation (2.2) with a multivariate standard Gaussian distribution and a burn-in period $n_b = 1000$. The results are presented in Figure 2.21. This algorithm enables us to draw samples from the pdf f.

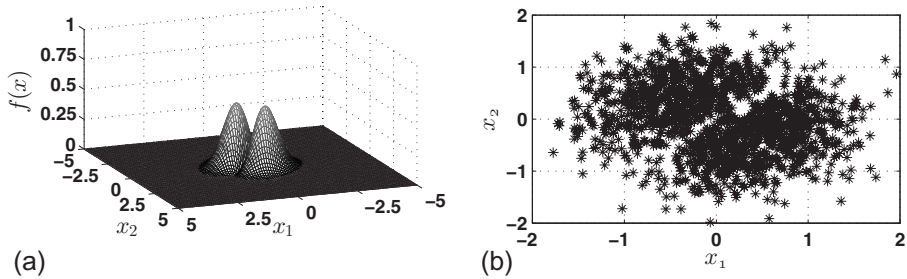

Figure 2.21 Application of MH to sample with the two-dimensional pdf f.
(a) Two-dimensional pdf f. (b) 1000 corresponding samples generated with MH.

2.3.3.4 Transformation of random variables

Some specific estimation methods can be applied only on standard multivariate Gaussian random variables. It might strongly limit the potential application of these methods on realistic cases. Nevertheless, some transformations τ have been proposed to transform a random variable $\mathbf{X}$ into a standard multivariate Gaussian random variable $\mathbf{U} = \tau(\mathbf{X})$. Depending on the available information on the pdf of $\mathbf{X}$, several transformations can be proposed (Hasofer & Lind, 1974; Lebrun & Dutfoy, 2009a, 2009b; Nataf, 1962; Pei-Ling & Kiureghian, 1991; Rosenblatt, 1952). See Table 2.1 for details on the correspondence between assumptions and transformations. We do not review all these transformations in this book, but let us consider a simple example of rv transformation τ using the diagonal transformation. We assume that $\mathbf{X} = (X_1, X_2)^{\mathrm{T}}$ follows a two-dimensional distribution where the pdf f_{X_1} of X_1 and the pdf f_{X_2} of X_2 are given by an exponential law $\mathcal{E}(1)$ and a log-normal law $\ln \mathcal{N}(0,1)$, respectively. F_{X_1} and F_{X_2} are their corresponding cdfs. The components X_1 and X_2 are independent. Then the variable $\mathbf{U} = (U_1, U_2)$ defined by

$$\mathbf{u} = \tau(\mathbf{x}) = \begin{pmatrix} \Phi_{0,1}^{-1}(F_{X_1}(x_1)) \\ \Phi_{0,1}^{-1}(F_{X_2}(x_2)) \end{pmatrix}$$

follows a two-dimensional standard normal distribution where $\Phi_{0,1}^{-1}$ is the inverse cdf of a standard univariate normal distribution. The inverse transformation τ^{-1} from $\mathbf{U}$ to $\mathbf{X}$ is described by

$$\mathbf{x} = \tau^{-1}(\mathbf{u}) = \begin{pmatrix} F_{X_1}^{-1}(\Phi_{0,1}(u_1)) \\ F_{X_2}^{-1}(\Phi_{0,1}(u_2)) \end{pmatrix}$$

where we recall that $F_{X_1}^{-1}$ and $F_{X_2}^{-1}$ are the inverse cdfs of X_1 and X_2, respectively. An application of the transformation τ on 1000 iid samples with the same distribution as $\mathbf{X}$ is proposed in Figure 2.22.

Table 2.1 Possible transformations τ depending on the assumptions on the pdf of X

Assumptions on the pdf of X	Corresponding transformations τ
$\mathbf{X}$ is Gaussian with uncorrelated components	Hasofer–Lind transformation
$\mathbf{X}$ has independent components (not assumed to be Gaussian)	Diagonal transformation
Only the marginal laws of $\mathbf{X}$ and their covariances are known	Nataf transformation
The complete law of $\mathbf{X}$ is known	Rosenblatt transformation

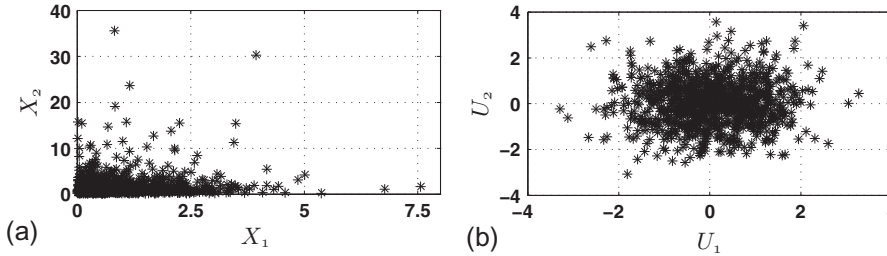

Figure 2.22 Application of diagonal transformation on 1000 iid samples with the same distribution as $\mathbf{X}$. (a) 1000 samples of $\mathbf{X}$. (b) 1000 corresponding samples of $\mathbf{U} = \tau(\mathbf{X})$.

References

Gordon, H. (1997). *Discrete probability*. New York, USA: Springer-Verlag.

Grinstead, C. M., & Snell, J. L. (1998). *Introduction to probability.* Providence, USA: American Mathematical Society.

Hasofer, A., & Lind, N. (1974). An exact and invariant first-order reliability format. *Journal of Engineering Mechanics*, *100*, 111-121.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109.

Jacod, J., & Protter, P. E. (2003). *Probability essentials*. Berlin Heidelberg, Germany: Springer.

Jaworski, P., Durante, F., Haerdle, W., & Rychlik, T. (2010). *Copula theory and its applications: Proceedings of the workshop held in Warsaw, 25–26 September 2009*. Berlin Heidelberg, Germany: Springer.

Kolmogorov, A. N. (1950). *Foundations of the theory of probability*. New York, USA: Chelsea Publishing Co.

Lebrun, R. (2013). *Contributions to stochastic dependence modeling*. Unpublished doctoral dissertation, Université Paris VII, Denis Diderot.

Lebrun, R., & Dutfoy, A. (2009a). A generalization of the Nataf transformation to distributions with elliptical copula. *Probabilistic Engineering Mechanics*, *24*(2), 172-178.

Lebrun, R., & Dutfoy, A. (2009b). An innovating analysis of the Nataf transformation from the copula viewpoint. *Probabilistic Engineering Mechanics*, *24*(3), 312-320.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, *21*, 1087–1092.

Nataf, A. (1962). Distribution des distributions dont les marges sont données (in French). *Comptes rendus de l'Académie des Sciences*, *225*, 42-43.

Nelsen, R. B. (2006). *An introduction to copulas (Springer series in statistics)*. Secaucus, USA: Springer-Verlag New York, Inc.

Pei-Ling, L., & Kiureghian, A. D. (1991). Optimization algorithms for structural reliability. *Structural Safety*, *9*(3), 161-177.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Annals of Mathematical Statistics*, *23*, 470-472.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London, UK: Chapman & Hall.

Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges (in French). *Publications de l'Institut de Statistiques de l'Université de Paris 8*, *8*, 221–231.

Stroock, D. W. (2014). *An introduction to Markov processes*. Berlin, Heidelberg, Germany: Springer.

Tijms, H. C. (2004). *Understanding probability: Chance rules in everyday life*. Cambridge, UK: Cambridge University Press.

Wand, M., & Jones, M. (1995). *Kernel smoothing*. Boca Raton, USA and London, UK and New York, USA: Chapman & Hall/CRC.

Wand, M. P., Marron, J. S., & Ruppert, D. (1991). Transformations in density estimation. *Journal of the American Statistical Association*, *86*, 343-361.

The formalism of rare event probability estimation in complex systems

3

J. Morio, M. Balesdent

3.1 Input–output system

3.1.1 Description

An input–output system can represent industrial aerospace (or other) complex codes (Keane & Nair, 2005) such as a trajectory estimation, structural design, propulsion analysis, computational fluid dynamics (CFD) simulation, and so on. This system is mathematically modeled as an input–output function $\phi(\cdot)$, considered as a "black-box," that is, the behavior of the system can be known only through a (finite) given number of responses of $\phi(\cdot)$. The system is considered as deterministic, so we assume that all the randomness can be described by the aleatory input variables.

Different kinds of elements can be of interest to characterize rare events. The most common is the probability that the output exceeds a given threshold T. Such probabilities are also called *probability of failure* in the system safety literature. Another quantity of interest is the quantile estimation of the system response relative to a given level of probability.

3.1.2 Formalism

Let $\mathbf{X}$ be a d-dimensional random vector with pdf f, and $\phi(\cdot)$ be a continuous scalar function:

$$\phi : \mathbb{R}^d \to \mathbb{R}$$
$$\mathbf{X} = (X_1, \ldots, X_d)^{\mathrm{T}} \to Y = \phi(\mathbf{X})$$

In some cases, the random variables X_i are considered as independent, but this assumption is not mandatory. The unknown pdf of Y is denoted by g. The probability of interest $\mathbb{P}_f$ is the probability that the output of the system exceeds a given threshold T:

Estimation of Rare Event Probabilities in Complex Aerospace and Other Systems
http://dx.doi.org/10.1016/B978-0-08-100091-5.00003-4

$$\begin{aligned}\mathbb{P}_f = \mathbb{P}(Y > T) &= \mathbb{P}(\phi(\mathbf{X}) > T) \\ &= \int_{\mathbb{R}^d} \mathbf{1}_{\phi(\mathbf{x})>T} f(\mathbf{x})\, \mathrm{d}\mathbf{x} \\ &= \int_{\Omega_f} f(\mathbf{x}) \mathrm{d}\mathbf{x}\end{aligned}$$

with Ω_f the failure domain of $\phi(\cdot)$, that is, $\Omega_f = \{\mathbf{x} \in \mathbb{R}^d | \phi(\mathbf{x}) > T\}$. This equation involves the indicator function that is defined as follows:

$$\mathbf{1}_{\phi(\mathbf{x})>T} = \begin{cases} 1, & \text{if } \phi(\mathbf{x}) > T \\ 0, & \text{if } \phi(\mathbf{x}) \leq T \end{cases}$$

Remark 3.1.1. *The probability that the output is underneath the threshold T can also be considered, but for the sake of clarity, we present only the different estimation algorithms with the probability for the output to be above the threshold. Of course, the methods detailed in this book are compatible with the two alternatives.*

In the aerospace field, for instance, the estimation of collision probability between a space debris and a satellite is modeled with this formalism. The model inputs are the uncertain debris and satellite position and speed, and the model output is the minimum distance between the satellite and the debris over a time interval. The function ϕ corresponds to a simulation code of spatial object trajectory propagation. Chapters 5–8 of this book are notably dedicated to the estimation of $\mathbb{P}_f$ for input–output systems from Monte Carlo simulations to more advanced algorithms such as adaptive splitting.

In addition to the probability of exceedance, we might also be interested in estimating some extreme quantiles. In that case, given a level of rare event probability α, we can determine the quantile q_α defined as:

$$q_\alpha = \inf_{v \in \mathbb{R}} \{\mathbb{P}(Y \leq v) \geq \alpha\}$$

Remark 3.1.2. *In some cases, Y can be multidimensional. Generalizations of the quantile notions can be used such as the minimum volume sets as described notably in Pastel, Morio, & Le Gland (2015).*

3.2 Time-variant system

3.2.1 Description

The second model considered in this book corresponds to a time-dependent system on which temporal evolution is modeled by a Markov process or a Markov chain (see Section 2.3.3.1). This kind of approach is of great interest in different scientific and engineering fields such as air traffic management (Bakker & Blom, 1993), fiber optics (Garnier & Del Moral, 2006) and so on. In the same way as input–output systems, the rare event is modeled by a threshold exceedance.

3.2.2 Formalism

Let $(\mathbf{X}_t)_{t\geq 0}$ be a continuous time Markov process in a state space $\mathbb{R}^d$ with initial law $\pi_0(\mathbf{x}_0)$ at time 0 and transition kernels $\pi_{s,t}(\mathbf{x}|\mathbf{x}')$. The quantity of interest is a probability $\mathbb{P}(\mathcal{R})$ with $\mathcal{R}$ defined by

$$\mathcal{R} = \{\mathbf{X}_t \in \mathbf{B}, \text{with } 0 \leq t \leq S\} \tag{3.1}$$

where S is a deterministic stopping time and $\mathbf{B} \subset \mathbb{R}^d$. Without loss of generality, it is assumed that the critical set is defined in terms of threshold exceedance of a real-valued function φ

$$\mathcal{R} = \{\mathbf{X}_t, \text{with } \varphi(\mathbf{X}_t) \geq T \text{ and } 0 \leq t \leq S\}$$

with T a given threshold. Thus, we have

$$\mathcal{R} = \{\mathbf{X}_t, \text{with } \sup_{0\leq t\leq S} \varphi(\mathbf{X}_t) \geq T\}$$

In practice, it is in fact not possible to directly simulate time-continuous Markov processes. The continuous problem can be expressed with a discrete time approach. We then approximate the time-continuous Markov processes by introducing a time discretization such that $t = j\Delta_t$ and $S = c\Delta_t$ where Δ_t is a positive constant as low as possible and $j = 1, \ldots, c$. We then obtain a discrete time Markov chain $(\mathbf{X}_j)_{0\leq j\leq c}$ with initial law $\pi_0(\mathbf{x}_0)$ at time 0 and transition kernel $\pi_j(\mathbf{x}_j|\mathbf{x}_{j-1})$. The estimation of $\mathbb{P}(\mathcal{R})$ in Equation (3.1) is similar in the discrete case to the determination of

$$\mathcal{R} = \{\mathbf{X}_j, \text{with } \varphi(\mathbf{X}_j) \geq T \text{ and } 0 \leq j\Delta_t \leq S\}$$

and thus

$$\mathcal{R} = \{\mathbf{X}_j, \text{with } \sup_{0\leq j\Delta_t\leq S} \varphi(\mathbf{X}_j) \geq T\}$$

Discrete and continuous notations can be used to describe estimation algorithms without any loss of generality.

Chapter 9 is dedicated to the estimation of $\mathbb{P}(\mathcal{R})$ for time-variant systems. For instance, conflict probability between two aircraft in uncontrolled airspace can be estimated with this kind of model. Indeed, a flight plan consists of a sequence of waypoints and speeds along an aircraft trajectory and is designed such that there will be no conflict risk between aircraft. Nevertheless, the trajectory of an aircraft is uncertain due to meteorology, pilot error, and so on. Thus, an aircraft does not follow its flight plan exactly. This uncertainty on the aircraft trajectory is time dependent and can be modeled by a Markov process. In that case, the goal is to estimate the probability that the distance between two aircraft with associated trajectories modeled by a Markov process is below T during a given time interval.

Remark 3.2.1. *The special case of Markov chain that evolves in a countable state space is not considered in this book. Nevertheless, the reference (Rubino & Tuffin, 2009) gives an overview of efficient algorithms for the estimation of rare event probabilities in this case.*

Remark 3.2.2. *A discrete Markov chain can be viewed as an input–output system with dependent inputs. This property is sometimes used when the length of the Markov chain is relatively short since it determines the dimension of the input space. However, in the general case, the rare event estimation methods that are specific to input–output systems cannot be applied in this way because of the length of a general Markov chain.*

In addition to the probability of exceedance, extreme quantiles are also of interest. In that case, given a level of rare event probability α, we determine the quantile q_α defined as

$$q_\alpha = \inf_{v \in \mathbb{R}} \left\{ \mathbb{P}(\{\varphi(\mathbf{X}_j) \leq v, \text{with } 0 \leq j\Delta_t \leq S\}) \geq \alpha \right\}$$

3.3 Characterization of a probability estimation

Most of the different estimation methods M described in this book propose an estimate $\hat{\mathbb{P}}^M$ of the probability $\mathbb{P}_f$ when this quantity is rare relative to the available simulation budget N, that is, when $\mathbb{P}_f << \frac{1}{N}$. Because $\hat{\mathbb{P}}^M$ is a random variable, we thus analyze its expectation $\mathbb{E}(\hat{\mathbb{P}}^M)$. In practice, we apply $\tilde{N}$ times a given method M on the system and thus obtain $\tilde{N}$ values of probability estimate, $\mathbb{P}_i^M$, $i = 1, \ldots, \tilde{N}$. The expectation $\mathbb{E}(\hat{\mathbb{P}}^M)$ is then determined by the mean of these $\tilde{N}$ probability estimates,

$$\mathbb{E}(\hat{\mathbb{P}}^M) \approx \hat{\mathbb{P}}^M_{\tilde{N}} = \frac{1}{\tilde{N}} \sum_{i=1}^{\tilde{N}} \mathbb{P}_i^M \tag{3.2}$$

It is mandatory to characterize the efficiency of an estimate $\hat{\mathbb{P}}^M$ of $\mathbb{P}_f$. In the case of rare event probability estimation, possible indicators of this estimation efficiency are notably its relative bias and its relative deviation. The relative bias describes how close the probability estimate $\hat{\mathbb{P}}^M$ is to $\mathbb{P}_f$ and is defined as

$$\mathrm{RB}(\hat{\mathbb{P}}^M) = \frac{\mathbb{E}(\hat{\mathbb{P}}^M) - \mathbb{P}_f}{\mathbb{P}_f}$$

In practice, for a given simulation method M, $\mathbb{E}(\hat{\mathbb{P}}^M)$ is approximated with Equation (3.2). The term $\mathrm{RB}(\hat{\mathbb{P}}^M)$ is not computable if we have no idea of the target probability value $\mathbb{P}_f$. The relative bias is also positive if $\hat{\mathbb{P}}^M$ tends to overestimate $\mathbb{P}_f$ and vice versa. The relative standard deviation or relative standard error RE of an estimator $\hat{\mathbb{P}}^M$ of $\mathbb{P}_f$ is given by the following ratio:

$$\mathrm{RE}(\hat{\mathbb{P}}^M) = \frac{\sqrt{\mathbb{V}(\hat{\mathbb{P}}^M)}}{\mathbb{E}(\hat{\mathbb{P}}^M)}$$

The parameter $\sqrt{\mathbb{V}(\hat{\mathbb{P}}^M)}$ is estimated by $\hat{\sigma}_{\hat{\mathbb{P}}^M_{\tilde{N}}}$ in the following way:

$$\hat{\sigma}_{\hat{\mathbb{P}}^M_{\tilde{N}}} = \sqrt{\frac{1}{\tilde{N}-1}\sum_{i=1}^{\tilde{N}}\left(\mathbb{P}_i^M - \hat{\mathbb{P}}^M_{\tilde{N}}\right)^2}$$

for a given simulation method M proposed in this book. The relative error is said bounded if $\mathrm{RE}(\hat{\mathbb{P}}^M)$ remains bounded when $\mathbb{P}_f \longrightarrow 0$ (L'Ecuyer, Blanchet, Tuffin, & Glynn, 2010; L'Ecuyer, Mandjes, & Tuffin, 2009). In that case, the number of samples needed to get a specified relative error is bounded whatever the rarity of the event of interest is. The logarithmic efficiency can also be defined. An unbiased estimator $\hat{\mathbb{P}}^M$ of $\mathbb{P}_f$ is said to be logarithmic efficient (L'Ecuyer et al., 2010; L'Ecuyer et al., 2009), if

$$\mathrm{LE}(\hat{\mathbb{P}}^M) = \lim_{\mathbb{P}_f \to 0} \frac{\ln(\mathbb{E}((\hat{\mathbb{P}}^M)^2))}{\ln(\mathbb{P}_f)} = 2$$

Logarithmic efficiency is a necessary but not sufficient condition for bounded relative error. Characterizing the rare event probability estimate with these concepts is very important even if they are often difficult to prove in the general case.

Moreover, if a simulation method M enables to estimate the probability $\hat{\mathbb{P}}^M$ for a given test case with relative error $\mathrm{RE}(\hat{\mathbb{P}}^M)$ and simulation budget N^M, we define the efficiency ν^M of this estimate relatively to crude Monte Carlo (CMC) estimate by the following ratio:

$$\nu^M = \frac{N^{\mathrm{CMC}}}{N^M}$$

where N^{CMC} is the number of required CMC samples to obtain $\mathrm{RE}(\hat{\mathbb{P}}^{\mathrm{CMC}})$=$\mathrm{RE}(\hat{\mathbb{P}}^M)$. If $\nu^M > 1$, then the method M is more efficient than CMC for the given test case. The computation of ν^M is of interest only when the probability estimate $\hat{\mathbb{P}}^M$ is not too far from $\mathbb{P}_f$.

References

Bakker, G. J., and Blom, H. A. P. (1993). Air traffic collision risk modelling. In *Proceedings of the 32nd IEEE conference on Decision and Control, 1993* (Vol. 2, p. 1464–1469).

Garnier, J., and Del Moral, P. (2006). Simulations of rare events in fiber optics by interacting particle systems. *Optics Communications*, *267*(1), 205–214.

Keane, A. J., and Nair, P. B. (2005). *Computational approaches for aerospace design: The pursuit of excellence*. Chichester, UK, Hoboken, USA: John Wiley & Sons.

L'Ecuyer, P., Blanchet, J. H., Tuffin, B., and Glynn, P. W. (2010). Asymptotic robustness of estimators in rare-event simulation. *ACM Transactions on Modeling and Computer Simulation*, *20*(1), 1–37.

L'Ecuyer, P., Mandjes, M., and Tuffin, B. (2009). Importance sampling in rare event simulation. In G. Rubino and B. Tuffin (Eds.), *Rare event simulation using Monte Carlo methods* (p. 17-38). New York: John Wiley & Sons, Ltd.

Pastel, R., Morio, J., and Le Gland, F. (2015). Extreme density level set estimation for input–output functions via the adaptive splitting technique. *Journal of Computational Science*, *6*, 40–46.

Rubino, G., and Tuffin, B. (2009). Markovian models for dependability analysis. In G. Rubino and B. Tuffin (Eds.), *Rare event simulation using Monte Carlo methods* (pp. 125–143). New York: John Wiley & Sons, Ltd.

Part Two

Practical Overview of the Main Rare Event Estimation Techniques

This page intentionally left blank

Introduction

M. Balesdent, J. Morio

4

4.1 Categories of estimation methods

This book assumes the following classification of the different rare event probability estimation techniques for input–output systems into four main categories:

- The simulation techniques that consist of several ways of efficiently sampling the input to decrease the probability estimate variance
- The statistical methods that enable the derivation of a probability estimate of output threshold exceedance from a fixed set of output samples
- The reliability-based approaches that take advantage of geometrical considerations on the function ϕ to estimate the rare event probability, sometimes with sampling
- Metamodeling and sensitivity analysis, which are of interest for characterizing rare events of high dimensional and time-consuming systems

A chapter of this book is devoted to each class of estimation method. The book gives the principle and the detailed algorithm to ease the implementation of each technique proposed and illustrates its performances on different toy cases. Depending on these results, we consider its potential application on the more complex use cases in Part Three.

The estimation of rare event probabilities corresponds to an integral computation. Nevertheless, numerical integration methods such as Gaussian quadrature (Novak, 1988) or sparse grids (Gerstner & Griebel, 2003; Smolyak, 1963) are not developed in this book because they are not adapted to the estimation of rare event probabilities. These techniques require some smoothness on the function to integrate whereas rare event probabilities are expressed with an indicator function. The application of some numerical integration methods can thus lead to significant errors in practice.

4.2 General notations

Because the same formalism will be used for all the estimation methods, let us recall the notations of Chapter 3 and consider a d-dimensional input random vector $\mathbf{X}$ with a joint pdf f, a continuous scalar function $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$ and a threshold T. The different components of $\mathbf{X}$ will be denoted $\mathbf{X} = (X_1, X_2, \ldots, X_d)^{\mathrm{T}}$. The function ϕ is static, that is, does not depend on time and represents, for instance, an input–output model. We assume that the output $Y = \phi(\mathbf{X})$ is a scalar random variable. The pdf of Y is denoted by g.

Estimation of Rare Event Probabilities in Complex Aerospace and Other Systems
http://dx.doi.org/10.1016/B978-0-08-100091-5.00004-6

4.3 Description of the toy cases

It is assumed in the different toy cases that $\mathbf{X}$ follows a multivariate standard normal distribution for the sake of simplicity but without any loss of generality because we recall that several transformations can be applied on the input distribution (see Section 2.3.3.4) to come down to a multivariate standard normal distribution for the input.

4.3.1 Identity function

The identity function is a one-dimensional toy case where $Y = X$. The input variable X is distributed according to a normal distribution $\mathcal{N}(0, 1)$. Even if this toy case does not bring any information on the algorithm performance in a general situation, it enables us to present the principle of some specific simulation algorithms. With a threshold T equal to 3, 4, and 5, the theoretical probability of exceedance (expressed as $\mathbb{P}(Y > T) = 1 - \Phi_{0,1}(T)$) is given by 1.35×10^{-3}, 3.17×10^{-5}, and 2.87×10^{-7}, respectively .

4.3.2 Polynomial square root function

This test case is a two-dimensional function and is illustrated in Figure 4.1. The input space is $\mathbb{R}^2$. The input variables are distributed according to a normal distribution $\mathcal{N}(\mathbf{0}_2, \mathbf{I}_2)$ where $\mathbf{I}_2$ refers to the identity matrix of size 2.

$$T = 6$$
$$\mathbf{X} \sim \mathcal{N}(\mathbf{0}_2, \mathbf{I}_2)$$
$$\phi(x_1, x_2) = -\sqrt{(-x_1 + 10)^2 + (x_2 + 7)^2 + 10 * (x_1 + x_2)^2} + 14.$$

This toy case allows us to illustrate the ability of a method to estimate a rare event probability in a relatively simple case. For a threshold T equal to 6, the estimated probability of exceedance, obtained with a huge CMC, is 2.35×10^{-6}.

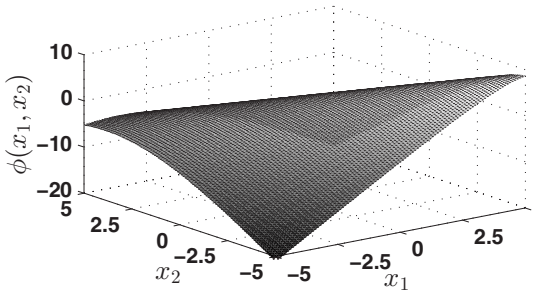

Figure 4.1 Polynomial square root function.

4.3.3 Four-branch system

The four-branch system consists of a two-dimensional test case. The input space is $\mathbb{X} = \mathbb{R}^2$. The input variables are distributed according to a normal distribution $\mathcal{N}(\mathbf{0}_2, \mathbf{I}_2)$.

$$T = 10$$
$$\mathbf{X} \sim \mathcal{N}(\mathbf{0_2}, \mathbf{I_2})$$
$$\phi(x_1, x_2) = 10 - \min \begin{cases} 3 + 0.1(x_1 - x_2)^2 - \frac{x_1 + x_2}{\sqrt{2}} \\ 3 + 0.1(x_1 - x_2)^2 + \frac{x_1 + x_2}{\sqrt{2}} \\ (x_1 - x_2) + \frac{7}{\sqrt{2}} \\ (x_2 - x_1) + \frac{7}{\sqrt{2}} \end{cases}$$

This toy case allows us to test the ability of the rare event estimation methods to accurately estimate the probability in the case of disconnected failure region Ω_f. This problem involves four zones for which $\phi(\mathbf{X}) > T$ (Figure 4.2). With a threshold T equal to 10, the estimated probability of exceedance obtained with a huge CMC is 2.22×10^{-3}. The estimated probability of exceedance with $T = 12$ is equal to 1.18×10^{-6}.

4.3.4 Polynomial product function

The polynomial product test case is a d-dimensional function and is adapted from the Styblinski–Tang function used in optimization. The input variables are distributed according to a normal distribution $\mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$.

$$\mathbf{X} \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$$
$$\phi(\mathbf{x}) = \frac{1}{2} \sum_{i=1}^{d} (x_i^4 + x_i^2 + 5x_i)$$

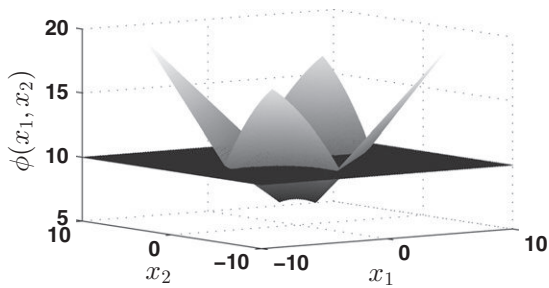

Figure 4.2 Four-branch system with threshold ($T = 10$).

Table 4.1 Toy cases based on the polynomial product function

d	T	$\mathbb{P}(\phi(\mathbf{X}) > T)$
5	400	8.44×10^{-7}
20	500	1.09×10^{-6}
50	700	3.56×10^{-7}
200	1000	4.85×10^{-6}

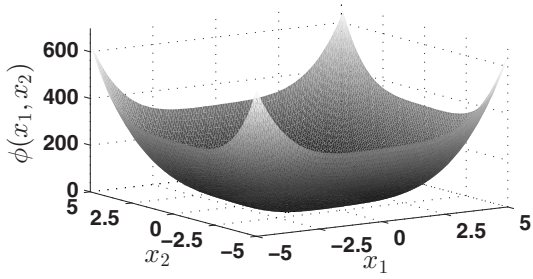

Figure 4.3 Polynomial product function (for $d = 2$).

This toy case is useful to evaluate the ability of the methods to cope with high dimensional problems. The different combinations of dimensions and thresholds used in this book to evaluate the rare event estimation methods are summarized in Table 4.1. The polynomial product function for $d = 2$ is illustrated in Figure 4.3.

References

Gerstner, T., and Griebel, M. (2003). Dimension-adaptive tensor-product quadrature. *Computing*, *71*(1), 65–87.

Novak, E. (1988). *Deterministic and stochastic error bounds in numerical analysis* (Vol. 1349). Berlin, Germany: Springer.

Smolyak, S. (1963). Quadrature and interpolation formulas for tensor products of certain classes of functions. *Soviet Mathematics, Doklady*, *4*, 240–243.

Simulation techniques

M. Balesdent, J. Morio, C. Vergé, R. Pastel

5

Simulation techniques consist in sampling the input and characterizing the uncertainty of the corresponding output. This is notably the case of the crude Monte Carlo method that is well suited to characterize events whose associated probabilities are not too low with respect to the simulation budget. However, for very seldom observed events, this approach does not lead to accurate results. In this chapter, we review different simulation techniques such as importance sampling or adaptive splitting that estimate rare event probabilities with a low variance and a reduced number of samples. These methods seek to generate more samples of $\mathbf{X}$ so that $\phi(\mathbf{X}) > T$.

5.1 Crude Monte Carlo

5.1.1 Principle

A simple way to estimate a probability is to consider crude Monte Carlo (CMC) (Fishman, 1996; Kroese & Rubinstein, 2012; Mikhailov, 1999; Niederreiter & Spanier, 2000; Robert & Casella, 2005; Silverman, 1986; Sobol, 1994). For that purpose, we generate N iid samples $\mathbf{X}_1, \ldots, \mathbf{X}_N$ with the joint pdf f of $\mathbf{X}$ and compute their outputs with the function ϕ : $\phi(\mathbf{X}_1), \ldots, \phi(\mathbf{X}_N)$. The probability $\mathbb{P}_f$ is then estimated by

$$\hat{\mathbb{P}}^{\text{CMC}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\phi(\mathbf{X}_i)>T}$$

where $\mathbf{1}_{\phi(\mathbf{X}_i)>T}$ is equal to 1 if $\phi(\mathbf{X}_i) > T$ and 0 otherwise. The CMC method for probability estimation is described in Algorithm 2.

This estimation converges almost surely to the target probability $\mathbb{P}_f$ as a consequence of the law of large numbers (see Section 2.3.1). There is no regularity condition concerning $\phi(\cdot)$ in order to apply the law of large numbers on this estimator, but $\mathbb{V}(\mathbf{1}_{\phi(\mathbf{X})>T}) < +\infty$ is required. The relative standard error of the estimator $\hat{\mathbb{P}}^{\text{CMC}}$ is provided by (Kroese & Rubinstein, 2012; Silverman, 1986)

$$\text{RE}(\hat{\mathbb{P}}^{\text{CMC}}) = \frac{\sigma_{\hat{\mathbb{P}}^{\text{CMC}}}}{\mathbb{P}_f} = \frac{1}{\sqrt{N}} \frac{\sqrt{\mathbb{P}_f - \mathbb{P}_f^2}}{\mathbb{P}_f} \tag{5.1}$$

The CMC convergence speed depends only on N and $\mathbb{P}_f$ whatever the dimension d of the input space is. Considering rare event probability estimation, that is, when $\mathbb{P}_f$ takes low values, we obtain

Estimation of Rare Event Probabilities in Complex Aerospace and Other Systems
http://dx.doi.org/10.1016/B978-0-08-100091-5.00005-8

$$\lim_{\mathbb{P}_f \to 0} \frac{\sigma_{\hat{\mathbb{P}}^{\text{CMC}}}}{\mathbb{P}_f} = \lim_{\mathbb{P}_f \to 0} \frac{1}{\sqrt{N\mathbb{P}_f}} = +\infty$$

The relative error is consequently unbounded. For instance, to estimate a probability $\mathbb{P}_f$ of the order of 10^{-4} with a 10% relative error, at least 10^6 samples are required. The simulation budget is thus an issue when the computation time required to obtain a sample $\phi(\mathbf{X}_i)$ is not negligible.

ALGORITHM 2 Crude Monte Carlo simulations for probability estimation

Input: The pdf f, the number of samples N, the function $\phi(\cdot)$, and the threshold T.
Output: The probability estimate $\hat{\mathbb{P}}^{\text{CMC}}$.

1 Sample $\mathbf{X}_1, \ldots, \mathbf{X}_N$ with density f.
2 Apply ϕ to $\mathbf{X}_1, \ldots, \mathbf{X}_N$ to determine the samples $\phi(\mathbf{X}_1), \ldots, \phi(\mathbf{X}_N)$.
3 Estimate $\hat{\mathbb{P}}^{\text{CMC}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\phi(\mathbf{X}_i) > T}$.
4 **return** $\hat{\mathbb{P}}^{\text{CMC}}$.

In addition to the probability of exceedance, it is also possible to estimate some quantiles of $\phi(\mathbf{X})$ with CMC. In that case, given a level of probability α, we can estimate the corresponding quantile with:

$$\hat{q}_\alpha^{\text{CMC}} = \inf_{y \in \mathbb{R}} \left\{ \hat{F}_N^{\text{CMC}}(y) \geq \alpha \right\}$$

where $\hat{F}_N^{\text{CMC}}(y)$ is the empirical cdf defined by

$$\hat{F}_N^{\text{CMC}}(y) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\phi(\mathbf{X}_i) \leq y}$$

It can be shown that $\hat{q}_\alpha^{\text{CMC}}$ is a biased estimate of q_α, but this bias is negligible compared to $\frac{1}{N}$, and the variance of $\hat{q}_\alpha^{\text{CMC}}$ can be approximated by (Arnold, Balakrishnan, & Nagaraja, 1992)

$$\mathbb{V}\left(\hat{q}_\alpha^{\text{CMC}}\right) \approx \frac{1-\alpha}{N f^2(q_\alpha)} \tag{5.2}$$

5.1.2 Application on a toy case

Four-branch system

CMC is applied to the four-branch system. An illustration of CMC sampling is given in Figure 5.1. The probability estimates obtained with CMC for this toy case are provided in Table 5.1 for different thresholds. Even with a budget of 10^6 samples, that is, a number of samples that is often not available in realistic applications, the relative

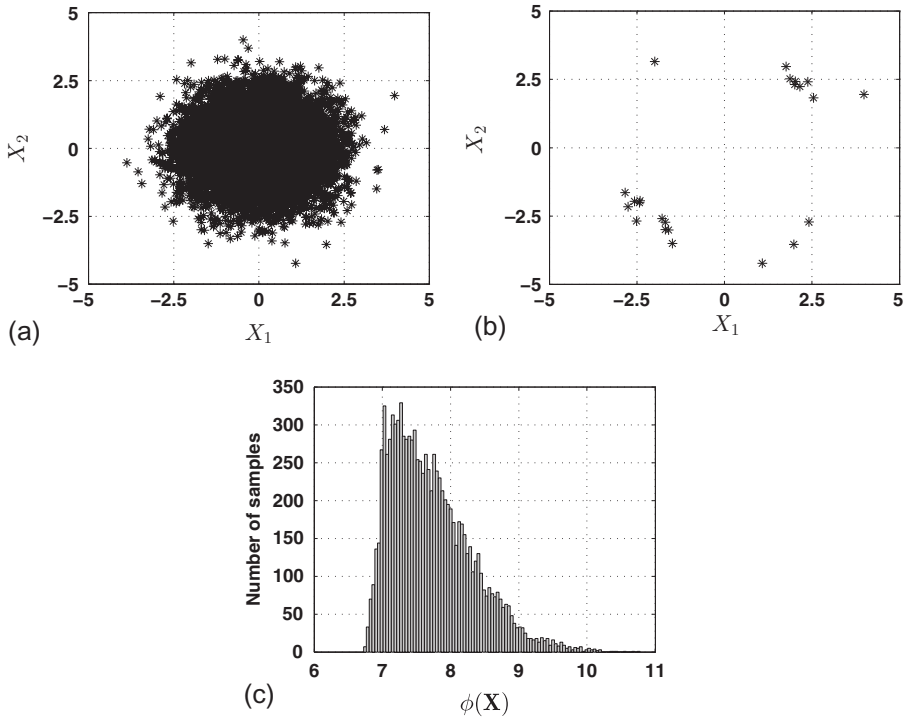

Figure 5.1 Probability estimation with 10,000 CMC samples: (a) CMC samples, (b) CMC samples above $T = 10$, and (c) output histogram.

Table 5.1 Results of CMC for the four-branch system

T	$\hat{\mathbb{P}}^{\text{CMC}}$	RB ($\hat{\mathbb{P}}^{\text{CMC}}$)	RE ($\hat{\mathbb{P}}^{\text{CMC}}$)	**Simulation budget**
10	2.3×10^{-3}	1%	22%	10^4
10	2.2×10^{-3}	−0.1%	2%	10^6
12	2.0×10^{-6}	69%	704%	10^4
12	1.0×10^{-6}	−11%	88%	10^6

error on $\mathbb{P}(\phi(\mathbf{X}) > 12)$ reaches a high level. Indeed, CMC is not able to generate enough samples $\mathbf{X}_i$ such that $\phi(\mathbf{X}_i) > T$.

5.1.3 Conclusion

The implementation of CMC on input–output functions is simple, but CMC requires a significant simulation budget to estimate a low probability with accuracy. Nevertheless, it will be interesting to apply CMC to the use cases of part Three because CMC can be seen as a reference. The main characteristics of CMC are summarized in Table 5.2.

Table 5.2 Summary of CMC rare event probability estimation

Criteria	CMC characteristics
Rarity of $\mathbb{P}_f$	Not adapted; slow convergence
Simulation budget	$\frac{10^2}{\mathbb{P}_f}$ required samples to estimate $\mathbb{P}_f$ with a 10% relative error
Type of function $\phi(\cdot)$	No regularity condition on $\phi(\cdot)$
Multimodality of f restricted to Ω_f	Without any influence
Dimension of $\mathbf{X}$	No influence on the convergence
Probability estimate error	Estimation with analytical formula depending on $\mathbb{P}_f$ and N or with retrials
Difficulty of implementation	None

The following sections describe different available alternatives to CMC to improve the probability estimations, reduce the number of required samples, and increase the estimation accuracy (i.e., decrease the relative error).

5.2 Simple variance reduction techniques

5.2.1 Quasi-Monte Carlo

The principle of quasi-Monte Carlo (QMC) (Niederreiter & Spanier, 2000) is to replace the random number sequences of the Monte Carlo method by deterministic number sequences $\mathbf{X}_1, \ldots, \mathbf{X}_N$, which have the property to be better uniformly distributed in the sense that they are more equidistributed. Figure 5.2 compares the sample distribution obtained with CMC samples and a Halton sequence (Niederreiter & Spanier, 2000) to sample a unit square with a uniform law. QMC reduces the variance of CMC and, depending on the deterministic number sequence used, an asymptotic variance can be determined. It can be applied in a low-moderate dimension and is a good compromise between numerical integration methods, such as Gaussian

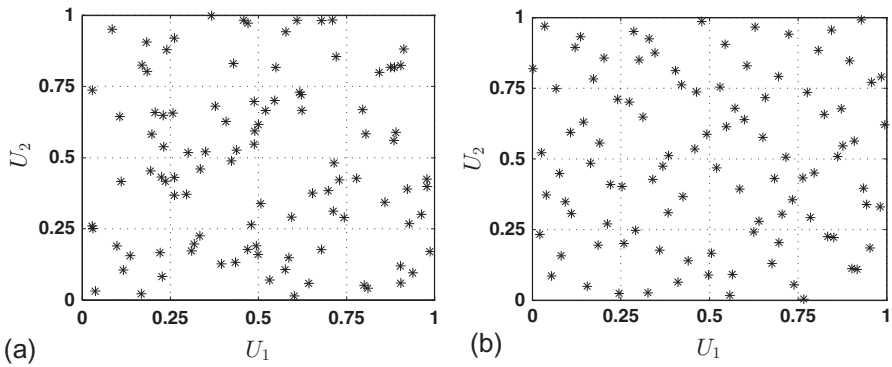

Figure 5.2 (a) 100 CMC samples and (b) 100 Halton sequence samples.

quadrature, and CMC. Nevertheless, QMC is not adapted for the estimation of a rare event probability because it does not focus on the tail of the probability distributions but on filling the gap that appears in CMC samples.

5.2.2 Conditional Monte Carlo

5.2.2.1 Principle

The idea of conditional Monte Carlo (Glasserman, 2003) is to decompose the input vector into $\mathbf{X} = (\mathbf{W}, \mathbf{Z})^T$ such that the conditional expectation $\mathbb{E}(\mathbf{X}|\mathbf{W})$ is known analytically. The probability $\mathbb{P}_f$ can be then expressed in the following way:

$$\mathbb{P}_f = \int_{\mathbb{R}^d} \mathbf{1}_{\phi(\mathbf{x})>T} f(\mathbf{x})\, \mathrm{d}\mathbf{x} = \int\int \mathbf{1}_{\phi(\mathbf{w},\mathbf{z})>T} f(\mathbf{w},\mathbf{z})\, \mathrm{d}\mathbf{w}\, \mathrm{d}\mathbf{z}$$

It can then be developed using conditional density and probability with

$$\begin{aligned}\mathbb{P}_f &= \int\int \mathbf{1}_{\phi(\mathbf{w},\mathbf{z})>T} f_{\mathbf{Z}|\mathbf{W}}(\mathbf{w},\mathbf{z}) f_{\mathbf{W}}(\mathbf{w})\, \mathrm{d}\mathbf{w}\, \mathrm{d}\mathbf{z}, \\ &= \int \mathbb{P}\left(\phi(\mathbf{W},\mathbf{Z}) > T|\mathbf{W} = \mathbf{w}\right) f_{\mathbf{W}}(\mathbf{w})\, \mathrm{d}\mathbf{w}\end{aligned}$$

where $f_{\mathbf{W}}$ is the marginal distribution of $\mathbf{W}$ and $f_{\mathbf{Z}|\mathbf{W}}$ is defined by

$$f_{\mathbf{Z}|\mathbf{W}}(\mathbf{w},\mathbf{z}) = \begin{cases} \frac{f(\mathbf{w},\mathbf{z})}{f_{\mathbf{W}}(\mathbf{w})} & \text{if } f_{\mathbf{W}}(\mathbf{w}) > 0, \\ 0 & \text{otherwise} \end{cases}$$

If $\mathbb{P}\left(\phi(\mathbf{W},\mathbf{Z}) > T|\mathbf{W}\right)$ is computable, we can then derive a Monte Carlo estimator of the probability $\mathbb{P}_f$ with

$$\hat{\mathbb{P}}^{\text{Cond MC}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{P}\left(\phi(\mathbf{W}_i,\mathbf{Z}) > T|\mathbf{W}_i\right)$$

where $\mathbf{W}_i$ are iid samples generated from the distribution of $\mathbf{W}$. The variance of $\hat{\mathbb{P}}^{\text{Cond MC}}$ is always lower than the CMC probability estimate variance. Indeed, with the law of total variance, we have:

$$\mathbb{V}\left(\hat{\mathbb{P}}^{\text{CMC}}\right) = \mathbb{V}\left(\hat{\mathbb{P}}^{\text{Cond MC}}\right) + \mathbb{E}\left(\mathbb{V}\left(1_{\phi(\mathbf{W},\mathbf{Z})>T}|\mathbf{W}\right)\right)$$

and thus

$$\mathbb{V}\left(\hat{\mathbb{P}}^{\text{Cond MC}}\right) \leq \mathbb{V}\left(\hat{\mathbb{P}}^{\text{CMC}}\right)$$

Nevertheless, the efficiency of this approach strongly depends on the decomposition of $\mathbf{X}$.

5.2.2.2 Conclusion

This method will not be applied in this book because we consider a very general failure function. Nevertheless, in aircraft design, a good approximation of the failure region is of the form $\mathbf{1}_{R(\mathbf{W})>C(\mathbf{Z})}$ where R is the response and C is the capacity. We run into it with stress calculations that may require great deal of computation for $R(\mathbf{W})$, where $\mathbf{W}$ mostly depends on load and geometry randomness, while the strength randomness $C(\mathbf{Z})$ is cheap to calculate and depends on material variability. Conditional Monte Carlo can then strongly decrease the CMC variance in this case (Smarslok, Haftka, Carraro, & Ginsbourger, 2010).

5.2.3 Control variates

5.2.3.1 Principle

The control variate (CV) method (Glasserman, 2003; Meyn, 2007) is a variance reduction technique based on the correlation of random variables. Let us define the random variable $H = \mathbf{1}_{\phi(\mathbf{X})>T}$ and a random variable M such that $\mathbb{E}(M) = \tau$ with $\tau \in \mathbb{R}$. M is assumed to be correlated with H. The quantity of interest is, of course, $\mathbb{P}_f = \mathbb{E}(H)$. We can also derive the rv H^* so that, given a coefficient c,

$$H^* = H + c(M - \tau)$$

H^* is an unbiased estimator of $\mathbb{P}_f$ for any choice of the coefficient c and we have thus $\mathbb{E}(H^*) = \mathbb{P}_f$. The variance of H^* is given by

$$\mathbb{V}(H^*) = \mathbb{V}(H) + c^2\mathbb{V}(M) + 2c\,\mathrm{Cov}(H, M)$$

It can be shown that choosing the optimal coefficient c^* defined by

$$c^* = \frac{-\mathrm{Cov}(H, M)}{\mathbb{V}(M)}$$

minimizes the variance of H^*. In that case, the variance H^* is equal to

$$\mathbb{V}(H^*) = (1 - \rho^2)\mathbb{V}(H)$$

where ρ is the correlation coefficient between H and M. Unfortunately, the optimal coefficient c^* is not often available analytically; thus, different techniques allow to choose efficient values of c. The most obvious technique is to estimate $\mathrm{Cov}(H, M)$ and $\mathbb{V}(M)$ with Monte Carlo simulations to approximate c^*. A general algorithm describing the CV principle is presented in Algorithm 3. The use of CV is difficult in practice for general complex systems. Indeed, finding an rv M that has a sufficient correlation with H is not obvious. A possible alternative is to consider a surrogate model (see Section 8.2) to generate samples of M (Cannamela, Garnier, & Iooss, 2008).

ALGORITHM 3 Control variate simulation algorithm

Input: The pdf f, the number of samples N, the function $\phi(\cdot)$, the threshold T, and a pdf h.

Output: The probability estimate $\hat{\mathbb{P}}^{\text{CV}}$.

1 Sample $\mathbf{X}_1, \ldots, \mathbf{X}_N$ with density f.
2 Apply ϕ to $\mathbf{X}_1, \ldots, \mathbf{X}_N$ to determine the samples $\phi(\mathbf{X}_1), \ldots, \phi(\mathbf{X}_N)$.
3 Sample $M_1, \ldots, M_N$ with density h such that they have a non-null covariance with $\mathbf{1}_{\phi(\mathbf{X}_1)>T}, \ldots, \mathbf{1}_{\phi(\mathbf{X}_N)>T}$.
4 Compute $\hat{Y}_N = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\phi(\mathbf{X}_i)>T}$ and $\hat{M}_N = \frac{1}{N}\sum_{i=1}^{N} M_i$.
5 Estimate $\hat{c}^*$ with $\hat{c}^* = \frac{-\left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\phi(\mathbf{X}_i)>T} M_i - \hat{M}_N \hat{Y}_N\right)}{\frac{1}{N}\sum_{i=1}^{N} M_i^2 - \hat{M}_N^2}$.
6 Set $\forall i \in \{1, \ldots, N\}$, $H_i^* = \mathbf{1}_{\phi(\mathbf{X}_i)>T} + \hat{c}^*(M_i - \hat{M}_N)$.
7 Estimate $\hat{\mathbb{P}}^{\text{CV}} = \frac{1}{N}\sum_{i=1}^{N} H_i^*$.
8 **return** $\hat{\mathbb{P}}^{\text{CV}}$

Table 5.3 Results of CV for the four-branch system toy case

T	$\hat{\mathbb{P}}^{\text{CV}}$	RB ($\hat{\mathbb{P}}^{\text{CV}}$)	RE ($\hat{\mathbb{P}}^{\text{CV}}$)	**Simulation budget**	ν^{CV}
10	2.2×10^{-3}	-2%	17%	10,000	1.3
10	2.2×10^{-3}	0.8%	6%	100,000	1.2

5.2.3.2 Application on a toy case

Four-branch system

Control variate algorithm is applied to the four-branch system. For that purpose, the variable M is defined as $M = \mathbf{1}_{||\mathbf{X}||>3.25}$ where $||\cdot||$ is the Euclidean norm. It can be shown that $\mathbb{E}(M) = \exp\left(-\frac{(3.25)^2}{2}\right)$. Probability estimation results with control variates are provided in Table 5.3. CV enables to slightly decrease the variance of CMC without increasing the number of calls to the function $\phi(\cdot)$. However, the efficiency of CV strongly depends on the choice of M.

5.2.3.3 Conclusion

The direct application of control variates is complicated in the general case because of the choice of the random variable M since this variable must be correlated with $\mathbf{1}_{\phi(\mathbf{X})>T}$. Joint use of surrogate models and CV is often advised. However, even with a surrogate model, CV is not very useful for estimating the probability of a rare event because the surrogate model must be accurate near the input region $\{\mathbf{x}|\phi(\mathbf{x}) = T\}$ that is not known *a priori*. The summary of CV characteristics is given in Table 5.4.

Table 5.4 Summary of CV for rare event probability estimation

Criteria	CV characteristics
Rarity of $\mathbb{P}_f$	Increase the required correlation between H and M
Simulation budget	At least 10^3 samples in most cases
Type of function $\phi(\cdot)$	No direct influence
Multimodality of f restricted to $\mathbb{P}_f$	No direct influence
Dimension of $\mathbf{X}$	No direct influence
Probability estimate error	Estimation with retrials
Difficulty of implementation	Low; difficult to determine an efficient rv M

5.2.4 Antithetic variates

5.2.4.1 Principle

The antithetic variate (AV) method (Hammersley & Handscomb, 1964; Sarkar & Prasad, 1992) is also a variance reduction technique. To describe the principles of this method, let us consider the two random variables H_1 and H_2 with the same probability law as $H = \mathbf{1}_{\phi(\mathbf{X})>T}$. Then

$$\mathbb{E}(H) = \frac{1}{2}(\mathbb{E}(H_1) + \mathbb{E}(H_2)) = \mathbb{E}\left(\frac{H_1 + H_2}{2}\right)$$

and

$$\mathbb{V}\left(\frac{H_1 + H_2}{2}\right) = \frac{\mathbb{V}(H_1) + \mathbb{V}(H_2) + 2\mathrm{Cov}(H_1, H_2)}{4}$$

If H_1 and H_2 are iid, then $\mathrm{Cov}(H_1, H_2) = 0$, and we obtain the same variance as the CMC estimate. The principle of AV is to obtain samples so that $\mathrm{Cov}(H_1, H_2) < 0$. The probability estimation is then obtained with CMC.

Nevertheless, AV cannot be easily applied on a complex system. Indeed, it is not obvious to propose $\mathbf{X}$ and $\mathbf{X}'$ such that

$$\mathrm{Cov}(\mathbf{1}_{\phi(\mathbf{X})>T}, \mathbf{1}_{\phi(\mathbf{X}')>T}) < 0$$

where $\mathbf{X}$ and $\mathbf{X}'$ have the same distribution.

In simple cases, AV can still be useful. If X follows a one-dimension normal pdf with mean μ and standard deviation σ, then $X' = 2\mu - X$ follows the same law as X. The variables $\mathbf{1}_{\phi(X)>T}$ and $\mathbf{1}_{\phi(X')>T}$ are then antithetic variates. See the general simulation algorithm for AV in Algorithm 4. Dagger sampling described in Kumamoto, Tanaka, Inoue, & Henley (1980) and more recently in Rongfu, Chanan, Lin, & Yuanzhang (2010) is an extension of the antithetic variate method. It improves CMC estimation for specific systems such as networks or fault trees.

ALGORITHM 4 Antithetic variate simulation algorithm

Input: The pdf f, the number of samples N assuming that N is even, the function $\phi(\cdot)$, and the threshold T.

Output: The probability estimate $\hat{\mathbb{P}}^{\text{AV}}$.

1 Sample $\mathbf{X}_1, \ldots, \mathbf{X}_{N/2}$ with density f.

2 Apply ϕ to $\mathbf{X}_1, \ldots, \mathbf{X}_{N/2}$ to determine the samples $\phi(\mathbf{X}_1), \ldots, \phi(\mathbf{X}_{N/2})$.

3 Sample $\mathbf{X}'_1, \ldots, \mathbf{X}'_{N/2}$ with density f such that

$$\left(\frac{1}{N/2} \sum_{i=1}^{N/2} \mathbf{1}_{\phi(\mathbf{X}_i)>T} \mathbf{1}_{\phi(\mathbf{X}'_i)>T} - \frac{1}{N/2} \sum_{i=1}^{N/2} \mathbf{1}_{\phi(\mathbf{X}_i)>T} \times \frac{1}{N/2} \sum_{i=1}^{N/2} \mathbf{1}_{\phi(\mathbf{X}'_i)>T} \right) < 0$$

Estimate $\hat{\mathbb{P}}^{\text{AV}} = \frac{1}{N} \sum_{i=1}^{N/2} \mathbf{1}_{\phi(\mathbf{X}_i)>T} + \mathbf{1}_{\phi(\mathbf{X}'_i)>T}$.

4 **return** $\hat{\mathbb{P}}^{\text{AV}}$.

Table 5.5 Results of AV for the identity function

T	$\hat{\mathbb{P}}^{\text{AV}}$	RB ($\hat{\mathbb{P}}^{\text{AV}}$)	RE ($\hat{\mathbb{P}}^{\text{AV}}$)	**Simulation budget**	ν^{AV}
3	1.4×10^{-3}	3%	24%	10,000	1.3
4	3.1×10^{-5}	−7%	51%	100,000	1.7

5.2.4.2 Application to a toy case

Identity function

Antithetic variates are applied to the identity function toy case. We propose to define $X' = -X$ so that X' and X have the same density, a standard one-dimension normal pdf. Probability estimation results with antithetic variates are given in Table 5.5. For a given simulation budget, the variance of CMC decreases with the use of antithetic variates.

5.2.4.3 Conclusion

Antithetic variate method is a variance reduction method that is not really adapted to the estimation of rare event probabilities (see Table 5.6). Indeed, exactly 50% of the AV samples are made of CMC samples. The application of such a method for a general function $\phi(\cdot)$ is also difficult because determining an antithetic variate to $\mathbf{1}_{\phi(\mathbf{X})>T}$ is not obvious.

5.3 Importance sampling

Remark 5.3.1. *Because the main mechanism involved in importance sampling is the modification of the sampling distribution, we will specify in this section for the*

Table 5.6 Summary of AV for rare event probability estimation

Criteria	AV characteristics
Rarity of $\mathbb{P}_f$	Not adapted to rare event probability
Simulation budget	At least 10^3 samples in most cases
Type of function $\phi(\cdot)$	No direct influence
Multimodality of f restricted to Ω_f	No direct influence
Dimension of $\mathbf{X}$	No direct influence
Probability estimate error	Estimation with retrials
Difficulty of implementation	Low; difficult to determine an efficient couple H_1 and H_2 in practice

sake of clarity the sampling distribution by a subscript determining the effective sampling distribution in the expression of the expectation and variance. For example, $\mathbb{E}_h(\mathbf{X})$ *means that the samples used for the calculation of the expectation have been generated with the pdf h.*

5.3.1 Principle of importance sampling

The main idea of importance sampling (IS) (Engelund & Rackwitz, 1993; Kroese & Rubinstein, 2012; L'Ecuyer, Mandjes, & Tuffin, 2009) is to use an auxiliary distribution h to generate more samples $\mathbf{X}_1, \ldots, \mathbf{X}_N$ such that $\phi(\mathbf{X}) > T$ than with the initial distribution f. A weight is then introduced in the probability estimate to take into account the change in the pdf which is used to generate the samples. IS takes advantage of the fact that

$$\mathbb{P}_f = \mathbb{E}_f\left(\mathbf{1}_{\phi(\mathbf{X})>T}\right) = \mathbb{E}_h\left(\mathbf{1}_{\phi(\mathbf{X})>T}\frac{f(\mathbf{X})}{h(\mathbf{X})}\right)$$

The IS probability estimate $\hat{\mathbb{P}}^{\text{IS}}$ is then given by

$$\hat{\mathbb{P}}^{\text{IS}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}_{\phi(\mathbf{X}_i)>T}\frac{f(\mathbf{X}_i)}{h(\mathbf{X}_i)} \tag{5.3}$$

where the rv $\mathbf{X}_i, i = 1, \ldots, N$ are sampled iid from pdf h. The estimate $\hat{\mathbb{P}}^{\text{IS}}$ of $\mathbb{P}_f$ is unbiased. Its variance is given by the following equation:

$$\mathbb{V}_h\left(\hat{\mathbb{P}}^{\text{IS}}\right) = \frac{\mathbb{V}_h\left(\sum_{i=1}^{N}\mathbf{1}_{\phi(\mathbf{X}_i)>T}\frac{f(\mathbf{X}_i)}{h(\mathbf{X}_i)}\right)}{N^2} = \frac{\mathbb{V}_h\left(\mathbf{1}_{\phi(\mathbf{X})>T}w(\mathbf{X})\right)}{N} \tag{5.4}$$

with $w(\mathbf{X}) = \frac{f(\mathbf{X})}{h(\mathbf{X})}$. The term $w(\mathbf{X})$ is often called the *likelihood function* in the importance sampling literature. The variance can be estimated using the classical Monte Carlo formula as follows:

$$\hat{\mathbb{V}}_h\left(\hat{\mathbb{P}}^{\mathrm{IS}}\right) = \frac{1}{N-1}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}_{\phi(\mathbf{X}_i)>T}w^2(\mathbf{X}_i) - (\hat{\mathbb{P}}^{\mathrm{IS}})^2\right) \tag{5.5}$$

The variance of $\hat{\mathbb{P}}^{\mathrm{IS}}$ strongly depends on the choice of h. If h is well chosen, the IS estimate then has a much smaller variance than the Monte Carlo estimate and vice versa. The objective of IS is to decrease the estimate variance. We can thus define an optimal IS auxiliary density h_{opt} as the density that minimizes the variance $\mathbb{V}_h\left(\hat{\mathbb{P}}^{\mathrm{IS}}\right)$. Because variances are non-negative quantities, the optimal auxiliary density h_{opt} is determined by canceling the variance in Equation (5.4). h_{opt} is then defined with (Bucklew, 2004)

$$h_{\mathrm{opt}}(\mathbf{x}) = \frac{\mathbf{1}_{\phi(\mathbf{x})>T}f(\mathbf{x})}{\mathbb{P}_f} \tag{5.6}$$

The optimal auxiliary density h_{opt} depends unfortunately on the probability $\mathbb{P}_f$ that we estimate and consequently is unusable in practice. Nevertheless, h_{opt} can be derived to determine an efficient sampling pdf. Indeed, a valuable auxiliary sampling pdf h will be close to the density h_{opt} with regard to a given criterion. An optimization of the auxiliary sampling pdf is then necessary.

This zero variance approach is difficult to implement directly for quantile estimation because the variance of quantile depends on the unknown quantity $f^2(q_\alpha)$ (see Equation 5.2). Nevertheless, using Equation (5.6), a simple and practically feasible alternative is to choose the following optimal IS auxiliary density to estimate the α-quantile q_α:

$$h_{\mathrm{opt}}(\mathbf{x}) = \frac{\mathbf{1}_{\phi(\mathbf{x})\geq q_\alpha}f(\mathbf{x})}{\alpha}$$

This auxiliary density is often suggested in different articles (Cannamela et al., 2008; Egloff & Leippold, 2010) without further theoretical justification but based on the similarity with Equation (5.6).

Specific surveys on IS have been proposed as in Smith, Shafi, & Gao (1997) and Tokdar & Kass (2010). Some possible importance sampling heuristics are efficient only in very restrictive cases and thus are not analyzed for the sake of conciseness. In this book, we review in the next sections only the main IS algorithms that can be used for a general input–output function $\phi(\cdot)$.

5.3.2 Nonadaptive importance sampling

The purpose of nonadaptive IS is to learn the optimal sampling density h_{opt} in a parametric way with a direct procedure. The required simulation budget is more limited than with adaptive algorithms. However, the potential application of nonadaptive importance sampling on real systems is more restricted than adaptive importance sampling algorithms described in Section 5.3.3.

5.3.2.1 Simple changes of measure

Principle

In simple cases of function $\phi(\cdot)$ or when we have some knowledge of the set Ω_f, conventional changes of density f can be efficient for decreasing the probability estimate variance such as scaling (SC) and mean translation (MT). SC consists in defining the auxiliary pdf h so that

$$h(\mathbf{x}) = \frac{1}{a} f\left(\frac{\mathbf{x}}{a}\right)$$

with $a \in \mathbb{R}^{+*}$ and usually $a > 1$. The new auxiliary density considered in MT is of type

$$h(\mathbf{x}) = f(\mathbf{x} - \mathbf{c})$$

with $\mathbf{c} \in \mathbb{R}^d$. An example of sampling procedure is given in Algorithm 5 for SC and MT. The choices of a and $\mathbf{c}$ for each method strongly influence the importance sampling efficiency. Effective values of a and $\mathbf{c}$ are not obvious to find without some knowledge of the function $\phi(\cdot)$. Because SC and MT correspond to a parametric model of the auxiliary density, it is possible to optimize the value of a or $\mathbf{c}$ with cross entropy (see Section 5.3.3).

Application to a toy case

Four-branch system In order to illustrate the mechanics of scaling and mean translation IS as well as the dependence of their accuracy with respect to the choice of a and $\mathbf{c}$, these two methods are applied to the four-branch system. For each method, two values of the auxiliary pdf parameter are selected. The results obtained for this case are provided in Table 5.7 (for $T = 10$). The results of MT and SC depend on the choice of the auxiliary pdf parameters a and $\mathbf{c}$. In this case, the best choice for MT parameters corresponds to the initial pdf (i.e., $[0, 0]$) because the regions of the input space that lead to the threshold exceedance are equidistant from the origin. MT is thus equivalent in this test case to a CMC simulation. Choosing an appropriate value of the scaling factor (i.e., $a = 1/3$) allows to reduce the relative error of the estimation.

ALGORITHM 5 Scaling (or mean translation) with importance sampling for probability estimation

Input: The pdf f, the number of samples N, the function $\phi(\cdot)$, the threshold T, and a constant a (or $\mathbf{c}$ in case of mean translation).

Output: The probability estimate $\hat{\mathbb{P}}^{\text{SC}}$ (or $\hat{\mathbb{P}}^{\text{MT}}$ in case of MT).

1 Sample $\mathbf{X}_1, \ldots, \mathbf{X}_N$ with density $h(\mathbf{x}) = \frac{1}{a} f\left(\frac{\mathbf{x}}{a}\right)$ (or $h(\mathbf{x}) = f(\mathbf{x} - \mathbf{c})$ in case of MT).

2 Apply ϕ to $\mathbf{X}_1, \ldots, \mathbf{X}_N$ to determine the samples $\phi(\mathbf{X}_1), \ldots, \phi(\mathbf{X}_N)$.

3 Estimate $\hat{\mathbb{P}}^{\text{SC}}$ (or $\hat{\mathbb{P}}^{\text{MT}}$) $= \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\phi(\mathbf{X}_i) > T} \frac{f(\mathbf{X}_i)}{h(\mathbf{X}_i)}$.

4 **return** $\hat{\mathbb{P}}^{\text{SC}}$ (or $\hat{\mathbb{P}}^{\text{MT}}$) .

Table 5.7 Results obtained by scaling (SC) and mean translation (MT) IS for the four-branch system and T=10

	Value of auxiliary pdf parameter	$\hat{\mathbb{P}}$	RB ($\hat{\mathbb{P}}$)	RE ($\hat{\mathbb{P}}$)	Simulation budget	v
SC	$a = 2$	2.0×10^{-3}	-10%	1571%	1000	1.8×10^{-3}
	$a = 1/3$	2.2×10^{-3}	-0.9%	17%	1000	16
MT	$\mathbf{c} = [0, 0]$	2.2×10^{-3}	-0.9%	68%	1000	1
	$\mathbf{c} = [-1, 1]$	2.2×10^{-3}	-0.9%	431%	1000	2.4×10^{-2}

Table 5.8 Summary of IS with simple changes of measure for rare event probability estimation

Criteria	MT and SC characteristics
Rarity of $\mathbb{P}_f$	No general influence on the methods
Simulation budget	At least 10^3 samples for a reasonable application
Type of function $\phi(\cdot)$	No regularity condition on $\phi(\cdot)$
Multimodality of f restricted to Ω_f	Monomodality preferable for MT
Dimension of $\mathbf{X}$	No theoretical influence on the IS convergence, but weight degeneracy can occur
Probability estimate error	Estimated with Equation (5.5) or with retrials
Difficulty of implementation	Low; *a priori* knowledge of efficient $\mathbf{c}$ and a

In most cases, the direct tuning of a and $\mathbf{c}$ can be obtained only with some *a priori* knowledge on the function $\phi(\cdot)$.

Conclusion

Simple changes of measure are basic applications of IS that require some *a priori* information for an efficient application. A summary on these techniques is given in Table 5.8. In practice, simple changes of measure are often considered in a preliminary auxiliary pdf design but are seldomly used for an accurate estimation, notably in complex and high-dimensional systems.

5.3.2.2 Exponential twisting

Principle

The main idea of exponential twisting is to define the auxiliary density on the output $Y = \phi(\mathbf{X})$ with

$$h_\theta(y) = \exp(\theta y - \lambda(\theta)) g(y) \tag{5.7}$$

where g is the density of random variable Y and $\lambda(\theta) = \ln\left(\mathbb{E}_g\left(\exp\left(\theta Y\right)\right)\right)$. The probability is then determined with

$$\mathbb{P}_f = \mathbb{E}_{h_\theta}\left(\mathbf{1}_{Y>T}\frac{g(Y)}{h_\theta(Y)}\right)$$

Exponential twisting can thus be applied only in some specific cases, notably if $Y = \sum_{i=1}^{d} X_i$ (function used in some queueing models (Heidelberger, 1995)) or if the density g is analytically known. Exponential twisting principle stems from the large deviation theory (see Section 6.2). The variable Y must have exponential moments so that $\lambda(\theta)$ can be finite for at least some values of θ. The pdf h depends on the parameter θ. We would like, of course, to choose θ to minimize variance or, equivalently, the second moment of the probability estimator. The second moment can be bounded with

$$\mathbb{E}_{h_\theta}\left(\mathbf{1}_{Y>T}\exp\left(-2\theta Y + 2\lambda(\theta)\right)\right) \leq \exp\left(-2\theta T + 2\lambda(\theta)\right)$$

Minimizing the second moment is complicated, but minimizing the upper bound is equivalent to maximizing $\theta T - \lambda(\theta)$. The function $\lambda(\theta)$ is convex; thus, an optimal value θ_{opt} can be obtained with saddle point approximation (Daniels, 1954; Goutis & Casella, 1999; Huzurbazar, 1999) such that

$$\left.\frac{\mathrm{d}\lambda(\theta)}{\mathrm{d}\theta}\right|_{\theta=\theta_{\text{opt}}} = T \tag{5.8}$$

The parameter θ_{opt} is then estimated numerically. A property of exponential twisting is given by

$$\mathbb{E}_{h_\theta}(Y) = \frac{\mathrm{d}\lambda(\theta)}{\mathrm{d}\theta}$$

If the density f is twisted with parameter θ_{opt}, we then have

$$\mathbb{E}_{h_\theta}(Y) = \left.\frac{\mathrm{d}\lambda(\theta)}{\mathrm{d}\theta}\right|_{\theta=\theta_{\text{opt}}} = T$$

The distribution of Y has been shifted so that T is now its mean when $\theta = \theta_{\text{opt}}$.

ALGORITHM 6 Importance sampling with exponentially twisted density

Input: The pdf g, the number of samples N, and the threshold T.
Output: The probability estimate $\hat{\mathbb{P}}^{\text{TW}}$.

1 Determine θ_{opt} such that $d\frac{\ln(\mathbb{E}_g(\exp(\theta Y)))}{d\theta}\Big|_{\theta=\theta_{\text{opt}}} = T$.
2 Sample $Y_1, \ldots, Y_N$ with density $h_{\theta_{\text{opt}}}(y) = \exp(\theta_{\text{opt}} y - \lambda(\theta_{\text{opt}}))g(y)$.
3 Estimate $\hat{\mathbb{P}}^{\text{TW}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}_{Y_i>T}\frac{g(Y_i)}{h_{\theta_{\text{opt}}}(Y_i)}$.
4 **return** $\hat{\mathbb{P}}^{\text{TW}}$.

When $Y = \sum_{i=1}^{d} X_i$, the estimator $\hat{\mathbb{P}}^{\text{TW}}$ has a bounded relative error if the variables X_i have a light tail (Asmussen, 2003; Siegmund, 1976). In case of probabilities that follow the large deviation principle (see Section 6.2) and under some general conditions, logarithmic efficiency is guaranteed with exponential twisting IS (Dieker & Mandjes, 2005).

Application to a toy case

Identity function We present an application of exponential twisting in the simple case of identity function. The density of Y, g is given by $\mathcal{N}(0, 1)$. Thus, h_θ is twisted from g in the following way:

$$h_\theta(y) = \exp(\theta y - \lambda(\theta)) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right)$$

with $\lambda(\theta) = \ln\left(\mathbb{E}_g\left(\exp\left(\theta Y\right)\right)\right)$. It could be shown that $\mathbb{E}_g\left(\exp\left(\theta Y\right)\right) = \exp\left(\frac{1}{2}\theta^2\right)$ and, thus, $\lambda(\theta) = \frac{1}{2}\theta^2$. The density h_θ can be rewritten such that

$$h_\theta(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y-\theta)^2\right)$$

The density h corresponds to a twisted normal distribution with a mean θ and variance equal to 1. In that case, exponential twisting is equivalent to a mean translation. Because $\frac{\mathrm{d}\lambda(\theta)}{\mathrm{d}\theta}\Big|_{\theta=\theta_{\text{opt}}} = \theta_{\text{opt}}$, we then determine that $\theta_{\text{opt}} = T$ to estimate the probability $\mathbb{P}_f$. The importance sampling density is thus twisted with a mean that equals the target threshold and is defined by

$$h(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y-T)^2\right)$$

Table 5.9 summarizes some probability estimates obtained with exponential twisting for different thresholds. We find efficient importance sampling probability estimates in this simple case with a low relative error. Nevertheless, it is often impossible to find such efficient sampling density for a complex system.

Table 5.9 Results obtained by exponential twisting for the identity function

T	$\hat{\mathbb{P}}^{\text{TW}}$	RB ($\hat{\mathbb{P}}^{\text{TW}}$)	RE ($\hat{\mathbb{P}}^{\text{TW}}$)	**Simulation budget**	ν^{TW}
3	1.3×10^{-3}	0.7%	6%	1000	213
4	3.2×10^{-5}	0.2%	7%	1000	6.4×10^3
5	2.8×10^{-7}	0.3%	7%	1000	7.3×10^5

Table 5.10 Summary of IS with exponential twisting for rare event probability estimation

Criteria	Exponential twisting characteristics
Rarity of $\mathbb{P}_f$	Increase the efficiency of the method
Simulation budget	At least 10^3 samples for a reasonable application
Type of function $\phi(\cdot)$	Essentially applied when $\phi(\mathbf{X}) = \sum_{i=1}^{d} X_i$ or when the density of $\phi(\mathbf{X})$ is known
Multimodality of f restricted to Ω_f	Not really adapted
Dimension of $\mathbf{X}$	No influence on the convergence
Probability estimate error	Estimated with Equation (5.5)
Difficulty of implementation	Low; possible complexity to find θ_{opt}

Conclusion

Exponential twisting, an algorithm based on large deviation theory, can be applied only to very specific cases of function $\phi(\cdot)$ or when the output density is analytically known. Thus, this method could not be applied to complex systems. A summary of IS with exponential twisting is presented in Table 5.10.

5.3.3 Adaptive importance sampling

The goal of adaptive IS is to learn the optimal sampling density h_{opt} with an iterative procedure. This approach can be applied for a parametric density, that is, one determines the parameters of a family density that minimize the variance. Nonparametric adaptive IS has also been proposed with the use of kernel density estimation (see Section 2.2.2). Both techniques are described in the following discussion.

5.3.3.1 Cross-entropy optimization of the importance sampling auxiliary density

Principle

Let us define $h_{\boldsymbol{\lambda}}$, a family of pdfs indexed by a parameter $\boldsymbol{\lambda} \in \boldsymbol{\Delta}$ where $\boldsymbol{\Delta}$ is the multidimensional space in which $\boldsymbol{\lambda}$ evolves. The parameter vector $\boldsymbol{\lambda}$, for instance, could be the mean and the covariance matrix in the case of Gaussian densities.

Definition 5.3.1 (Kullback–Leibler divergence (Kullback & Leibler, 1951))**.** *Let P and Q be two probability distributions defined by their pdf p and q with support $\mathbb{R}^d$. The Kullback–Leibler divergence between P and Q is defined by*

$$D_{\text{KL}}(P, Q) = \int_{\mathbb{R}^d} \ln\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right) p(\mathbf{x})\, \mathrm{d}\mathbf{x}$$

D_{KL} is a positive quantity and is equal to 0 if and only if $P = Q$ almost everywhere. $D_{\text{KL}}(P, Q)$ is not symmetric because $D_{\text{KL}}(P, Q) \neq D_{\text{KL}}(Q, P)$. The Kullback–Leibler divergence, also known as *relative entropy*, comes from the field of information

theory as the continuous entropy defined in Chapter 2. The objective of IS with cross entropy (CE) is to determine the parameter $\boldsymbol{\lambda}_{\text{opt}}$ that minimizes the Kullback–Leibler divergence between $h_{\boldsymbol{\lambda}}$ and h_{opt} (de Mello & Rubinstein, 2002; Rubinstein & Kroese, 2004). The value of $\boldsymbol{\lambda}_{\text{opt}}$ is thus obtained with

$$\boldsymbol{\lambda}_{\text{opt}} = \underset{\boldsymbol{\lambda}\in\boldsymbol{\Delta}}{\text{argmin}} \left\{ D_{\text{KL}}(h_{\text{opt}}, h_{\boldsymbol{\lambda}}) \right\} \tag{5.9}$$

Determining the parameter $\boldsymbol{\lambda}_{\text{opt}}$ with Equation (5.9) is not obvious because it depends on the unknown pdf h_{opt}. In fact, it can be shown (Rubinstein & Kroese, 2004) that Equation (5.9) is equivalent to Equation (5.10):

$$\boldsymbol{\lambda}_{\text{opt}} = \underset{\boldsymbol{\lambda}\in\boldsymbol{\Delta}}{\text{argmax}} \left\{ \mathbb{E}_f \left[\mathbf{1}_{\phi(\mathbf{X})>T} \ln\left(h_{\boldsymbol{\lambda}}(\mathbf{X})\right) \right] \right\} \tag{5.10}$$

In practice, we do not focus directly on Equation (5.10) but iteratively proceed to estimate $\boldsymbol{\lambda}_{\text{opt}}$ with an increasing sequence of thresholds:

$$\gamma_0 < \gamma_1 < \gamma_2 < \cdots < \gamma_k < \cdots \leq T$$

chosen adaptively using quantile definition. At the iteration k, the value $\boldsymbol{\lambda}_{k-1}$ is available, and we determine:

$$\begin{aligned} \boldsymbol{\lambda}_k &= \underset{\boldsymbol{\lambda}\in\boldsymbol{\Delta}}{\text{argmax}} \left\{ \mathbb{E}_{h_{\boldsymbol{\lambda}_{k-1}}} \left[\mathbf{1}_{\phi(\mathbf{X})>T} \frac{f(\mathbf{X})}{h_{\boldsymbol{\lambda}_{k-1}}(\mathbf{X})} \ln\left(h_{\boldsymbol{\lambda}}(\mathbf{X})\right) \right] \right\}, \\ &= \underset{\boldsymbol{\lambda}\in\boldsymbol{\Delta}}{\text{argmax}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\phi(\mathbf{X}_i)>\gamma_k} \frac{f(\mathbf{X}_i)}{h_{\boldsymbol{\lambda}_{k-1}}(\mathbf{X}_i)} \ln(h_{\boldsymbol{\lambda}}(\mathbf{X}_i)) \right\} \end{aligned}$$

where the samples $\mathbf{X}_1, \ldots, \mathbf{X}_N$ are generated with $h_{\boldsymbol{\lambda}_{k-1}}$. The probability $\hat{\mathbb{P}}^{\text{CE}}$ is then estimated with IS at the last iteration (when $\gamma_k \geq T$). The cross-entropy optimization algorithm for the IS density is described more precisely in Algorithm 7.

For some particular density families (e.g., exponential), the pdf optimal parameters can be found analytically by canceling the gradient of the following expression (under mild regularity conditions)

$$\frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\phi(\mathbf{X}_i)>\gamma_k} \frac{f(\mathbf{X}_i)}{h_{\boldsymbol{\lambda}_{k-1}}(\mathbf{X}_i)} \ln(h_{\boldsymbol{\lambda}}(\mathbf{X}_i))$$

with respect to $\boldsymbol{\lambda}$, which can be written as

$$\frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\phi(\mathbf{X}_i)>\gamma_k} \frac{f(\mathbf{X}_i)}{h_{\boldsymbol{\lambda}_{k-1}}(\mathbf{X}_i)} \nabla \ln(h_{\boldsymbol{\lambda}}(\mathbf{X}_i)) \tag{5.11}$$

ALGORITHM 7 IS optimized by cross entropy for probability estimation

Input: The pdf f, the number of samples N, the function $\phi(\cdot)$, the threshold T, a value $\rho \in (0,1)$, and $\boldsymbol{\lambda}_0 \in \boldsymbol{\Delta}$.

Output: The probability estimate $\hat{\mathbb{P}}^{\mathrm{CE}}$.

Set $k = 1$.
Generate the iid samples $\mathbf{X}_1, \ldots, \mathbf{X}_N$ with pdf $h_{\boldsymbol{\lambda}_0}$.
Apply ϕ to $\mathbf{X}_1, \ldots, \mathbf{X}_N$ to determine $Y_1 = \phi(\mathbf{X}_1), \ldots, Y_N = \phi(\mathbf{X}_N)$.
Compute the empirical ρ-quantile γ_1 of the samples $Y_1, \ldots, Y_N$.
while $\gamma_k < T$ **do**
    Optimize the parameters of the auxiliary pdf family with

$$\boldsymbol{\lambda}_k = \underset{\boldsymbol{\lambda} \in \boldsymbol{\Delta}}{\operatorname{argmax}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left[\mathbf{1}_{\phi(\mathbf{X}_i) > \gamma_k} \frac{f(\mathbf{X}_i)}{h_{\boldsymbol{\lambda}_{k-1}}(\mathbf{X}_i)} \ln \left[h_{\boldsymbol{\lambda}}(\mathbf{X}_i) \right] \right] \right\}$$

    Set $k \leftarrow k + 1$.
    Generate the iid samples $\mathbf{X}_1, \ldots, \mathbf{X}_N$ with pdf $h_{\boldsymbol{\lambda}_{k-1}}$.
    Apply ϕ to $\mathbf{X}_1, \ldots, \mathbf{X}_N$ to determine $Y_1 = \phi(\mathbf{X}_1), \ldots, Y_N = \phi(\mathbf{X}_N)$.
    Compute the empirical ρ-quantile γ_k of the samples $Y_1, \ldots, Y_N$.
Estimate the probability $\hat{\mathbb{P}}^{\mathrm{CE}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\phi(\mathbf{X}_i) > T} \frac{f(\mathbf{X}_i)}{h_{\boldsymbol{\lambda}_{k-1}}(\mathbf{X}_i)}$.
return $\hat{\mathbb{P}}^{\mathrm{CE}}$.

which involves the classical "score function" $\nabla \ln(h_{\boldsymbol{\lambda}}(\mathbf{X}))$. For example, if the auxiliary distribution is composed of d independent Gaussian random variables of which the mean μ_j and variance $v_j, j = 1, \ldots, d$ are optimized by CE, the updating formula of these parameters at the kth iteration of the algorithm, obtained by canceling Equation (5.11), are given by

$$\mu_j = \frac{\sum_{i=1}^{N} \mathbf{1}_{\phi(\mathbf{X}_i) > \gamma_k} \frac{f(\mathbf{X}_i)}{h_{\boldsymbol{\lambda}_{k-1}}(\mathbf{X}_i)} X_{ij}}{\sum_{i=1}^{N} \mathbf{1}_{\phi(\mathbf{X}_i) > \gamma_k} \frac{f(\mathbf{X}_i)}{h_{\boldsymbol{\lambda}_{k-1}}(\mathbf{X}_i)}},$$

$$v_j = \frac{\sum_{i=1}^{N} \mathbf{1}_{\phi(\mathbf{X}_i) > \gamma_k} \frac{f(\mathbf{X}_i)}{h_{\boldsymbol{\lambda}_{k-1}}(\mathbf{X}_i)} \left(X_{ij} - \mu_j \right)^2}{\sum_{i=1}^{N} \mathbf{1}_{\phi(\mathbf{X}_i) > \gamma_k} \frac{f(\mathbf{X}_i)}{h_{\boldsymbol{\lambda}_{k-1}}(\mathbf{X}_i)}}$$

We recall here that the notation X_{ij} stands for the jth component of the ith sample of $\mathbf{X}$. CE is a very practical algorithm to approximate the optimal sampling density. Nevertheless, the choice of the parametric family density $h_{\boldsymbol{\lambda}}$ must be made carefully to obtain valuable results. Because of the adaptiveness of the algorithm, it is difficult to ensure the robustness (logarithmic efficiency) of the CE estimate in general. A new

concept of probabilistic bounded relative error has notably been proposed in Tuffin & Ridder (2012).

The CE algorithm for α-quantile estimation is relatively similar to Algorithm 7. Nevertheless, at each iteration k, instead of testing $\gamma_k < T$, we compare $\gamma_k < \hat{q}_\alpha^{\mathrm{CE}}$ where $\hat{q}_\alpha^{\mathrm{CE}}$ is the α-quantile CE estimate obtained with

$$\hat{q}_\alpha^{\mathrm{CE}} = \inf_{y \in \mathbb{R}} \left\{ F_N^{\mathrm{CE}}(y) \geq \alpha \right\}$$

where $F_N^{\mathrm{CE}}(y)$ is the empirical cdf defined by

$$F_N^{\mathrm{CE}}(y) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\phi(\mathbf{X}_i) \leq y} \frac{f(\mathbf{X}_i)}{h_{\boldsymbol{\lambda}_{k-1}}(\mathbf{X}_i)}$$

Application to toy cases

Four-branch system In order to illustrate the mechanisms involved in CE, the CE method is applied first to the four-branch system with classical unimodal parametric multivariate Gaussian density. The results obtained for this case are provided in Table 5.11 for different thresholds. For this toy case, the value of ρ-quantile parameter has been set to 0.97, the number of samples per iteration is equal to 500, and only the bandwidth of the auxiliary pdf is optimized. The auxiliary sampling distribution is illustrated in Figure 5.3. The relative error of CE is relatively high in this case (32%) although it improves the CMC estimation. This can be explained by the choice of the

Table 5.11 Results of CE for the four-branch system toy case (with parametric multivariate Gaussian as auxiliary density)

T	$\hat{\mathbb{P}}^{\mathrm{CE}}$	RB ($\hat{\mathbb{P}}^{\mathrm{CE}}$)	RE ($\hat{\mathbb{P}}^{\mathrm{CE}}$)	**Simulation budget**	ν^{CE}
10	2.4×10^{-3}	−8%	29%	1400	4
12	1.3×10^{-6}	10%	32%	10,000	819

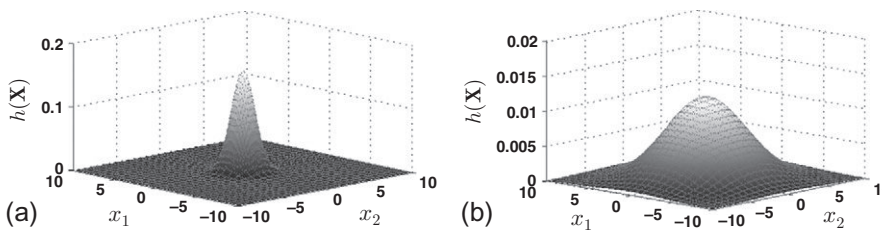

Figure 5.3 CE auxiliary pdf evolution: (a) initialization; (b) final iteration.

parameterized pdf family (here, a two-dimension normal density function) that is not appropriate to a disconnected failure region Ω_f.

Polynomial product function CE is applied to the polynomial product function with different dimensions d and thresholds T. Two auxiliary density families have also been tested: Laplace pdf and Gaussian pdf. In these two cases, only the bandwidths are optimized for each dimension of the random vector (we suppose that the dimensions of the auxiliary pdf are independent). The results obtained for this toy case are summarized in Table 5.12. The choice of the auxiliary density is crucial in CE. Indeed, in this toy case, the results obtained with the parametric multivariate Gaussian pdf as auxiliary density are worse than those obtained with parametric Laplace pdf. The choice of Gaussian auxiliary density also induces an increase of the simulation budget in order to provide an estimation of the target probabilities. Conversely, the use of Laplace pdf results in very accurate probability estimations in a reduced simulation budget even for relatively high-dimensional problems.

Conclusion

CE optimization is an effective generic approach to determine a valuable parametric auxiliary IS distribution. Nevertheless, the choice of a parametric density must be made carefully *a priori* to fit the unknown optimal auxiliary IS distribution. Valuable results with a well-chosen parametric density applying CE on realistic test cases could give valuable results. See the main characteristics of CE in Table 5.13.

5.3.3.2 Nonparametric adaptive importance sampling

Principle

The objective of the nonparametric adaptive importance sampling (NAIS) technique (Morio, 2012; Neddermeyer, 2009, 2010; Zhang, 1996) is to approximate the IS optimal auxiliary density given in Equation (5.6) with kernel density function (see Section 2.2.2). NAIS only requires the choice of a kernel density family and

Table 5.12 Results of CE for the polynomial product toy case (for different dimensions and auxiliary densities)

d	T	$\hat{\mathbb{P}}^{\text{CE}}$	RB ($\hat{\mathbb{P}}^{\text{CE}}$)	RE ($\hat{\mathbb{P}}^{\text{CE}}$)	**Simulation budget**	ν^{CE}
Auxiliary pdf family: Gaussian						
5	400	8.5×10^{-7}	0.7%	22%	110×10^3	223
20	500	7.4×10^{-7}	−32%	151%	124×10^3	3
50	700	2.7×10^{-7}	−24%	309%	210×10^3	1.4
200	1000	2.7×10^{-6}	−43%	129%	586×10^3	0.4
Auxiliary pdf family: Laplace						
5	400	8.4×10^{-7}	−0.5%	20%	8×10^3	3700
20	500	1.1×10^{-6}	−0.9%	21%	20×10^3	1030
50	700	3.5×10^{-7}	−1.7%	19%	60×10^3	1320
200	1000	4.4×10^{-6}	−6%	80%	300×10^3	1.1

Table 5.13 Summary of IS optimized with CE for rare event probability estimation

Criteria	CE characteristics
Rarity of $\mathbb{P}_f$	Increased number of iterations before reaching convergence
Simulation budget	At least 10^3 samples in most cases
Type of function $\phi(\cdot)$	No influence
Multimodality of f restricted to Ω_f	High influence on the choice of h_λ
Dimension of $\mathbf{X}$	Indirect influence through the dimension of Δ
Probability estimate error	Estimated with Equation (5.5) or with retrials
Difficulty of implementation	Medium; possible complex optimization to determine λ_{opt}

is thus more flexible than a parametric model such as CE. Its iterative principle is relatively similar to the CE optimization and is described in Algorithm 8. At each iteration of NAIS, we iteratively proceed with an increasing sequence of thresholds

$$\gamma_0 < \gamma_1 < \gamma_2 < \cdots < \gamma_k < \cdots \leq T$$

adaptively chosen using quantile definition. At the iteration k, we approximate the optimal sampling density for estimating $\mathbb{P}(\phi(\mathbf{X}) > \gamma_k)$, that is, $\frac{\mathbf{1}_{\phi(\mathbf{x})>\gamma_k} f(\mathbf{x})}{\mathbb{P}(\phi(\mathbf{X})>\gamma_k)}$ from the current available samples with a kernel density h_{k+1}. The use of kernel density function enables to approximate a large range of optimal auxiliary densities. The efficiency of NAIS strongly decreases when the dimension of the input space is greater than 10 (10 as an order of magnitude) because of the numerical cost induced by the use of kernel density (Morio, 2012) which highly suffers from the "curse of dimensionality." The NAIS algorithm for α-quantile estimation is relatively similar to Algorithm 8. Nevertheless, at each iteration k, instead of testing $\gamma_k < T$, we compare $\gamma_k < \hat{q}_\alpha^{\text{NAIS}}$, where $\hat{q}_\alpha^{\text{NAIS}}$ is the α-quantile NAIS estimate obtained with

$$\hat{q}_\alpha^{\text{NAIS}} = \inf_{y \in \mathbb{R}} \left\{ \hat{F}_N^{\text{NAIS}}(y) \geq \alpha \right\}$$

where $\hat{F}_N^{\text{NAIS}}(y)$ is the empirical cdf defined by

$$\hat{F}_N^{\text{NAIS}}(y) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\phi(\mathbf{X}_i) \leq y} \frac{f(\mathbf{X}_i)}{h_k(\mathbf{X}_i)}$$

Application to toy cases

Four-branch system NAIS is applied to the four-branch system in order to illustrate its ability to perform probability estimations when the failure region Ω_f is disconnected. The results obtained for this case are provided in Table 5.14 for different

ALGORITHM 8 NAIS for probability estimation

Input: The pdf f, the number of samples N, the function $\phi(\cdot)$, the threshold T, a value $\rho \in (0,1)$, and a kernel K.

Output: The probability estimate $\hat{\mathbb{P}}^{\text{NAIS}}$.

Set $k = 1$ and $h_0 = f$.

Generate the iid samples $\mathbf{X}_1^{(1)}, \ldots, \mathbf{X}_N^{(1)}$ with pdf h_0.

Apply ϕ to $\mathbf{X}_1^{(1)}, \ldots, \mathbf{X}_N^{(1)}$ to determine $Y_1^{(1)} = \phi(\mathbf{X}_1^{(1)}), \ldots, Y_N^{(1)} = \phi(\mathbf{X}_N^{(1)})$.

Compute the empirical ρ-quantile γ_1 of the samples $Y_1^{(1)}, \ldots, Y_N^{(1)}$.

while $\gamma_k < T$ **do**

  Estimate $l_k = \frac{1}{kN} \sum_{j=1}^{k} \sum_{i=1}^{N} \mathbf{1}_{\phi(\mathbf{X}_i^{(j)}) \geq \gamma_k} \frac{f(\mathbf{X}_i^{(j)})}{h_{j-1}(\mathbf{X}_i^{(j)})}$ and set

  $w_j(\mathbf{X}_i^{(j)}) = \mathbf{1}_{\phi(\mathbf{X}_i^{(j)}) \geq \gamma_k} \frac{f(\mathbf{X}_i^{(j)})}{h_{j-1}(\mathbf{X}_i^{(j)})}$.

  Update the kernel sampling pdf with

$$h_{k+1}(\mathbf{x}) = \frac{1}{kNl_k \det\left(\mathbf{B}_k^{1/2}\right)} \sum_{j=1}^{k} \sum_{i=1}^{N} w_j(\mathbf{X}_i^{(j)}) K_d\left(\mathbf{B}_k^{-1/2}\left(\mathbf{x} - \mathbf{X}_i^{(j)}\right)\right)$$

  where $\mathbf{B}_k$ is a symmetric positive definite bandwidth matrix optimized with asymptotic integrated square error (AMISE) criterion (see Section 2.2.2).

  Set $k \leftarrow k + 1$.

  Generate the iid samples $\mathbf{X}_1^{(k)}, \ldots, \mathbf{X}_N^{(k)}$ with pdf h_k.

  Apply ϕ to $\mathbf{X}_1^{(k)}, \ldots, \mathbf{X}_N^{(k)}$ to determine $Y_1^{(k)} = \phi(\mathbf{X}_1^{(k)}), \ldots, Y_N^{(k)} = \phi(\mathbf{X}_N^{(k)})$.

  Compute the empirical ρ-quantile γ_k of the samples $Y_1^{(k)}, \ldots, Y_N^{(k)}$.

Estimate the probability $\hat{\mathbb{P}}^{\text{NAIS}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\phi(\mathbf{X}_i^{(k)}) > T} \frac{f(\mathbf{X}_i^{(k)})}{h_k(\mathbf{X}_i^{(k)})}$.

return $\hat{\mathbb{P}}^{\text{NAIS}}$.

Table 5.14 Results of NAIS for the four-branch system toy case

T	$\hat{\mathbb{P}}^{\text{NAIS}}$	RB ($\hat{\mathbb{P}}^{\text{NAIS}}$)	RE ($\hat{\mathbb{P}}^{\text{NAIS}}$)	**Simulation budget**	ν^{NAIS}
10	2.2×10^{-3}	-0.9%	15%	998	20
12	1.2×10^{-6}	1.7%	13%	7120	7137

thresholds. NAIS enables to efficiently estimate the rare event probability. For this toy case, the ρ-quantile is set to 0.75 and the number of samples per iteration is equal to 250. The auxiliary sampling distribution evolution is illustrated in Figure 5.4. The four modes of the optimal auxiliary distribution are determined during the algorithm process.

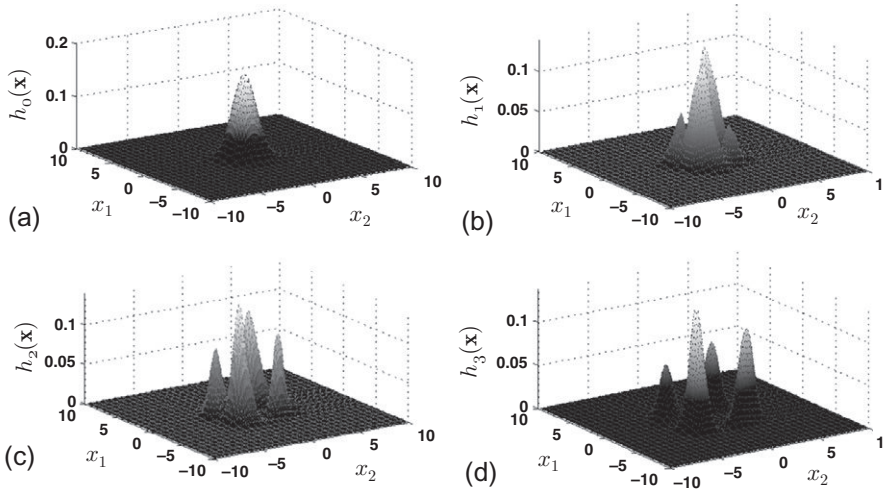

Figure 5.4 NAIS auxiliary pdf evolution. (a) Initialization. (b) First iteration. (c) Second iteration. (d) Final iteration.

Polynomial product function NAIS is applied to the polynomial product function for different dimensions d and thresholds T with Gaussian kernels. The results obtained for this toy case are summarized in Table 5.15. The efficiency of NAIS sharply decreases when the dimension increases. This toy case illustrates the difficulty to apply NAIS on high-dimensional use cases because it does not always converge in such situations.

Conclusion

NAIS approximates the optimal auxiliary IS distribution with a kernel density estimator. NAIS is thus particularly adapted when the optimal auxiliary IS distribution is complex (for instance, with several modes) with a dimension that is not too high. A summary of the main NAIS characteristics is presented in Table 5.16. This adaptiveness of NAIS makes it interesting to consider on realistic test cases.

Table 5.15 Results of NAIS for the polynomial product toy case (for different dimensions and thresholds)

d	T	$\hat{\mathbb{p}}^{\text{NAIS}}$	RB ($\hat{\mathbb{p}}^{\text{NAIS}}$)	RE ($\hat{\mathbb{p}}^{\text{NAIS}}$)	**Simulation budget**	ν^{NAIS}
5	400	8.0×10^{-7}	−5%	23%	8730	2754
20	500	4.0×10^{-7}	−63%	252%	60,000	6
50	700	Not affordable	n/a	n/a	n/a	n/a
200	1000	Not affordable	n/a	n/a	n/a	n/a

Table 5.16 Summary of NAIS for rare event probability estimation

Criteria	NAIS characteristics
Rarity of $\mathbb{P}_f$	Increased number of iterations needed to converge
Simulation budget	At least 10^3 samples in most cases
Type of function $\phi(\cdot)$	No influence
Multimodality of f restricted to Ω_f	Particularly adapted to multimodal pdf comparatively to parametric algorithms
Dimension of $\mathbf{X}$	Great decrease of efficiency when $d > 10$ (curse of dimensionality)
Probability estimate error	Estimated with Equation (5.5) or with retrials
Difficulty of implementation	Medium; NAIS bandwidth estimation can be complex

5.4 Adaptive splitting technique

5.4.1 Description

The purpose of importance splitting, also called *subset sampling*, *subset simulation*, or *sequential Monte Carlo*, is to decompose the target probability in a product of conditional probabilities that can be estimated with a reasonable simulation budget. Importance splitting was first proposed in a physical context in Kahn & Harris (1951), and numerous variants (Au, 2005; Au & Beck, 2001; Botev & Kroese, 2012; Del Moral, 2004; L'Ecuyer, Demers, & Tuffin, 2006; L'Ecuyer, Le Gland, Lezaud, & Tuffin, 2009) have been worked out since. Considering the set $\mathbf{A} = \{\mathbf{x} \in \mathbb{R}^d | \phi(\mathbf{x}) > T\}$, the objective of adaptive splitting technique (AST) (Cérou, Del Moral, Furon, & Guyader, 2012) is to determine the probability $\mathbb{P}(\mathbf{X} \in \mathbf{A}) = \mathbb{P}(\phi(\mathbf{X}) > T)$. For that purpose, the principle of AST is to iteratively estimate supersets of $\mathbf{A}$ and then to estimate $\mathbb{P}(\mathbf{X} \in \mathbf{A})$ with conditional probabilities. Let us define $\mathbf{A}_0 = \mathbb{R}^d \supset \mathbf{A}_1 \supset \cdots \supset \mathbf{A}_{n-1} \supset \mathbf{A}_n = \mathbf{A}$, a decreasing sequence of $\mathbb{R}^d$ subsets with the smallest element $\mathbf{A} = \mathbf{A}_n$. The probability $\mathbb{P}(\mathbf{X} \in \mathbf{A})$ can then be rewritten in the following way through the Bayes' theorem (see Section 2.1.2):

$$\mathbb{P}(\mathbf{X} \in \mathbf{A}) = \prod_{k=1}^{n} \mathbb{P}(\mathbf{X} \in \mathbf{A}_k | \mathbf{X} \in \mathbf{A}_{k-1})$$

where we recall that $\mathbb{P}(\mathbf{X} \in \mathbf{A}_k | \mathbf{X} \in \mathbf{A}_{k-1})$ is the probability that $\mathbf{X} \in \mathbf{A}_k$ knowing that $\mathbf{X} \in \mathbf{A}_{k-1}$. An optimal choice of the sequence $\mathbf{A}_k$, $k = 0, \ldots, n$ is found when $\mathbb{P}(\mathbf{X} \in \mathbf{A}_k | X \in \mathbf{A}_{k-1}) = \rho$, where ρ is a constant, that is, when all the conditional probabilities are equal. The variance of $\mathbb{P}(\mathbf{X} \in \mathbf{A})$ is indeed minimized in this configuration as shown in Lagnoux (2006) and Cérou, Del Moral, Le Gland, & Lezaud (2006). Consequently, if each $\mathbb{P}(\mathbf{X} \in \mathbf{A}_k | \mathbf{X} \in \mathbf{A}_{k-1})$ is well estimated, then

the probability $\mathbb{P}(\mathbf{X} \in \mathbf{A})$ is estimated more accurately with AST than with a direct estimation by Monte Carlo (Cérou et al., 2012).

Let us define f_k the density of $\mathbf{X}$ restricted to the set $\mathbf{A}_k$. The subset $\mathbf{A}_k$ can be defined with $\mathbf{A}_k = \{\mathbf{x} \in \mathbb{R}^d | \phi(\mathbf{x}) > T_k\}$ for $k = 0, \ldots, n$ with $T = T_n > T_{n-1} > \cdots > T_k > \cdots > T_0$. Determining the sequence $\mathbf{A}_k$ is equivalent to choosing some values for T_k, with $k = 0, \ldots, n$. The value of T_k can be determined in an adaptive manner to obtain valuable results (Cérou et al., 2012) using ρ-quantile of samples generated with the pdf f_k. The different stages of AST to estimate $\mathbb{P}(\phi(\mathbf{X}) > T)$ are described in Algorithm 9.

Generating directly independent samples from the conditional densities f_k is in most cases impossible because they are usually unknown. We can then consider the Metropolis–Hastings algorithm (see Section 2.3.3.3). Indeed, the pdf f_k is known up to a constant because

$$f_k(\mathbf{x}) = \frac{\mathbf{1}_{\phi(\mathbf{x}) > \gamma_\rho^{(k-1)}} f(\mathbf{x})}{\mathbb{P}(\phi(\mathbf{X}) > \gamma_\rho^{(k-1)})}$$

Moreover, in the AST procedure, the samples $\mathbf{X}_i^{(k-1)}$, $i = 1, \ldots, N$ such that $\phi(\mathbf{X}_i^{(k-1)}) > \gamma_\rho^{(k-1)}$ are distributed with pdf f_k. They can be used as starting samples of the MH procedure. A transition kernel $\pi(\mathbf{X}' | \mathbf{X}_i^{(k-1)})$ must also be defined to apply a proposal/refusal procedure with the acceptance rate

ALGORITHM 9 Adaptive splitting technique for probability estimation

Input: The pdf f, the number of samples N, the function $\phi(\cdot)$, the threshold T, and a constant $\rho \in (0, 1)$.

Output: The probability estimate $\hat{\mathbb{P}}^{\text{AST}}$.

Set $k = 0$ and $f_0 = f$.

Generate N samples $\mathbf{X}_1^{(0)}, \ldots, \mathbf{X}_N^{(0)}$ from f_0 and apply the function ϕ in order to determine $\phi(\mathbf{X}_1^{(0)}), \ldots, \phi(\mathbf{X}_N^{(0)})$.

Estimate the ρ-quantile $\gamma_\rho^{(0)}$ of the samples $\phi(\mathbf{X}_1^{(0)}), \ldots, \phi(\mathbf{X}_N^{(0)})$.

while $\gamma_\rho^{(k)} < T$ **do**

  Determine the subset $\mathbf{A}_{k+1} = \{\mathbf{x} \in \mathbb{R}^d | \phi(\mathbf{x}) > \gamma_\rho^{(k)}\}$ and the conditional density f_{k+1}.

  Set $k \leftarrow k + 1$.

  Generate N samples $\mathbf{X}_1^{(k)}, \ldots, \mathbf{X}_N^{(k)}$ with pdf f_k and apply the function ϕ in order to determine $\phi(\mathbf{X}_1^{(k)}), \ldots, \phi(\mathbf{X}_N^{(k)})$.

  Estimate the ρ-quantile $\gamma_\rho^{(k)}$ of the samples $\phi(\mathbf{X}_1^{(k)}), \ldots, \phi(\mathbf{X}_N^{(k)})$.

Estimate $\hat{\mathbb{P}}^{\text{AST}} = (1 - \rho)^k \times \frac{1}{N} \sum_{i=1}^N \mathbf{1}_{\phi(\mathbf{X}_i^{(k)}) > T}$.

return $\hat{\mathbb{P}}^{\text{AST}}$.

$$\min\left(1, \mathbf{1}_{\phi(\mathbf{X}')>\gamma_\rho^{(k-1)}} \frac{f_k(\mathbf{X}')\pi(\mathbf{X}'|\mathbf{X}_i^{(k-1)})}{f_k(\mathbf{X}_i^{(k-1)})\pi(\mathbf{X}_i^{(k-1)}|\mathbf{X}')}\right)$$

Then the following random variable Ξ_k is distributed with pdf f_k thanks to this proposal/refusal method (Cérou et al., 2012):

$$\Xi_k = \begin{cases} \mathbf{X}' & \text{with probability} \\ & \min\left(1, \mathbf{1}_{\phi(\mathbf{X}')>\gamma_\rho^{(k-1)}} \frac{f(\mathbf{X}')\pi(\mathbf{X}'|\mathbf{X}_i^{(k-1)})}{f(\mathbf{X}_i^{(k-1)})\pi(\mathbf{X}_i^{(k-1)}|\mathbf{X}')}\right), \\ \mathbf{X}_i^{(k-1)} & \text{with } \phi(\mathbf{X}_i^{(k-1)}) > \gamma_\rho^{(k-1)}, \quad \text{otherwise} \end{cases}$$

If π is an f-reversible Markovian kernel, the proposal/refusal procedure is simplified because

$$\Xi_k = \begin{cases} \mathbf{X}' & \text{if } \phi(\mathbf{X}') > \gamma_\rho^{(k-1)}, \\ \mathbf{X}_i^{(k-1)} & \text{with } \phi(\mathbf{X}_i^{(k-1)}) > \gamma_\rho^{(k-1)}, \quad \text{otherwise} \end{cases}$$

This proposal/refusal algorithm enables us to generate any number of samples according to f_k in a simple manner. The number of samples to estimate each $\mathbb{P}(\mathbf{X} \in \mathbf{A}_{k+1}|\mathbf{X} \in \mathbf{A}_k)$ is also kept constant. This operation must be applied for each density f_k. The generated samples are unfortunately dependent and identically distributed according to f_k. Up to now, it is not possible to do this in an independent fashion. However, under mild conditions, it can be shown (Tierney, 1994) that applying this proposal/refusal method several times can decrease the variance.

The splitting algorithm for α-quantile estimation is relatively similar to Algorithm 9. Nevertheless, at each iteration k, instead of testing $\gamma_\rho^{(k)} < T$, we evaluate $(1-\rho)^k < \alpha$, and then the estimate $\hat{q}_\alpha^{\text{AST}}$ is the $\frac{\alpha}{(1-\rho)^k}$-quantile of the samples $\phi(\mathbf{X}_1^{(k)}), \ldots, \phi(\mathbf{X}_N^{(k)})$ in the last AST iteration.

AST is often applied to estimate very rare events ($\mathbb{P}_f < 10^{-4}$). For higher probabilities, other simulation methods such as IS are more efficient than AST (Cérou et al., 2012). The logarithmic efficiency has been proved for splitting with fixed levels in Cérou, Del Moral, & Guyader (2011).

5.4.2 Application to toy cases

Four-branch system

In order to illustrate the mechanisms involved in AST, it is applied to the four-branch system for two different values of T ($T = 10$ and $T = 20$) with a ρ-quantile set at 0.85 and 1000 samples per iteration. The corresponding results are provided in Table 5.17. In this case, three iterations of AST are necessary to obtain the estimation of the target probability. The evolution of the samples for the different iterations is provided in Figure 5.5. The two main steps of AST (i.e., selection of samples according to the ρ-quantile and resampling using the MH process) are represented in this figure. In this

Table 5.17 Results of AST for the four-branch system

T	$\hat{\mathbb{P}}^{\text{AST}}$	RB ($\hat{\mathbb{P}}^{\text{AST}}$)	RE ($\hat{\mathbb{P}}^{\text{AST}}$)	**Simulation budget**	ν^{AST}
10	2.2×10^{-3}	-0.9%	26%	4400	1.5
12	1.2×10^{-6}	1.7%	32%	37,000	226

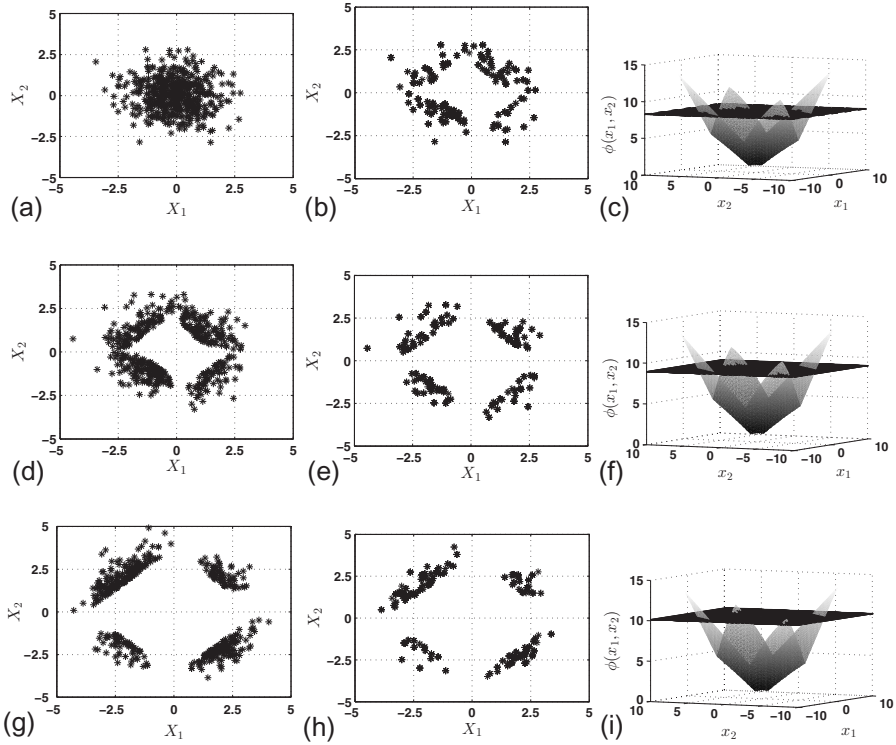

Figure 5.5 AST-generated samples and corresponding quantiles: (a) iteration 1—initial samples; (b) iteration 1—selection of samples; (c) iteration 1—intermediate quantile; (d) iteration 2—resampling; (e) iteration 2—selection of samples; (f) iteration 2—intermediate quantile; (g) iteration 3—resampling; (h) iteration 3—selection of samples; and (i) iteration 3—intermediate quantile.

example, AST succeeds in determining the four regions of the input space that lead to the threshold exceedance.

As seen in Table 5.17, the efficiency of AST relatively to CMC depends on the order of magnitude of the target probability. In general, the lower the target probability is, the higher the ν^{AST} coefficient will be.

Polynomial product

AST is applied to the polynomial product function with different dimensions d and thresholds T in order to test its ability to cope with high-dimensional problems. The results obtained for this toy case are summarized in Table 5.18 with a ρ-quantile set at 0.85 and 2000 samples per iteration. The target probabilities are indicated in Table 4.1.

AST does not suffer highly from the curse of dimensionality and can be applied with reliability on high-dimensional systems. We might also notice that the simulation budget of AST is not impacted by the dimension of the input vector. The required simulation budget essentially depends on the magnitude order of the probability to estimate and on the choice of the ρ-quantile parameter (Balesdent, Morio, & Marzat, 2015).

Table 5.18 Results of AST for the polynomial product toy case (for different dimensions and thresholds)

d	T	$\hat{\mathbb{P}}^{\text{AST}}$	RB ($\hat{\mathbb{P}}^{\text{AST}}$)	RE ($\hat{\mathbb{P}}^{\text{AST}}$)	**Simulation budget**	ν^{AST}
5	400	8.4×10^{-7}	−0.5%	23%	28,000	804
20	500	1.1×10^{-6}	0.9%	22%	33,000	569
50	700	3.5×10^{-7}	1.7%	22%	50,000	1180
200	1000	4.7×10^{-6}	−0.2%	21%	44,000	110

5.4.3 Conclusion

AST is a major algorithm for the estimation of rare event probabilities. It is adapted to a large number of test cases (multidimensional, multimodal failure region, etc.) even if its parameter tuning requires some experience. The characteristics of AST are summarized in Table 5.19. The application of AST on test cases is often mandatory for comparison with other efficient algorithms.

Table 5.19 Summary of AST for rare event probability estimation

Criteria	**AST characteristics**
Rarity of $\mathbb{P}_f$	Increased number of iterations needed to converge; increased efficiency relative to CMC
Simulation budget	At least 10^4 samples in most cases
Type of function $\phi(\cdot)$	No influence
Multimodality of f restricted to Ω_f	Particularly adapted to multimodal failure region
Dimension of $\mathbf{X}$	Still efficient on high-dimensional systems
Probability estimate error	Estimation with retrials
Difficulty of implementation	Medium to high; tuning the different AST parameters

References

Arnold, B. C., Balakrishnan, N., and Nagaraja, H. N. (1992). *A first course in order statistics*. New York, USA: Wiley.

Asmussen, S. (2003). *Applied probability and queues*. New York, USA: Springer.

Au, S. K. (2005). Reliability-based design sensitivity by efficient simulation. *Computers and Structures*, *83*, 1048–1061.

Au, S. K., and Beck, J. L. (2001). Estimation of small failure probabilities in high dimensions by subset simulations. *Probabilistic Engineering Mechanics*, *16*(4), 263–277.

Balesdent, M., Morio, J., and Marzat, J. (2015). Recommendations for the tuning of rare event probability estimators. *Reliability Engineering and System Safety*, *133*(0), 68–78.

Botev, Z. I., and Kroese, D. P. (2012). Efficient Monte-Carlo simulation via the generalized splitting method. *Statistics and Computing*, *22*(1), 1–16.

Bucklew, J. A. (2004). *Introduction to rare event simulation*. New York, USA: Springer.

Cannamela, C., Garnier, J., and Iooss, B. (2008). Controlled stratification for quantile estimation. *Annals of Applied Statistics*, *2*(4), 1554–1580.

Cérou, F., Del Moral, P., Furon, T., and Guyader, A. (2012). Sequential Monte Carlo for rare event estimation. *Statistics and Computing*, *22*, 795–808.

Cérou, F., Del Moral, P., and Guyader, A. (2011). A nonasymptotic theorem for unnormalized Feynman–Kac particle models. *Annales de l'Institut Henri Poincaré—Probabilités et Statistiques*, *47*(3), 629–649.

Cérou, F., Del Moral, P., Le Gland, F., and Lezaud, P. (2006). Genetic genealogical models in rare event analysis. *ALEA, Latin American Journal of Probability and Mathematical Statistics*, *1*, 181–203.

Daniels, H. E. (1954). Saddlepoint approximations in statistics. *The Annals of Mathematical Statistics*, *25*(4), 631–650.

de Mello, T. H. , and Rubinstein, R. Y. (2002). *Rare event estimation for static models via cross-entropy and importance sampling*. New York, USA: John Wiley.

Del Moral, P. (2004). *Feynman–Kac formulae, genealogical and interacting particle systems with applications. Probability and its applications.* New York, USA: Springer.

Dieker, A. B., and Mandjes, M. (2005). On asymptotically efficient simulation of large deviation probabilities. *Advances in Applied Probability*, *37*(2), 539–552.

Egloff, D., and Leippold, M. (2010). Quantile estimation with adaptive importance sampling. *The Annals of Statistics*, *38*, 1244–1278.

Engelund, S., and Rackwitz, R. (1993). A benchmark study on importance sampling techniques in structural reliability. *Structural Safety*, *12*(4), 255–276.

Fishman, G. S. (1996). *Monte-Carlo: Concepts, algorithms, and applications*. New York, USA: Springer.

Glasserman, P. (2003). *Monte-Carlo methods in financial engineering*. New York, USA: Springer.

Goutis, C., and Casella, G. (1999). Explaining the saddlepoint approximation. *The American Statistician*, *53*(3), 216–224.

Hammersley, J., and Handscomb, D. (1964). *Monte-Carlo methods*. London, UK: Methuen.

Heidelberger, P. (1995). Fast simulation of rare events in queueing and reliability models. *ACM Transactions on Modeling and Computer Simulation*, *5*(1), 43–85.

Huzurbazar, S. (1999). Practical saddlepoint approximations. *The American Statistician*, *53*(3), 225–232.

Kahn, H., and Harris, T. (1951). Estimation of particle transmission by random sampling. *National Bureau of Standards Applied Mathematical Series*, *12*, 27–30.

Kroese, D. P., and Rubinstein, R. Y. (2012). Monte-Carlo methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(1), 48–58.

Kullback, S., and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86.

Kumamoto, H., Tanaka, K., Inoue, K., and Henley, E. J. (1980). Dagger-sampling Monte Carlo for system unavailability evaluation. *Reliability, IEEE Transactions on*, *R-29*(2), 122–125.

Lagnoux, A. (2006). Rare event simulation. *Probability in the Engineering and Informational Science*, *20*(1), 45–66.

L'Ecuyer, P., Demers, V., and Tuffin, B. (2006). Splitting for rare event simulation. In *Proceedings of the 2006 winter simulation conference* (p. 137–148).

L'Ecuyer, P., Le Gland, F., Lezaud, P., and Tuffin, B. (2009). Splitting techniques. In G. Rubino and B. Tuffin (Eds.), *Rare event simulation using Monte Carlo methods* (p. 39–61). New York, USA: John Wiley & Sons, Ltd.

L'Ecuyer, P., Mandjes, M., and Tuffin, B. (2009). Importance sampling in rare event simulation. In G. Rubino and B. Tuffin (Eds.), *Rare event simulation using Monte Carlo methods* (p. 17–38). New York, USA: John Wiley & Sons, Ltd.

Meyn, S. P. (2007). *Control techniques for complex networks*. New York, USA: Cambridge University Press.

Mikhailov, G. A. (1999). *Parametric estimates by the Monte Carlo method*. Utrecht, the Netherlands: VSP.

Morio, J. (2012). Extreme quantile estimation with nonparametric adaptive importance sampling. *Simulation Modelling Practice and Theory*, *27*, 76–89.

Neddermeyer, J. C. (2009). Computationally efficient nonparametric importance sampling. *Journal of the American Statistical Association*, *104*, 788–802.

Neddermeyer, J. C. (2010). Non-parametric partial importance sampling for financial derivative pricing. *Quantitative Finance*, *11*, 1193–1206.

Niederreiter, H., and Spanier, J. (2000). *Monte Carlo and quasi-Monte Carlo methods*. Berlin, Germany: Springer.

Robert, C., and Casella, G. (2005). *Monte Carlo statistical methods*. New York, USA: Springer.

Rongfu, S., Chanan, S., Lin, C., and Yuanzhang, S. (2010). Short-term reliability evaluation using control variable based dagger sampling method. *Electric Power Systems Research*, *80*(6), 682–689.

Rubinstein, R. Y., and Kroese, D. P. (2004). *The cross-entropy method: A unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning*. Secaucus, USA: Springer-Verlag.

Sarkar, P. K., and Prasad, M. A. (1992). Variance reduction in Monte-Carlo radiation transport using antithetic variates. *Annals of Nuclear Energy*, *19*(5), 253–265.

Siegmund, D. (1976). Importance sampling in the Monte Carlo study of sequential tests. *The Annals of Statistics*, *4*(4), 673–684.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London, UK: Chapman & Hall.

Smarslok, B. P., Haftka, R. T., Carraro, L., and Ginsbourger, D. (2010). Improving accuracy of failure probability estimates with separable Monte Carlo. *International Journal of Reliability and Safety 4*(4), 393–414.

Smith, P. J., Shafi, M., and Gao, H. (1997). Quick simulation: A review of importance sampling techniques in communications systems. *Selected Areas in Communications, IEEE Journal on*, *15*(4), 597–613.

Sobol, I. M. (1994). *A primer for the Monte Carlo method*. Boca Raton, USA: CRC Press.
Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics*, *22*, 1701–1762.
Tokdar, S. T., and Kass, R. E. (2010). Importance sampling: A review. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(1), 54–60.
Tuffin, B., and Ridder, A. (2012). Probabilistic bounded relative error for rare event simulation learning techniques. In *Proceedings of the 2012 winter simulation conference* (p. 1–12).
Zhang, P. (1996). Nonparametric importance sampling. *Journal of the American Statistical Association*, *91*(435), 1245–1253.

This page intentionally left blank

Statistical techniques

J. Morio, D. Jacquemart, M. Balesdent

6

Statistical techniques enable to derive a probability estimate and associated confidence interval with a fixed set of samples $\phi(\mathbf{X}_1), \ldots, \phi(\mathbf{X}_N)$. The main statistical approaches, extreme value theory and large deviation theory, model the behavior of the pdf tails. We review in this chapter the theoretical features of these techniques.

6.1 Extreme value theory

Extreme value theory (EVT) (Embrechts, Kluppelberg, & Mikosch, 1994; Kotz & Nadarajah, 2000) characterizes the distribution tail of a random variable based on a finite reasonable number of observations. Thanks to its general applicative conditions, this theory has been widely used in the description of extreme meteorological phenomena (Towler et al., 2010), in finance and insurance (Embrechts et al., 1994; Reiss & Thomas, 2001), and in engineering (Castillo, Hadi, & Sarabia, 2005). EVT is notably very useful when only a fixed set of data can be used to estimate the probability $\mathbb{P}(\phi(\mathbf{X}) > T)$ for a given threshold T. We consequently assume in this chapter that a finite set of iid samples $\phi(\mathbf{X}_1), \ldots, \phi(\mathbf{X}_N)$ of the output is available and that we cannot generate new samples of $\phi(\mathbf{X})$. It is not mandatory that the samples $\phi(\mathbf{X}_i)$ be strictly independent, but this hypothesis is often considered for the sake of simplicity. Indeed, EVT can still be applied on correlated samples if the extreme samples are sufficiently separated in the sample sequence (Davis, Mikosch, & Zhao, 2013). In that case, extreme samples can then be considered to be independent.

6.1.1 Law of sample maxima

The associated ordered sample set is defined with $\phi(\mathbf{X}_{(1)}) \leq \phi(\mathbf{X}_{(2)}) \leq \cdots \leq \phi(\mathbf{X}_{(N)})$. The basic theorem of EVT (Embrechts et al., 1994; Gnedenko, 1943; Resnick, 1987) is that, under certain conditions, the maxima of an iid sequence converge to a generalized extreme value (GEV) distribution G_ξ, which results in the following cdf

$$G_\xi(x) = \begin{cases} \exp(-\exp(-x)), & \text{for } \xi = 0, \\ \exp\left(-(1+\xi x)^{-1/\xi}\right), & \text{for } \xi \neq 0 \end{cases}$$

The set of GEV distributions is composed of three distinct cdf types characterized by $\xi = 0$, $\xi > 0$, and $\xi < 0$ that correspond to the Gumbel, Fréchet, and Weibull distributions, respectively. Let us define G, the cdf of the iid samples $\phi(\mathbf{X}_1), \ldots, \phi(\mathbf{X}_N)$.

Estimation of Rare Event Probabilities in Complex Aerospace and Other Systems
http://dx.doi.org/10.1016/B978-0-08-100091-5.00006-X

Theorem 6.1.1 (Resnick, 1987)**.** *Let us suppose that there exist a_N and b_N with $a_N > 0$ such that, for all $y \in \mathbb{R}$,*

$$\mathbb{P}\left(\frac{\phi(\mathbf{X}_{(N)}) - b_N}{a_N} \leq y\right) = G^N(a_N y + b_N)$$

with

$$\lim_{N \to +\infty} G^N(a_N y + b_N) = G(y)$$

where G is a nondegenerate cdf, and G is a GEV distribution G_ξ. In this case, we denote $G \in MDA(\xi)$ where MDA stands for maximum domain of attraction.

The sequences a_N and b_N are given in Embrechts et al. (1994) for most well-known pdfs. An approximation of $\mathbb{P}(\phi(\mathbf{X}) > T)$ for large values of T and N may also be obtained (Embrechts et al., 1994):

$$\hat{\mathbb{P}}^{\text{EVT}} \approx \frac{1}{N}\left(1 + \xi\left(\frac{T - b_N}{a_N}\right)\right)^{-1/\xi} \tag{6.1}$$

The GEV approach is primarily used when only samples of maxima are available. In that case, the different parameters of the GEV distribution are obtained by determining maximum likelihood or probability weighted moment estimators. When samples of maxima are not available, one has to group the samples $\phi(\mathbf{X}_1), \ldots, \phi(\mathbf{X}_N)$ into blocks and fit the GEV using the maximum of each block (block maxima method). The main difficulty is in determining an efficient sample size for the different blocks.

6.1.2 Peak over threshold approach

6.1.2.1 Principle

Instead of grouping the samples into block maxima, peak over threshold (POT) considers the largest samples $\phi(\mathbf{X}_i)$ to estimate the probability $\mathbb{P}(\phi(\mathbf{X}) > T)$. The most common way of analyzing extremes with POT is to characterize the distribution of samples above a threshold u, which is given by the generalized Pareto cdf. The first paper linking the EVT with the distribution of a threshold exceedance was Pickands (1975). Later DeHaan obtained a result of the same type with a slightly simplified conclusion by using slowly varying functions (de Haan, 1984). The following theorem can then be obtained:

Theorem 6.1.2 (Pickands, 1975)**.** *Let us assume that the cdf G of iid samples $\phi(\mathbf{X}_1), \ldots, \phi(\mathbf{X}_N)$ is continuous. Set $y^* = \sup\{y, G(y) < 1\} = \inf\{y, G(y) = 1\}$. Then, the two following assertions are equivalent:*

1. *$G \in MDA(\xi)$,*
2. *A positive and measurable function $u \to \beta(u)$ exists such that*

$$\lim_{u \to y^*} \sup_{0 < y < y^* - u} |G^u(y) - H_{\xi,\beta(u)}(y)| = 0$$

where $G^u(y) = \mathbb{P}(\phi(\mathbf{X}) - u \leq y | \phi(\mathbf{X}) > u)$*, and* $H_{\xi,\beta(u)}$ *is the cdf of a generalized Pareto distribution (GPD) with shape parameter* ξ *and scale parameter* $\beta(u)$*.*

The expression of the GPD distribution function is given by

$$H_{\xi,\beta}(x) = \begin{cases} 1 - \exp\left(-\frac{x}{\beta}\right), & \text{for } \xi = 0, \\ 1 - \left(1 + \frac{\xi x}{\beta}\right)^{-1/\xi}, & \text{for } \xi \neq 0 \end{cases}$$

This theorem is in fact useful for estimating a probability of exceedance. Indeed, the probability $\mathbb{P}(\phi(\mathbf{X}) > T)$ can be rewritten as

$$\mathbb{P}(\phi(\mathbf{X}) > T) = \mathbb{P}(\phi(\mathbf{X}) > T | \phi(\mathbf{X}) > u)\mathbb{P}(\phi(\mathbf{X}) > u)$$

for $T > u$. A natural estimate of $\mathbb{P}(\phi(\mathbf{X}) > u)$ is given by the CMC estimate (see Section 5.1)

$$\hat{\mathbb{P}}^{\text{CMC}}(\phi(\mathbf{X}) > u) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\phi(\mathbf{X}_i) > u}$$

With Theorem 6.1.2 and for significant value of u, we obtain

$$\hat{\mathbb{P}}(\phi(\mathbf{X}) > T | \phi(\mathbf{X}) > u) = 1 - H_{\xi,\beta(u)}(T - u) \tag{6.2}$$

The estimate of $\mathbb{P}(\phi(\mathbf{X}) > T)$ is then built with

$$\hat{\mathbb{P}}^{\text{POT}} = \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\phi(\mathbf{X}_i) > u} \right) \times \left(1 - H_{\xi,\beta(u)}(T - u)\right) \tag{6.3}$$

The mathematical justification of Equation (6.2) and Equation (6.3) is discussed in D'Agostino & Stephens (1996), Fraga Alves & Ivette Gomes (1996), Kotz & Nadarajah (2000), and Drees, de Haan, & Li (2006) for a given set of samples to determine whether this set is suitable for the application of POT. Three parameters must be determined in the POT probability estimate of Equation (6.3): the threshold u and the couple of parameters $(\xi, \beta(u))$. The choice of u is very influential because it determines the samples that are used in the estimation of $(\xi, \beta(u))$. Indeed, a high threshold leads to consider only a small number of samples in the estimation of $(\xi, \beta(u))$; thus, their estimate can be spoiled by a large variance whereas a low threshold introduces a bias in the probability estimate (Dekkers & De Haan, 1999). Several methods have been proposed to determine a valuable threshold u knowing the samples. The most well-known ones are the Hill plot and the mean excess plot (Embrechts et al., 1994), that is defined by $(u, \mathbb{E}(Y - u | Y > u))$. These methods are nevertheless quite empirical because they are based on graphical interpretation. It is often necessary in practice to compare the estimates of u given by the different methods. Once a threshold estimate $\hat{u}$ is set, the GPD parameters are often estimated by maximum likelihood (Coles, 2001) or more occasionally by the method of moments

(Hosking & Wallis, 1987). The estimate $\hat{\mathbb{P}}^{\text{POT}}(\phi(\mathbf{X}) > T)$ given in Equation (6.3) for $T > u$ is then completely defined. A review of these different methods may be found in Neves & Fraga Alves (2004). POT is described in Algorithm 10.

ALGORITHM 10 Peak over threshold technique for probability estimation

Input: The samples $\phi(\mathbf{X}_1), \ldots, \phi(\mathbf{X}_N)$.
Output: The probability estimate $\hat{\mathbb{P}}^{\text{POT}}$.

1 Estimate $\hat{u}$ with Hill plot or the mean excess plot from $\phi(\mathbf{X}_1), \ldots, \phi(\mathbf{X}_N)$.
2 Estimate $(\hat{\xi}, \hat{\beta}(\hat{u}))$ by maximum likelihood or moment method with samples $\{\phi(\mathbf{X}_i), \text{with } \phi(\mathbf{X}_i) > \hat{u})\}$, $i = 1, \ldots, N$.
3 Estimate $\hat{\mathbb{P}}^{\text{POT}} = \left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\phi(\mathbf{X}_i)>\hat{u}}\right) \times \left(1 - H_{\hat{\xi},\hat{\beta}(\hat{u})}(T - \hat{u})\right)$.
4 **return** $\hat{\mathbb{P}}^{\text{POT}}$.

It is not possible, to our knowledge, to control the probability error estimate in EVT. Nevertheless, the use of bootstrap on samples $\phi(\mathbf{X}_1), \ldots, \phi(\mathbf{X}_N)$ (Geluk & de Haan, 2002) can provide some information on the accuracy of EVT.

α-quantiles of $\phi(\mathbf{X}_1), \ldots, \phi(\mathbf{X}_N)$ can also be obtained with POT. It can be shown that (McNeil & Saladin, 1997)

$$\hat{q}_\alpha^{\text{POT}} = \hat{u} + \frac{\hat{\beta}}{\hat{\xi}}\left(\left(\frac{N_{\hat{u}}}{N(1-\alpha)}\right)^{\hat{\xi}} - 1\right)$$

where $N_{\hat{u}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\phi(\mathbf{X}_i)>\hat{u}}$. In this case, $\hat{u}$ can be defined as a γ-quantile estimate of $\phi(\mathbf{X}_1), \ldots, \phi(\mathbf{X}_N)$ with $0 \ll \gamma < \alpha$.

6.1.2.2 Block maxima versus POT

POT takes into account all relevant high samples $\phi(\mathbf{X}_1), \ldots, \phi(\mathbf{X}_N)$ whereas the block maxima method can miss some of these high samples and, at the same time, consider some lower samples in its probability estimation. Thus, POT seems to be more appropriate for the design of the sample pdf tail. Nevertheless, the block maxima method is preferable when the available samples are not exactly iid or when only samples of maxima are available. For instance, the samples of a monthly river maximum height corresponds to this situation. Finally, the tuning of block maxima size turns out to be easier than the tuning of POT threshold u in many situations (Ferreira, Haan, et al., 2014). Because the output samples are always considered independent in the different test cases, the following discussion in this book considers only the application of the POT approach to estimate rare event probabilities with EVT.

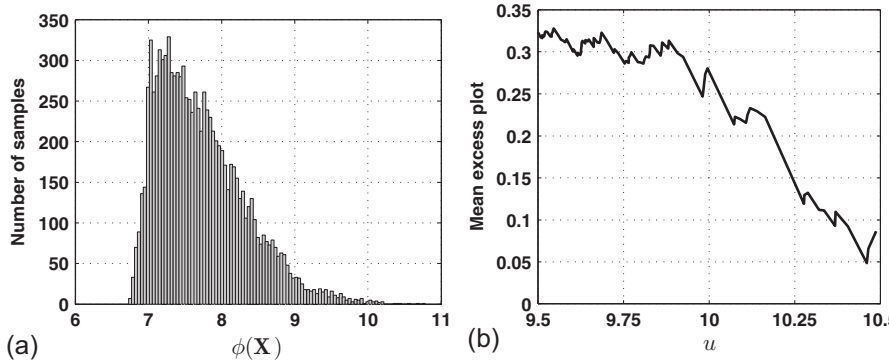

Figure 6.1 Probability estimation with EVT based on 10,000 CMC samples: (a) output histogram; (b) mean excess plot.

Table 6.1 Results of POT for the four-branch system

T	$\hat{\mathbb{P}}^{\text{POT}}$	RB ($\hat{\mathbb{P}}^{\text{POT}}$)	RE ($\hat{\mathbb{P}}^{\text{POT}}$)	**Simulation budget**	ν^{POT}
10	2.1×10^{-3}	-3%	35%	1000	4
10	2.3×10^{-3}	3%	9%	10,000	5
12	3.3×10^{-5}	2600%	324%	1000	n/a
12	2.9×10^{-6}	145%	242%	10,000	n/a

6.1.3 Application to a toy case

Four-branch system

POT is applied to the four-branch system. For that purpose, we generate a given number of Monte Carlo simulations of $\phi(\mathbf{X})$ and then apply POT to these samples. We estimate the parameter $\hat{u}$ with mean excess plot as illustrated in Figure 6.1 and evaluate the parameters $(\hat{\xi}, \hat{\beta}(\hat{u}))$ of the GPD with maximum likelihood. The results obtained with POT are presented in Table 6.1. POT improves CMC performances even if POT probability estimations can be biased for a rare event. When the simulation budget is low, the GPD parameters are not well estimated. Moreover, the gap between GPD distribution and the true model of pdf of $\phi(\mathbf{X})$ becomes also too large when the event rarity increases.

6.1.4 Conclusion

Extreme value theory is the only method that can be applied when resampling is not possible, that is, when we must deal with a fixed set of samples $\phi(\mathbf{X}_1), \ldots, \phi(\mathbf{X}_N)$. Nevertheless, one has to be cautious in the use of POT because a bias can appear in the probability estimation of rare events even if this estimation is theoretically available. The main characteristics of POT are presented in Table 6.2. The application of POT on complex systems is required if resampling is impossible.

Table 6.2 Summary of POT for rare event probability estimation

Criteria	POT characteristics
Rarity of $\mathbb{P}_f$	Possible bias with a weak simulation budget
Simulation budget	Can be theoretically applied with any sample size
Type of function $\phi(\cdot)$	No influence
Multimodality of f restricted to Ω_f	No influence
Dimension of $\mathbf{X}$	No influence
Probability estimate error	Estimation with bootstrap samples
Difficulty of implementation	Low; approximate estimation of $\hat{u}$ with graphical methods

6.2 Large deviation theory

The large deviation theory (LDT) characterizes the asymptotic behavior of sequence tails (Dembo & Zeitouni, 1998; den Hollander, 2008; Varadhan, 2008) and more precisely, it analyzes how a sequence tail deviates from its typical behavior described by the law of large numbers (see Section 2.3.1). Let us define S_N as an infinite sequence of random variables indexed by N. No assumptions about the dependency structure are made. We say that S_N satisfies the principle of large deviations with a continuous rate function I if the following limit exists for $\gamma > \mathbb{E}(S_N)$:

$$\lim_{N\to\infty} \frac{1}{N} \ln[\mathbb{P}(S_N > \gamma)] = -I(\gamma)$$

For a large value of N, the existence of this limit implies that

$$\mathbb{P}(S_N > \gamma) \approx \exp(-NI(\gamma))$$

The probability decays exponentially as N grows to infinity at a rate depending on γ. This approximation is a well-known result of LDT. If the limit does not exist, then $\mathbb{P}(S_N > \gamma)$ has a behavior that is too singular or decreases faster than exponential decay. If the limit is equal to 0, then the tail $\mathbb{P}(S_N > \gamma)$ decreases with N slower than $\exp(-Na)$ with $a > 0$. The computation of the rate function I is not obvious but can be obtained through the Gärtner–Ellis theorem (Touchette, 2009). Let us define the cumulant generating function $\lambda(\theta)$ of S_N with

$$\lambda(\theta) = \lim_{N\to\infty} \frac{1}{N} \ln\left[\mathbb{E}\left(\exp\left(N\theta S_N\right)\right)\right] \tag{6.4}$$

with $\theta \in \mathbb{R}$.

Theorem 6.2.1 (Gärtner–Ellis' theorem (Ellis, 1984; Gärtner, 1977))**.** *If the function* $\lambda(\theta)$ *of the variable* S_N *exists and is differentiable for all* $\theta \in \mathbb{R}$*, then* S_N *satisfies the principle of large deviations and* $I(\gamma)$ *is given by*

$$I(\gamma) = \sup_{\theta \in \mathbb{R}} [\theta\gamma - \lambda(\theta)]$$

In the specific case of a sum of iid random variables, we can derive the Cramér's theorem from Gärtner–Ellis' theorem (Touchette, 2009).

Theorem 6.2.2 (Cramér's theorem (Cramér, 1938))**.** *If* $S_N = \frac{1}{N}\sum_{i=1}^{N} Y_i$ *where the random variables* Y_i *are iid with the same probability law as* Y*, then* S_N *satisfies the principle of large deviations and the rate function is given by*

$$I(\gamma) = \sup_{\theta \in \mathbb{R}} [\theta\gamma - \lambda(\theta)]$$

with

$$\lambda(\theta) = \ln\left[\mathbb{E}\left(\exp\left(\theta Y\right)\right)\right]$$

This theorem holds only for light tail distributions. Let us consider the CMC probability estimate (see Section 5.1). In that case, we have $Y_i = \mathbf{1}_{\phi(\mathbf{X}_i)>T}$. The random variable Y_i follows a Bernoulli distribution of mean $\mathbb{P}_f = \mathbb{P}(\phi(\mathbf{X}) > T)$. The sequence S_N is defined with

$$S_N = \left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\phi(\mathbf{X}_i)>T}\right)$$

The functions $\lambda(\theta)$ and $I(\gamma)$ can be derived for some well-known pdfs. In the case of Bernoulli distributions of mean $\mathbb{P}_f$, we have

$$\lambda(\theta) = \ln\left(\mathbb{P}_f \exp(\theta) + 1 - \mathbb{P}_f\right)$$

and

$$I(\gamma) = \gamma \ln\left(\frac{\gamma}{\mathbb{P}_f}\right) + (1-\gamma)\ln\left(\frac{1-\gamma}{1-\mathbb{P}_f}\right)$$

The quantity $I(\gamma)$ corresponds to the relative entropy (Kullback–Leibler divergence) of two Bernoulli laws of parameter γ and $\mathbb{P}_f$. In many situations, the large deviation rate function corresponds to a Kullback–Leibler divergence between probability laws (Touchette, 2009). We can then obtain the convergence speed of the Monte Carlo probability estimate in function of the number of samples with the following equation for $\gamma > \mathbb{P}_f$:

$$\lim_{N\to\infty} \frac{1}{N} \ln[\mathbb{P}(S_N > \gamma)] = -\gamma \ln\left(\frac{\gamma}{\mathbb{P}_f}\right) - (1-\gamma) \ln\left(\frac{1-\gamma}{1-\mathbb{P}_f}\right) \tag{6.5}$$

6.2.1 Conclusion

LDT cannot in fact be applied directly to realistic test cases to determine a rare event probability in a situation when the density of Y is not known *a priori*. LDT can be useful to analyze the deviation of a probability estimate, notably if the probability estimate is a sum of random variables as shown in Equation (6.5) for the CMC estimate. Specific surveys on LDT can be found in Juneja & Shahabuddin (2006) and Blanchet & Lam (2012). Nevertheless, different methods are based on LDT considerations such as exponential twisting (see Section 5.3.2.2) or weighted importance resampling (see Section 9.6).

References

Blanchet, J., and Lam, H. (2012). State-dependent importance sampling for rare-event simulation: An overview and recent advances. *Surveys in Operations Research and Management Science*, *17*(1), 38–59.

Castillo, E., Hadi, A., and Sarabia, J. (2005). *Extreme value and related models with applications in engineering and science*. New Jersey, USA: John Wiley & Sons.

Coles, S. G. (2001). *An introduction to statistical modeling of extreme values*. New York, USA: Springer Verlag.

Cramér, H. (1938). Sur un nouveau théorème-limite de la théorie des probabilités (in French). *Actualités scientifiques et industrielles*, *736*, 5–23.

D'Agostino, R. B., and Stephens, M. A. (1996). Goodness-of-fit Techniques. In *Statistics: A series of textbooks and monographs* (Vol. 68). New York, USA: Dekker.

Davis, R. A., Mikosch, T., and Zhao, Y. (2013). Measures of serial extremal dependence and their estimation. *Stochastic Processes and their Applications*, *123*(7), 2575–2602.

de Haan, L. (1984). Slow variation and characterization of domains of attraction. In J. T. de Oliveira (Ed.), *Statistical extremes and applications* (Vol. 131, p. 31–48). Dordrecht, The Netherlands: Springer Netherlands.

Dekkers, A. L. M., and De Haan, L. (1999). On the estimation of the extreme-value index and large quantile estimation. *The Annals of Statistics*, *17*(4), 1795–1832.

Dembo, A., and Zeitouni, O. (1998). *Large deviations techniques and applications*. New York, USA: Springer-Verlag.

den Hollander, F. (2008). *Large deviations*. New York, USA: American Mathematical Society.

Drees, H., de Haan, L., and Li, D. (2006). Approximations to the tail empirical distribution function with application to testing extreme value conditions. *Journal of Statistical Planning and Inference*, *136*(10), 3498–3538.

Ellis, R. S. (1984). Large deviations for a general class of random vectors. *The Annals of Probability*, *12*(1), 1–12.

Embrechts, P., Kluppelberg, C., and Mikosch, T. (1994). Modelling extremal events for insurance and finance. *ZOR Zeitschrift for Operations Research Mathematical Methods of Operations Research*, *97*(1), 1–34.

Ferreira, A., Haan, L. de, et al. (2014). On the block maxima method in extreme value theory: PWM estimators. *The Annals of Statistics*, *43*(1), 276–298.

Fraga Alves, M. I., and Ivette Gomes, M. (1996). Statistical choice of extreme value domains of attraction—A comparative analysis. *Communications in Statistics—Theory and Methods*, *25*(4), 789–811.

Gärtner, J. (1977). On large deviations from the invariant measure. *Theory of Probability & Its Applications*, *22*(1), 24–39.

Geluk, J., and de Haan, L. (2002). *On bootstrap sample size in extreme value theory* (Tech. Rep. No. EI 2002-40). Erasmus University Rotterdam, Erasmus School of Economics (ESE).

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Annals of Mathematics*, *44*(3), 423–453.

Hosking, J., and Wallis, J. (1987). Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics*, *29*(3), 339–349.

Juneja, S., and Shahabuddin, P. (2006). Chapter 11 rare-event simulation techniques: An introduction and recent advances. In S. G. Henderson and B. L. Nelson (Eds.), *Simulation* (Vol. 13, p. 291–350). Amsterdam: Elsevier.

Kotz, S., and Nadarajah, S. (2000). *Extreme value distributions. Theory and applications*. London, UK: Imperial College Press.

McNeil, A. J., and Saladin, T. (1997). The peaks over thresholds method for estimating high quantiles of loss distributions. In *Proceedings of the* 27*th international ASTIN colloquium* (pp. 23–43).

Neves, C., and Fraga Alves, M. (2004). Reiss and Thomas' automatic selection of the number of extremes. *Computational Statistics and Data Analysis*, *47*(4), 689–704.

Pickands, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics*, *3*(1), 119–131.

Reiss, R. D., and Thomas, M. (2001). *Statistical analysis of extreme values, from insurance, finance, hydrology and other field*. Basel, Boston, Berlin: Birkhaüser Verlag.

Resnick, S. I. (1987). *Extreme values, regular variation and point processes*. New York, USA: Springer-Verlag.

Touchette, H. (2009). The large deviation approach to statistical mechanics. *Physics Reports*, *478*(1–3), 1–69.

Towler, E., Rajagopalan, B., Gilleland, E., Summers, R. S., Yates, D., and Katz, R. W. (2010). Modeling hydrologic and water quality extremes in a changing climate: A statistical approach based on extreme value theory. *Water Resources Research*, *46*(11).

Varadhan, S. R. S. (2008). Special invited paper: Large deviations. *The Annals of Probability*, *36*(2), 397–419.

This page intentionally left blank

Reliability based approaches

7

J. Morio, M. Balesdent

Researchers in reliability and system safety have proposed specific algorithms for rare event estimation methods. In most cases, these methods consider sampling strategies based on geometrical approximations of the limit state function $\{\mathbf{x}|\phi(\mathbf{x}) = T\}$. They may be more efficient than simulation techniques presented in the previous chapter but some assumptions may reduce their applicability.

7.1 First-order and second-order reliability methods

7.1.1 Principle

First- and second-order reliability method (FORM/SORM) (R. Bjerager, 1991; Lassen & Recho, 2006; Madsen, Krenk, & Lind, 1986; Yan-Gang & Tetsuro, 1999) are considered as efficient computational methods for structural reliability estimation. FORM/SORM give an analytical approximation of the curve $\{\mathbf{x}|T - \phi(\mathbf{x}) = 0\}$ (also called *limit state function*) at the most probable point of failure in the input space $\mathbf{x}$, that is, the point on the limit state function curve that has the highest probability content. It is also assumed that $\mathbf{X}$ follows a multivariate standard normal distribution. If this is not the case, several transformations can be applied on the input distribution (see Section 2.3.3.4). The most likely failure point, also called the *design point*, $\mathbf{x}^*$ is obtained by solving the following optimization problem

$$\begin{array}{ll} \text{minimize} & \|\mathbf{x}\|, \\ \text{with respect to} & \mathbf{x}, \\ \text{subject to} & T - \phi(\mathbf{x}) = 0 \end{array}$$

where $||.||$ is the Euclidean norm. The constraint $T - \phi(\mathbf{x}) = 0$ defines the limit of failure space for input vector $\mathbf{x}$. The parameter $\beta = \|\mathbf{x}^*\|$ is the reliability index. Several algorithms have been proposed to find $\mathbf{x}^*$ and solve this optimization problem as proposed in Dietlevsen & Madsen (1996); Hasofer & Lind (1974); Pei-Ling & Kiureghian (1991); Rackwitz & Flessler (1978). The surface $\{\mathbf{x}|T - \phi(\mathbf{x}) = 0\}$ at the solution $\mathbf{x}^*$ is approximated as first-order by a hyperplane in the case of FORM by using the Taylor series expansion. Accuracy problems can occur when this surface is strongly nonlinear. Thus, SORM has been established as an attempt to improve the accuracy of FORM because SORM approximates the limit of failure space at the design point by a quadratic surface. The rare event probability is then estimated with FORM by

$$\hat{\mathbb{P}}^{\text{FORM}} = \Phi_{0,1}(-\beta)$$

Estimation of Rare Event Probabilities in Complex Aerospace and Other Systems
http://dx.doi.org/10.1016/B978-0-08-100091-5.00007-1

Recall that $\Phi_{0,1}(\cdot)$ is the cdf of the standard normal distribution. In the case of SORM, the failure probability is given by Breitung (1984)

$$\hat{\mathbb{P}}^{\text{SORM}} = \Phi_{0,1}(-\beta) \prod_{i=1}^{d-1} (1 - \beta\kappa_i)^{-1/2}$$

where κ_i denotes the principal curvature of $T - \phi(\mathbf{x})$ at the design point β. The term κ_i is defined with

$$\kappa_i = \left. \frac{\partial^2 (T - \phi(\mathbf{x}))}{\partial^2 x^i} \right|_{\mathbf{x}=\mathbf{x}^*}$$

FORM/SORM is described in Algorithm 11. A first-order saddle point approximation (FOSPA) (Du & Sudjianto, 2004; Huanh & Du, 2008) method has also been proposed as an improvement of FORM/SORM. It consists of using LDT and the saddle point approximation (Daniels, 1954; Goutis & Casella, 1999; Huzurbazar, 1999; Jensen, 1995), but it is rarely applied in practice. In the same way, the method of moments (Huang & Du, 2006) that consists in a approximating the cdf of $\mathbf{1}_{\phi(\mathbf{X})>T}$ with its moments can sometimes be an alternative when the design point is difficult to find.

FORM/SORM do not require a large simulation budget to obtain a valuable result. Nevertheless, their different assumptions require that one has to be careful when one applies FORM/SORM to a realistic case of function $\phi(\cdot)$, notably if the most probable point of failure is not unique (Sudret, 2012). Nevertheless, a heuristic method has been developed to find the multiple design points of a reliability problem when they exist (Der Kiureghian & Dakessian, 1998). It consists in penalizing previously found solutions to force the algorithm to find a new solution. Once these points are found, FORM or SORM approximations are constructed at these points and the probability of interest is computed as the probability of the union of the approximated events.

ALGORITHM 11 FORM (or SORM) technique for probability estimation

Input: The pdf f, the function $\phi(\cdot)$, the threshold T.
Output: The probability estimate $\hat{\mathbb{P}}^{FORM}$ (or $\hat{\mathbb{P}}^{SORM}$).

1 Evaluate $\beta = \|\mathbf{x}^*\|$ with $\mathbf{x}^* = \underset{\mathbf{x}}{\text{argmin}} ||\mathbf{x}||$ subject to $T - \phi(\mathbf{x}) = 0$.

2 Estimate $\hat{\mathbb{P}}^{FORM} = \Phi_{0,1}(-\beta)$ (or $\hat{\mathbb{P}}^{SORM} = \Phi_{0,1}(-\beta) \prod_{i=1}^{d-1} (1 - \beta\kappa_i)^{-\frac{1}{2}}$ with

$$\kappa_i = \left. \frac{\partial^2 (T - \phi(\mathbf{x}))}{\partial^2 x^i} \right|_{\mathbf{x}=\mathbf{x}^*}).$$

return $\hat{\mathbb{P}}^{FORM}$ ***(or*** $\hat{\mathbb{P}}^{SORM}$***).***

There is also no control of the error in FORM/SORM. It is possible from FORM/SORM to determine an importance sampling auxiliary density around the design point and then to sample with it to estimate the rare event probability (de Rocquigny, Devictor, & Tarantola, 2008).

Even if it is possible in theory, it is also important to notice that FORM/SORM are not efficient in practice to estimate quantiles of $\phi(\mathbf{X})$ (de Rocquigny et al., 2008).

7.1.2 Application to toy cases

Polynomial square root function

FORM and SORM are applied to the polynomial square root function. The results obtained with these two methods are given in Table 7.1.

In this test case, SORM is successful in performing the estimation whereas FORM overestimates the target probability. These results can be explained by the relatively high curvature of the limit state (i.e., $\left\{\mathbf{x} \in \mathbb{R}^d | \phi(\mathbf{x}) - T = 0\right\}$). The linear approximation of the failure region used in FORM is not valid on this toy case and leads to the biased obtained results. Let us also notice that because these methods are deterministic, $RE(\hat{\mathbb{P}})$ is not computable. The different points evaluated on $\phi(\cdot)$ during the search of the most probable point are illustrated in Figure 7.1.

Table 7.1 Results obtained with FORM and SORM for the polynomial square root toy case

Method	$\hat{\mathbb{P}}$	Simulation budget	RB($\hat{\mathbb{P}}$)
FORM	8.83×10^{-6}	894	+4104%
SORM	2.40×10^{-7}	899	+14%

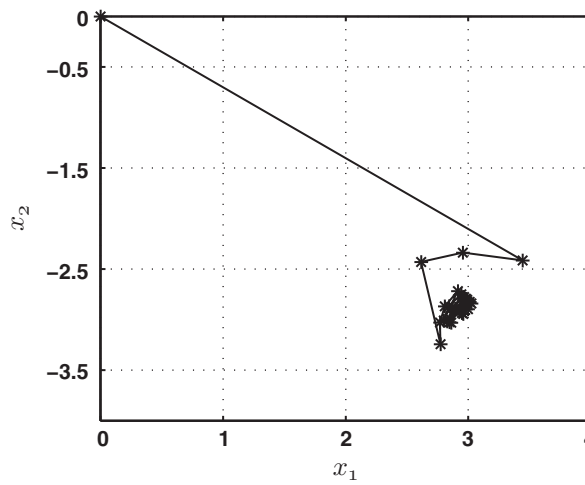

Figure 7.1 Points evaluated on $\phi(\cdot)$ during the search of the most probable point.

Four-branch system

FORM and SORM are also applied to the four-branch system. Results are given in Table 7.2. For this multimodal test case, neither FORM nor SORM allow making an accurate estimate of the target probability. This is principally the result of the multimodality of this test case. Indeed, classical FORM and SORM find only the most probable point and approximate the total probability from it ([3.54, 3.54] in this example, Figure 7.2). In other words, FORM and SORM consider only one of the four branches when estimating the probability. This explains the low values of the probability estimates with respect to the true probability.

7.1.3 Conclusion

FORM/SORM estimate rare event probabilities by approximating the limit state of failure domain $\{\mathbf{x}|\phi(\mathbf{x}) - T = 0\}$. These methods can be efficiently applied with a low simulation budget, but one has to be careful in the validity of the different FORM/SORM assumptions to avoid a bias in the probability estimation. The main characteristics of FORM/SORM are summarized in Table 7.3. Applying FORM/-SORM to practical cases is often interesting because a low simulation budget is required to reach the algorithm convergence.

Table 7.2 Results obtained with FORM and SORM for the four-branch system toy case

Method	T	$\hat{\mathbb{P}}$	Simulation budget	RB($\hat{\mathbb{P}}$)
FORM	12	2.8×10^{-7}	12	−77%
SORM	12	1.7×10^{-7}	56	−86%

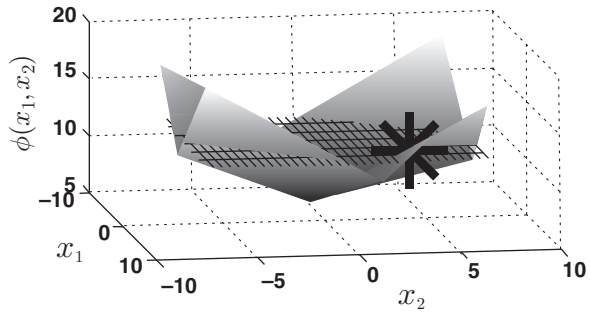

Figure 7.2 Four-branch system, threshold (12) and most probable point (∗).

Table 7.3 Summary of FORM/SORM for rare event probability estimation

Criteria	FORM/SORM characteristics
Rarity of $\mathbb{P}_f$	No clear influence on the simulation budget
Simulation budget	From 10 to 10^3 samples in most cases
Type of function $\phi(\cdot)$	Regularity and differentiability assumptions
Multimodality of f restricted to Ω_f	Not adapted; bias in the estimate if all the multiple design points are not found
Dimension of $\mathbf{X}$	Can induce a decrease of the estimator efficiency (depending on the regularity of $\phi(\cdot)$)
Probability estimate error	Unknown
Difficulty of implementation	Low; optimization required to determine β

7.2 Line sampling

7.2.1 Principle

The underlying idea of line sampling (LS) (P. Koutsourelakis, 2004; P. S. Koutsourelakis, Pradlwarter, & Schueller, 2004; Schueller, Pradlwarter, & Koutsourelakis, 2004) is to employ lines instead of random points in order to probe the input failure domain of the system, that is, $\mathbf{x}$ so that $\phi(\mathbf{x}) > T$. LS has to be applied on input random vector $\mathbf{X}$ that has multivariate standard normal density. If that is not the case, several transformations can be applied on the input distribution (see Section 2.3.3.4). Let us define the set $\mathbf{A} = \{\mathbf{x} \in \mathbb{R}^d | \phi(\mathbf{x}) > T\}$. The set $\mathbf{A}$ can be also expressed in the following way:

$$\mathbf{A} = \{\mathbf{x} \in \mathbb{R}^d | x^1 \in A_1(\mathbf{x}^{-1})\} \tag{7.1}$$

where the function A_1 is a function defined on $\mathbb{R}^{d-1}$ that depends on $\mathbf{x}^{-1} = (x^2, x^3, \ldots, x^d)^T$. The output of A_1 is a subset of $\mathbb{R}$. Similar functions A_1 may be defined with respect to any direction in the random parameter space and for all measurable $\mathbf{A}$. In practice, we determine a unit important direction vector $\vec{\alpha} \in \mathbb{R}^d$. It is the direction that enables to reach the curve $T - \phi(\mathbf{x}) = 0$ with the shortest path to the origin. This direction can be found with Monte Carlo Markov chain methods (Pradlwarter et al., 2005) or with the design point $\mathbf{x}^*$ given by FORM/SORM (see the previous section). Then we can decompose $\mathbf{x}$ in the following way:

$$\mathbf{x} = \mathbf{x}^{\perp} + \mathbf{x}^{\parallel}$$

where $\mathbf{x}^{\parallel}$ is the projection of $\mathbf{x}$ on $\vec{\alpha}$, that is, $\mathbf{x}^{\parallel} = r\vec{\alpha}$, and thus $\mathbf{x}^{\perp} = \mathbf{x} - r\vec{\alpha}$. r is the scalar product between $\vec{\alpha}$ and $\mathbf{x}$. By applying a coordinate transformation to $\mathbf{x}$, we can modify $\mathbf{x}$ so that $\vec{\alpha}$ corresponds to the first coordinate of the new input coordinates. In that case, we have $x^1 = r$ and $\mathbf{x}^{-1} = \mathbf{x}^{\perp}$. Without loss of generality and to obtain more readable equations, we assume that this is the considered situation.

The rare event probability $\mathbb{P}_f$ can be written with integrals in the following way:

$$\begin{aligned}\mathbb{P}_f &= \int_{\mathbb{R}^d} \mathbf{1}_{\phi(\mathbf{x})>T} f(\mathbf{x})\,\mathrm{d}\mathbf{x},\\ &= \int_{\mathbb{R}^d} \mathbf{1}_{\mathbf{x}\in\mathbf{A}} f(\mathbf{x})\,\mathrm{d}\mathbf{x},\\ &= \int_{\mathbb{R}^{d-1}} \int_{\mathbb{R}} \mathbf{1}_{x^1\in A_1(\mathbf{x}^{-1})} f(\mathbf{x})\,\mathrm{d}x^1\,\mathrm{d}\mathbf{x}^{-1}\end{aligned}$$

It can then be rewritten with mathematical expectation over the variable $\mathbf{X}^{-1}$ with

$$\mathbb{P}_f = \mathbb{E}\left(\mathbb{P}(X^1 \in A_1(\mathbf{X}^{-1}))\right) \tag{7.2}$$

The rare event probability is described as the expectation of the continuous random variable $\mathbb{P}(X^1 \in A_1(\mathbf{X}^{-1}))$ relative to the random vector $\mathbf{X}^{-1}$. This expectation is replaced practically in LS by its Monte Carlo estimate:

$$\hat{\mathbb{P}}^{\mathrm{LS}} = \frac{1}{N_C}\sum_{i=1}^{N_C}(\mathbb{P}(X^1 \in A_1(\mathbf{X}_i^{-1}))) \tag{7.3}$$

where $(\mathbf{X}_1^{-1}), \ldots, (\mathbf{X}_{N_C}^{-1})$ are iid samples of the rv $\mathbf{X}^{-1}$. It is still necessary to estimate the probability $\mathbb{P}(X^1 \in A_1(\mathbf{X}_i^{-1}))$, that is

$$\mathbb{P}(X^1 \in A_1(\mathbf{X}_i^{-1})) = \int_{\mathbb{R}} \mathbf{1}_{x^1\in A_1(\mathbf{x}_i^{-1})} f_{X^1}(x^1)\,\mathrm{d}x^1 \tag{7.4}$$

where we recall that f_{X^1} is an univariate standard normal variable. It is possible to show that this integral can be approximated (P. S. Koutsourelakis et al., 2004) in the same way as it is done in FORM with

$$\mathbb{P}(X^1 \in A_1(\mathbf{X}_i^{-1})) \approx \Phi_{0,1}(-c_i) \tag{7.5}$$

where c_i is the value of x^1 such that $\phi(c_i, \mathbf{x}_i^{-1}) = T$. This approximation is valuable only if there is only one intersection point between the input failure region and the chosen sampling direction. The variance of the LS estimate is always lower than or equal to the CMC estimation (P. S. Koutsourelakis et al., 2004). The LS algorithm is described in Algorithm 12. To our knowledge, LS has never been applied in practice to estimate quantiles of $\phi(\mathbf{X})$.

7.2.2 Application to toy cases

Polynomial square root function

LS is applied to the polynomial square root function. The unit important direction is determined with 5 SORM simulations with $\vec{\alpha} = (3, -3)^T/\sqrt{(18)}$. Moreover, the

ALGORITHM 12 Line Sampling algorithm for probability estimation

Input: The pdf f of $\mathbf{X}$, the function $\phi(\cdot)$, and the threshold T.
Output: The probability estimate $\hat{\mathbb{P}}^{\text{LS}}$.

1 Determine the unit important direction vector $\vec{\alpha} \in \mathbb{R}^d$.
2 Generate N_C iid samples $\mathbf{X}_1, \ldots, \mathbf{X}_{N_C}$ of the variable $\mathbf{X}$.
3 Determine the components $\mathbf{X}_1^{\perp}, \ldots, \mathbf{X}_{N_C}^{\perp}$ and $\mathbf{X}_1^{\parallel}, \ldots, \mathbf{X}_{N_C}^{\parallel}$ from $\mathbf{X}_1, \ldots, \mathbf{X}_{N_C}$.
4 **for** $i \leftarrow 1$ ***to*** N_C **do**
5 Determine c_i such as $\phi(\mathbf{X}_i^{\perp} + c_i\vec{\alpha}) = T$.
6 Estimate $\hat{\mathbb{P}}(X^1 \in \mathbf{A}_1(\mathbf{X}_i^{-1})) = \Phi_{0,1}(-c_i)$.
7 Estimate $\hat{\mathbb{P}}^{\text{LS}} = \frac{1}{N_C}\sum_{i=1}^{N_C}(\hat{\mathbb{P}}(X^1 \in \mathbf{A}_1(\mathbf{X}_i^{-1})))$.
8 **return** $\hat{\mathbb{P}}^{\text{LS}}$.

Table 7.4 Results obtained with LS for the polynomial square root test case

$\hat{\mathbb{P}}^{\text{LS}}$	$\text{RB}(\hat{\mathbb{P}}^{\text{LS}})$	$\text{RE}(\hat{\mathbb{P}}^{\text{LS}})$	**Simulation budget**	ν^{LS}
2.34×10^{-6}	-0.4%	33%	200	2.2×10^6
2.33×10^{-6}	-1%	12%	2000	1.6×10^4
2.34×10^{-6}	-0.4%	3%	20,000	2.2×10^4

determination of c_i for each sample $\mathbf{X}_i$ is made by dichotomy with 20 calls to the function $\phi(\cdot)$. The results obtained with LS are given in Table 7.4. An example of 2000 LS samples is presented in Figure 7.3. This toy case is well adapted to line sampling because it outperforms the results that can be obtained with CMC. One LS probability estimation is a mean of N_C probability estimations and leads to a low relative error for a given simulation budget.

Four-branch system

LS is also applied to the four-branch function. The unit's important direction is the one given by FORM with 12 calls to the function $\phi(\cdot)$ with $\vec{\alpha} = (3.54, 3.54)^T$. The parameter c_i is still estimated by a dichotomy with 20 samples. The corresponding results are summarized in Table 7.5. In the same way as FORM/SORM, LS cannot cope with multimodal optimal auxiliary distribution. It can concentrate on only one of the failure modes, and the resulting probability estimate is thus biased.

7.2.3 Conclusion

LS enables to estimate rare event probabilities by sampling on lines parallel to an important direction toward the failure domain. It can be considered as a complementary

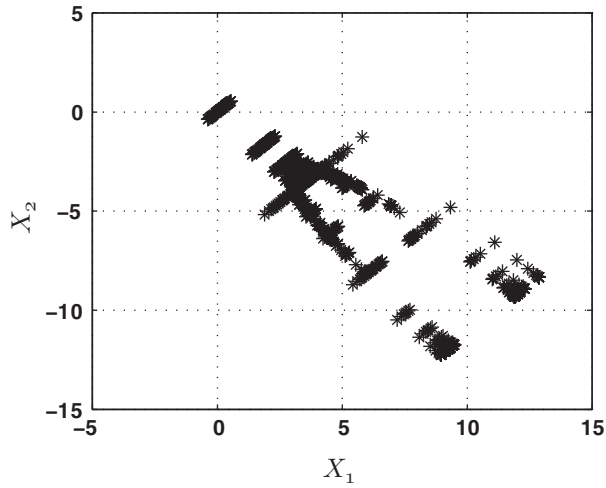

Figure 7.3 A run of LS with 2000 samples on the polynomial square root function.

Table 7.5 Results obtained with LS for the polynomial square root test case

$\hat{\mathbb{P}}^{\mathrm{LS}}$	$\mathrm{RB}(\hat{\mathbb{P}}^{\mathrm{LS}})$	$\mathrm{RE}(\hat{\mathbb{P}}^{\mathrm{LS}})$	**Simulation budget**	ν^{LS}
1.7×10^{-7}	−85%	22%	200	n/a
1.7×10^{-7}	−85%	6%	2000	n/a
1.6×10^{-7}	−86%	1.5%	20,000	n/a

method to FORM/SORM with a control of the probability error because FORM/SORM can be used to find the important direction to sample. When LS is adapted, its probability estimates have a high accuracy even when the simulation budget is low because this estimate corresponds to a mean of failure probabilities. As with FORM/SORM, LS is not able to cope with multimodal optimal sampling distribution. The main characteristics of LS are summarized in Table 7.6. The application of LS on the test cases in part Three of the book could be of interest if FORM/SORM is applicable.

7.3 Directional sampling

7.3.1 Principle

Directional sampling (DS) (P. Bjerager, 1988), also called *directional simulation*, can be viewed as a joint use of Monte Carlo simulations and line sampling. It

Table 7.6 Summary of LS for rare event probability estimation

Criteria	LS characteristics
Rarity of $\mathbb{P}_f$	Low influence on the simulation budget
Simulation budget	From 100 to 10^4 samples in most cases
Type of function $\phi(\cdot)$	Regularity and differentiability assumptions
Multimodality of f restricted to Ω_f	Not adapted; bias in the estimate due to unadapted $\vec{\alpha}$
Dimension of $\mathbf{X}$	Increase the difficulty to find $\vec{\alpha}$; if $\vec{\alpha}$ is well determined, then LS is able to cope with high dimensions
Probability estimate error	Estimation with retrials
Difficulty of implementation	Medium as soon as $\vec{\alpha}$ is known

must be applied on an input random vector $\mathbf{X}$ that has multivariate standard normal distribution. If this is not the case, several transformations can be applied on the input distribution (see Section 2.3.3.4).

The vector $\mathbf{X}$ can be expressed as

$$\mathbf{X} = R\mathbf{A}$$

where R^2 is a chi-squared random variable with d degrees of freedom with density f_R and $\mathbf{A}$ is a random unit vector uniformly distributed on the d-dimension unit sphere Λ^d with density $f_{\mathbf{A}}$. R^2 and $\mathbf{A}$ are independent random variables. The rare event probability $\mathbb{P}_f$ can be then written as follows:

$$\begin{aligned}\mathbb{P}_f &= \int_{\mathbf{A}\in\Lambda^n}\int_0^{+\infty} \mathbf{1}_{\phi(r\mathbf{a})>T} f_R(r) f_{\mathbf{A}}(\mathbf{a})\,\mathrm{d}\mathbf{a}\,\mathrm{d}r,\\ &= \int_{\mathbf{A}\in\Lambda^n} \mathbb{P}(\phi(R\mathbf{A}) > T|\mathbf{A}=\mathbf{a}) f_{\mathbf{A}}(\mathbf{a})\,\mathrm{d}\mathbf{a}\end{aligned}$$

The DS estimation consists of a CMC estimation of $\mathbb{P}(\phi(R\mathbf{A}) > T|\mathbf{A}=\mathbf{a})$. In practice, a sequence of N iid random direction vectors $\mathbf{A}_j$ for $j = 1,\ldots,N$ is generated and then we determine r_j such that $\phi(r_j\mathbf{A}_j) = T$ by dichotomy, for instance. An estimate of $\mathbb{P}(\phi(R\mathbf{A}) > T|\mathbf{A}=\mathbf{A}_j)$ is given by $1 - F_{R^2}(r_j^2)$ where F_{R^2} is the cdf of a chi-squared random variable with d degrees of freedom. This approximation is valuable only if there is one intersection point between the input failure region and the chosen sampling direction. The DS probability estimate $\hat{\mathbb{P}}^{\mathrm{DS}}$ is then obtained with

$$\hat{\mathbb{P}}^{\mathrm{DS}} = \frac{1}{N}\sum_{j=1}^{N}\left(1 - F_{R^2}(r_j^2)\right)$$

ALGORITHM 13 Directional sampling algorithm for probability estimation

Input: The pdf f of $\mathbf{X} = (R\mathbf{A})$, the function $\phi(\cdot)$, and the threshold T.
Output: The probability estimate $\hat{\mathbb{P}}^{\mathrm{DS}}$.

1 Generate N iid samples $\mathbf{A}_1, \ldots, \mathbf{A}_N$ of variable $\mathbf{A}$.
2 Determine $r_1, \ldots, r_N$ such that $\phi(r_j \mathbf{A}_j) = T$ for $j = 1, \ldots, N$ by optimization.
3 Estimate $\hat{\mathbb{P}}^{\mathrm{DS}} = \frac{1}{N} \sum_{j=1}^{N} \left(1 - F_{R^2}(r_j^2)\right)$.
4 **return** $\hat{\mathbb{P}}^{\mathrm{DS}}$.

The DS algorithm is described in Algorithm 13. To our knowledge, DS has never been applied in practice to estimate quantiles of $\phi(\mathbf{X})$.

7.3.2 Application to toy cases

Four-branch system

DS is applied to the four-branch function. Several numbers of sampling directions are tested and involve different simulation budgets. The results obtained with this method for different numbers of directions are given in Table 7.7. DS outperforms CMC simulations on this test case whatever the simulation budget as show the values of $RE(\hat{\mathbb{P}}^{\mathrm{DS}})$ and ν^{DS}. A set of samples obtained with DS is presented in Figure 7.4. Nevertheless, notice that the probability estimate could be biased when too low a number of sampling directions is chosen. One might also notice that DS is able to cope with disconnected failure regions Ω_f.

Polynomial product

DS is also applied to the polynomial product function, with different dimensions d and thresholds T in order to test its ability to cope with high-dimensional problems. The results obtained for this toy case are summarized in Table 7.8. 10,000 directions are generated for estimating all the target probabilities. This number of directions is sufficient and results in unbiased estimated probabilities. The accuracy of DS is

Table 7.7 Results obtained with DS for the four-branch system test case

Number of directions	$\hat{\mathbb{P}}^{\mathrm{DS}}$	$\mathrm{RB}(\hat{\mathbb{P}}^{\mathrm{DS}})$	$\mathrm{RE}(\hat{\mathbb{P}}^{\mathrm{DS}})$	Simulation budget	ν^{DS}
10	1.1×10^{-6}	−7%	48%	171	21,781
100	1.2×10^{-6}	2%	18%	1701	14,788
1000	1.2×10^{-6}	2%	4%	17,001	26,834
10,000	1.2×10^{-6}	1%	2%	170,001	21,295

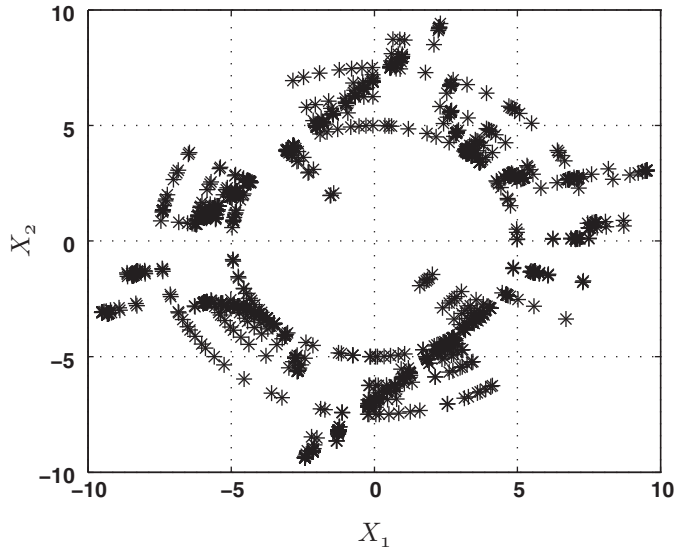

Figure 7.4 A run of DS with 100 directions on the four-branch function.

Table 7.8 Results of directional sampling for the polynomial product toy case (for different dimensions and thresholds)

d	T	$\hat{\mathbb{P}}^{\mathrm{DS}}$	RB($\hat{\mathbb{P}}^{\mathrm{DS}}$)	RE($\hat{\mathbb{P}}^{\mathrm{DS}}$)	**Simulation budget**	ν^{DS}
5	400	8.5×10^{-7}	+1%	4%	204,301	3831
20	500	1.05×10^{-6}	−5%	29%	238,601	46
50	700	3.4×10^{-7}	−2%	90%	220,001	16
200	1000	4.6×10^{-6}	−2%	97%	190,001	12

influenced by the dimension of the problem. It still performs better than CMC on the same problem, but this gap decreases when the dimension increases.

7.3.3 Conclusion

The DS method generates samples over a large number of different directions in the input space and estimates the target probability as a mean of the probabilities obtained for each sampling direction. The efficiency and accuracy of DS depend directly on the number of generated directions as shown in the results of the toy cases. The characteristics of DS are given in Table 7.9. DS can be efficiently applied to complex cases.

Table 7.9 Summary of DS for rare event probability estimation

Criteria	DS characteristics
Rarity of $\mathbb{P}_f$	Increased number of required sampling directions to avoid a biased probability
Simulation budget	At least 10^3 samples; depends on the number of sampling directions
Type of function $\phi(\cdot)$	Regularity and differentiability assumptions
Multimodality of f restricted to Ω_f	No influence
Dimension of $\mathbf{X}$	Not adapted to high-dimensional systems
Probability estimate error	Estimation with retrials
Difficulty of implementation	Low; tuning of the number sampling directions

7.4 Stratified sampling

7.4.1 Principle

The principle of stratified sampling (SS) is very similar to CMC (Cochran, 1977) but with the idea of partitioning the input space. An extended version of SS called the *coverage Monte Carlo method* (Karp & Luby, 1983; Kumamoto, Tanaka, & Inoue, 1987) has been proposed for very specific systems represented by a fault tree or a network using its minimal cuts to improve the probability estimation. For the same kind of systems, recursive variance reduction methods described in Cancela & El Khadiri (1995, 2003) have also been proposed and have some links with SS.

Stratified sampling consists in partitioning support of $\mathbf{X}$, defined by $\mathbb{R}^d$ in several subsets $\mathbb{Q}_i$, $i = 1, \ldots, m$ such that $\mathbb{Q}_i \bigcap \mathbb{Q}_j = \emptyset$ for $i \neq j$ and $\bigcup_i \mathbb{Q}_i = \mathbb{R}^d$. For a given subset $\mathbb{Q}_i$, we then generate n_i iid samples $\mathbf{X}_1^i, \ldots, \mathbf{X}_{n_i}^i$ from the pdf $h_{\mathbb{Q}_i}$ defined by

$$h_{\mathbb{Q}_i}(\mathbf{x}) = \mathbf{1}_{\mathbf{x} \in \mathbb{Q}_i} \frac{f(\mathbf{x})}{d_i}$$

where d_i is defined by

$$d_i = \int_{\mathbb{Q}_i} f(\mathbf{x})\, \mathrm{d}\mathbf{x}$$

The required number of samples N in SS is thus equal to $N = \sum_{i=1}^{m} n_i$. The principle of stratified sampling is to use conditional probabilities and the law of total probability (see Chapter 2) to estimate $\mathbb{P}_f$ such that

$$\mathbb{P}_f = \sum_{i=1}^{m} \mathbb{P}(\phi(\mathbf{X}) > T | \mathbf{X} \in \mathbb{Q}_i) \mathbb{P}(\mathbf{X} \in \mathbb{Q}_i)$$

The SS probability estimate $\hat{\mathbb{P}}^{\text{SS}}$ is then obtained in the following way:

$$\begin{aligned} \hat{\mathbb{P}}^{SS} &= \sum_{i=1}^{m} \hat{\mathbb{P}}(\phi(\mathbf{X}) > T | \mathbf{X} \in \mathbb{Q}_i) \mathbb{P}(\mathbf{X} \in \mathbb{Q}_i), \\ &= \sum_{i=1}^{m} d_i \hat{\mathbb{P}}_{h_{\mathbb{Q}_i}} \end{aligned}$$

where $\hat{\mathbb{P}}_{h_{\mathbb{Q}_i}}$ is defined as

$$\hat{\mathbb{P}}_{h_{\mathbb{Q}_i}} = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{1}_{\phi(\mathbf{X}_j^i) > T}$$

The estimator $\hat{\mathbb{P}}^{\text{SS}}$ is unbiased and its variance $\sigma^2_{\hat{\mathbb{P}}^{\text{SS}}}$ depends notably on n_i and $h_{\mathbb{Q}_i}$, and is given by the following equation (Keramat & Kielbasa, 1998):

$$\sigma^2_{\hat{\mathbb{P}}^{\text{SS}}} = \sum_{i=1}^{m} d_i \frac{\mathbb{P}_{h_{\mathbb{Q}_i}}(1 - \mathbb{P}_{h_{\mathbb{Q}_i}})}{n_i}$$

where $\mathbb{P}_{h_{\mathbb{Q}_i}}$ is the mathematical expectation of $\hat{\mathbb{P}}_{h_{\mathbb{Q}_i}}$. If $m = 1$, the previous equation corresponds to the CMC estimator variance. SS technique for probability estimation is described in Algorithm 14. A possible allocation of the samples n_i with respect to $N = \sum_{i=1}^{m} n_i$ is to set $n_i = d_i N$. It can be shown in this case that the variance of $\hat{\mathbb{P}}^{\text{SS}}$ is then always lower than the corresponding variance of $\hat{\mathbb{P}}^{\text{CMC}}$. Nevertheless, this proportional allocation of samples is not optimal because the ideal allocation $\{n_1, \ldots, n_m\}$ is the one that minimizes the variance $\sigma^2_{\hat{\mathbb{P}}^{\text{SS}}}$ with the constraint $N = \sum_{i=1}^{m} n_i$. The optimal allocation can be found with a Lagrange multiplier method such that

$$n_i = n \frac{d_i \mathbb{P}_{h_{\mathbb{Q}_i}}(1 - \mathbb{P}_{h_{\mathbb{Q}_i}})}{\sum_{i=1}^{m} \mathbb{P}_{h_{\mathbb{Q}_i}}(1 - \mathbb{P}_{h_{\mathbb{Q}_i}})}$$

In a similar way, it is also possible to estimate some quantiles of $\phi(\mathbf{X})$ with SS. In that case, given a level of probability α, we estimate the quantile with

ALGORITHM 14 Stratified sampling for probability estimation

Input: The pdf f, the number of samples N, the function $\phi(\cdot)$, the threshold T, the partition $\mathbb{Q}_i$, $i = 1, \ldots, m$, of $\mathbb{R}^d$, and the sample numbers n_i, $i = 1, \ldots, m$, with the constraint $N = \sum_{i=1}^{m} n_i$.

Output: The probability estimate $\hat{\mathbb{P}}^{\mathrm{SS}}$.

1 Define the density $h_{\mathbb{Q}_i}(\mathbf{x}) = \mathbf{1}_{\mathbf{x}\in\mathbb{Q}_i} \frac{f(\mathbf{x})}{d_i}$ with $d_i = \int_{\mathbb{Q}_i} f(\mathbf{x})\, d\mathbf{x}$.

2 **for** $i \leftarrow 1$ ***to*** m **do**

3 Sample $\mathbf{X}_1^i, \ldots, \mathbf{X}_{n_i}^i$ with density $h_{\mathbb{Q}_i}$.

4 Apply ϕ to $\mathbf{X}_1^i, \ldots, \mathbf{X}_{n_i}^i$ to determine the samples $\phi(\mathbf{X}_1^i), \ldots, \phi(\mathbf{X}_{n_i}^i)$.

5 Estimate $\hat{\mathbb{P}}^{\mathrm{SS}} = \sum_{i=1}^{m} \frac{d_i}{n_i} \sum_{j=1}^{n_i} \mathbf{1}_{\phi(\mathbf{X}_j^i)>T}$.

6 **return** $\hat{\mathbb{P}}^{\mathrm{SS}}$.

$$q_\alpha^{\mathrm{SS}} = \inf_{y\in\mathbb{R}} \left\{ F_N^{\mathrm{SS}}(y) \geq \alpha \right\}$$

where $F_N^{\mathrm{SS}}(y)$ is the empirical cdf defined by

$$F_N^{\mathrm{SS}}(y) = \sum_{i=1}^{m} \frac{d_i}{n_i} \sum_{j=1}^{n_i} \mathbf{1}_{\phi(\mathbf{X}_j^i)\leq y}$$

The choice of the subsets $\mathbb{Q}_i$ and of n_i is very important in order to reduce the Monte Carlo estimator variance with SS but requires some information on the input–output function $\phi(\cdot)$. If we have no clue concerning the set Ω_f, the method of stratified sampling is not applicable to the specific context of rare event estimation and could increase the Monte Carlo relative error if $\mathbb{Q}_i$ and n_i are not adapted to $\phi(\cdot)$ and f.

7.4.2 Monte Carlo method with Latin hypercube sampling

Latin hypercube sampling (LHS) (Ayyub & Kwan-Ling, 1989; Inman, Helson, & Campbell, 1981; Keqin D. & Chuntu, 1998; MacKay, 1992; McKay, Beckman, & Conover, 1979; Zhang, Breitkopf, Knopf-Lenoir, & Zhang, 2011) can be used instead of stratified sampling when the partition $\mathbb{Q}_i$, $i = 1, \ldots, m$ is difficult to estimate. The principle is to independently stratify each of the d input dimensions $\mathbf{x} = (x^1, x^2, \ldots, x^d)^T$ into N equi-possible intervals of probability $\frac{1}{N}$. For a given dimension k, we generate one sample in each interval according to the conditional joint law of f for the dimension k and thus obtain N scalar samples. The random matching between the scalar samples in the different dimensions enables to obtain a N d-tuple $\mathbf{X}_1, \ldots, \mathbf{X}_N$ that describes an LHS. The probability with LHS is estimated in the same way as the Monte Carlo with

$$\hat{\mathbb{P}}^{\text{LHS}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\phi(\mathbf{X}_i)>T}.$$

This estimate is unbiased, and its relative deviation is always lower than CMC (Keramat & Kielbasa, 1997, 1999). Nevertheless, Latin hypercube sampling is a space-filling method and is thus particularly adapted for estimating nonsmall probabilities. For rare event probability estimation, applying LHS is not advised because the obtained variance reduction can be very low.

7.4.3 Adaptive directional sampling

7.4.3.1 Principle

An adaptive allocation of the samples in stratified sampling for rare event probability estimation has been proposed in Munoz Zuniga, Garnier, Remy, & Rocquigny (2011) with an algorithm called *adaptive directional sampling* (ADS). It estimates the SS optimal allocation in two stages when directional sampling and stratified sampling are combined in a directional stratified sampling (DSS) probability estimate.

If R^2 is a chi-squared random variable with d degrees of freedom with density f_R and $\mathbf{A}$ is a random unit vector uniformly distributed on the d-dimension unit sphere Λ^d with density $f_{\mathbf{A}}$, then the vector $\mathbf{X}$ can be expressed as $\mathbf{X} = R\mathbf{A}$ where R and $\mathbf{A}$ are independent random variables (See Section 7.3). The rare event probability $\mathbb{P}_f$ can then be written as $\mathbb{P}_f = \mathbb{E}(\xi(\mathbf{A}))$ where $\xi(\mathbf{a}) = \mathbb{P}(\phi(R\mathbf{A}) > T|\mathbf{A} = \mathbf{a})$. The idea of ADS is to stratify the support of the variable $\mathbf{A}$ that takes values on Λ^d. A natural approach for that purpose is to partition Λ^d with cones. Let us define $\mathbb{Q}_i$, $i = 1, \ldots, m$, a cone partition of Λ^d. If we apply the principle of stratified sampling, we then obtain:

$$\begin{aligned}\mathbb{P}_f &= \sum_{i=1}^{m} \mathbb{P}(\mathbf{A} \in \mathbb{Q}_i)\, \mathbb{P}\left(\phi(R\mathbf{A}) > T|\mathbf{A} = \mathbf{A}_{\mathbb{Q}_i}\right), \\ &= \sum_{i=1}^{m} \mathbb{P}(\mathbf{A} \in \mathbb{Q}_i)\, \mathbb{E}\left(\xi(\mathbf{A}_{\mathbb{Q}_i})\right)\end{aligned}$$

where the random variable $\mathbf{A}_{\mathbb{Q}_i}$ follows the law of $\mathbf{A}$ restricted to $\mathbf{A} \in \mathbb{Q}_i$. The probability $\mathbb{P}(\mathbf{A} \in \mathbb{Q}_i)$ is easily computed by definition of the partition of the unit sphere, but $\mathbb{E}\left(\xi(\mathbf{A}_{\mathbb{Q}_i})\right)$ must be estimated. We generate n_i samples of $\mathbf{A}_{\mathbb{Q}_i}$ in each subset $\mathbb{Q}_i$. The total number of samples N is thus equal to $N = \sum_{i=1}^{m} n_i$. The probability $\mathbb{P}\left(\phi(R\mathbf{A}) > T|\mathbf{A} = \mathbf{A}_{\mathbb{Q}_i}\right) = \mathbb{E}\left(\xi(\mathbf{A}_{\mathbb{Q}_i})\right)$ is estimated with CMC:

$$\hat{\mathbb{P}}\left(\phi(R\mathbf{A}) > T|\mathbf{A} = \mathbf{A}_{\mathbb{Q}_i}\right) = \hat{\xi}(\mathbf{A}_{\mathbb{Q}_i}) = \frac{1}{n_i}\sum_{j=1}^{n_i} \xi(\mathbf{A}_{j,\mathbb{Q}_i})$$

where the samples $\mathbf{A}_{j,\mathbb{Q}_i}$, $j = 1, \ldots, n_i$ follow the law of $\mathbf{A}$ restricted to $\mathbf{A} \in \mathbb{Q}_i$. Rejection sampling is a possible way to obtain samples such as $\mathbf{A}_{j,\mathbb{Q}_i}$. The DSS probability estimator is then given by

$$\hat{\mathbb{P}}^{\mathrm{DSS}} = \sum_{i=1}^{m} \mathbb{P}\left(\mathbf{A} \in \mathbb{Q}_i\right) \hat{\xi}\left(\mathbf{A}_{\mathbb{Q}_i}\right)$$

It can be shown that the optimal variance of $\hat{\mathbb{P}}^{\mathrm{DSS}}$ is obtained when the number of samples in each subset $\mathbb{Q}_i$ is given by

$$n_i = N \frac{\mathbb{P}\left(\mathbf{A} \in \mathbb{Q}_i\right) \sigma_i}{\sum_{j=1}^{m} \mathbb{P}\left(\mathbf{A} \in \mathbb{Q}_j\right) \sigma_j}$$

where $\sigma_i = \sqrt{\mathbb{V}\left(\xi\left(\mathbf{A}_{\mathbb{Q}_i}\right)\right)}$. This allocation is unknown at the beginning of the DSS algorithm. ADS thus consists of a two-stage technique based on DSS probability estimate. We first estimate the optimal sample allocation n_i, $i = 1, \ldots, m$, and then generate new samples in a second stage following this estimated optimal allocation. We assume that $\gamma_1(N)N$ and $\gamma_2(N)N$ samples are available for each ADS stage with $\gamma_1(N) + \gamma_2(N) = 1$. The $\gamma_1(N)N$ first samples enable to estimate an optimal allocation:

$$\hat{n}_i = \gamma_1(N) N \frac{\mathbb{P}\left(\mathbf{A} \in \mathbb{Q}_i\right) \hat{\sigma}_i}{\sum_{j=1}^{m} \mathbb{P}\left(\mathbf{A} \in \mathbb{Q}_j\right) \hat{\sigma}_j}$$

where $\hat{\sigma}_i$ is an estimator of σ_i. Then, in the second stage, this allocation is used to estimate the rare event probability in the following way:

$$\hat{\mathbb{P}}^{\mathrm{ADS}} = \sum_{i=1}^{m} \mathbb{P}\left(\mathbf{A} \in \mathbb{Q}_i\right) \frac{1}{N_i} \sum_{j=1}^{N_i} \xi\left(\mathbf{A}_{j,\mathbb{Q}_i}\right)$$

where $N_i = \lfloor \frac{\hat{n}_i}{\gamma_1(N)} \gamma_2(N) \rfloor$ and $\lfloor . \rfloor$ is the floor function. The ADS probability estimate $\hat{\mathbb{P}}^{\mathrm{ADS}}$ is proved to be consistent and unbiased. An expression of its variance can be derived but must be estimated through retrials of the algorithm. The different stages of ADS are described in Algorithm 15.

7.4.3.2 Application to toy cases

ADS is applied to three different test cases with the following tuning $\gamma_1(N) = \gamma_2(N) = 0.5$ that it seems to provide the best results according to Munoz Zuniga et al. (2011). The number m of cones in the stratification of variable $\mathbf{A}$ is set to 2^d; we recall that d is the input dimension.

ALGORITHM 15 Adaptive directional sampling for probability estimation

Input: The pdf f, the number of samples N, the sample fraction for each ADS stage $\gamma_1(N)$ and $\gamma_2(N)$ with $\gamma_1(N) + \gamma_2(N) = 1$, the function $\phi(\cdot)$, the threshold T, and the partition $\mathbb{Q}_i$, $i = 1, \ldots, m$, of $\mathbb{R}^d$.

Output: The probability estimate $\hat{\mathbb{P}}^{\text{ADS}}$.

for $i \leftarrow 1$ ***to*** m **do**

- Generate $n_i = \lfloor \frac{\gamma_1(N)N}{m} \rfloor$ iid samples $\mathbf{A}_{1,\mathbb{Q}_i}, \ldots, \mathbf{A}_{n_i,\mathbb{Q}_i}$ of variable $\mathbf{A}_{\mathbb{Q}_i}$ with rejection sampling.
- Determine $r_1, \ldots, r_{n_i}$ such that $\phi(r_j \mathbf{A}_j) = T$ for $j = 1, \ldots, n_i$ by optimization.
- Estimate $\xi(\mathbf{A}_{j,\mathbb{Q}_i}) = \left(1 - F_{R^2}(r_j^2)\right)$ for $j = 1, \ldots, n_i$.
- Estimate the mean $\hat{\xi}(\mathbf{A}_{\mathbb{Q}_i})$ and the variance $\hat{\sigma}_i^2$ of the samples $\xi(\mathbf{A}_{j,\mathbb{Q}_i})$ for $j = 1, \ldots, n_i$

Evaluate the optimal allocation

$$\hat{n}_i = \gamma_1(N) N \frac{\mathbb{P}(\mathbf{A} \in \mathbb{Q}_i)\,\hat{\sigma}_i}{\sum_{j=1}^{m} \mathbb{P}\left(\mathbf{A} \in \mathbb{Q}_j\right)\hat{\sigma}_j}$$

for $i = 1, \ldots, m$.

for $i \leftarrow 1$ ***to*** m **do**

- Generate $N_i = \lfloor \frac{\hat{n}_i}{\gamma_1(N)} \gamma_2(N) \rfloor$ iid samples $\mathbf{A}_{1,\mathbb{Q}_i}, \ldots, \mathbf{A}_{N_i,\mathbb{Q}_i}$ of variable $\mathbf{A}_{\mathbb{Q}_i}$ with rejection sampling.
- Determine $r_1, \ldots, r_{N_i}$ such that $\phi(r_j \mathbf{A}_j) = T$ for $j = 1, \ldots, N_i$ by optimization.
- Estimate $\hat{\xi}(\mathbf{A}_{\mathbb{Q}_i}) = \frac{1}{N_i} \sum_{j=1}^{N_i} \left(1 - F_{R^2}(r_j^2)\right)$

Evaluate the probability estimate $\hat{\mathbb{P}}^{\text{ADS}}$ with

$$\hat{\mathbb{P}}^{\text{ADS}} = \sum_{i=1}^{m} \mathbb{P}(\mathbf{A} \in \mathbb{Q}_i)\,\hat{\xi}(\mathbf{A}_{\mathbb{Q}_i})$$

return $\hat{\mathbb{P}}^{\text{ADS}}$.

Polynomial square root function

ADS is applied to the polynomial square root function. Several values of simulation budget N are tested. See Table 7.10 for all the results obtained with this method. A set of samples obtained with ADS is provided in Figure 7.5. ADS enables to reach a very low relative error with a reasonable simulation budget. ADS probability estimation results are thus promising.

Four-branch system

ADS is applied to the four-branch system. Results are given in Table 7.11 with a threshold equal to 12. The ADS algorithm is able to cope with multimodal failure

Table 7.10 Results obtained with ADS for the polynomial square root test case

$\hat{\mathbb{P}}^{\text{ADS}}$	$\text{RB}(\hat{\mathbb{P}}^{\text{ADS}})$	$\text{RE}(\hat{\mathbb{P}}^{\text{ADS}})$	**Simulation budget**	ν^{ADS}
2.2×10^{-6}	-7%	20%	200	5.4×10^4
2.3×10^{-6}	-1%	8%	2000	3.4×10^4
2.3×10^{-6}	-1%	2%	20,000	4.1×10^4

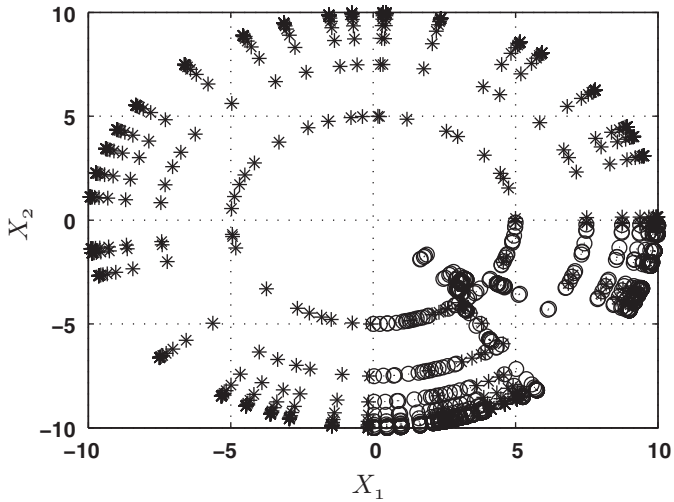

Figure 7.5 Set of samples obtained with ADS for polynomial square root function ("*" corresponds to samples of the first ADS stage and "o" to samples of the second ADS stage).

Table 7.11 Results obtained with ADS for the four-branch system test case

$\hat{\mathbb{P}}^{\text{ADS}}$	$\text{RB}(\hat{\mathbb{P}}^{\text{ADS}})$	$\text{RE}(\hat{\mathbb{P}}^{\text{ADS}})$	**Simulation budget**	ν^{ADS}
1.2×10^{-6}	3%	14%	200	2.2×10^5
1.2×10^{-6}	3%	3%	2000	4.0×10^5
1.2×10^{-6}	2%	1%	20,000	2.8×10^5

regions. Moreover, even if there is no principal direction of failure in the input space, ADS still performs probability estimation with high efficiency. It is then nearly equivalent to a DS probability estimation because the stratification is not useful in that case.

Polynomial product

ADS is also applied to the polynomial product function with different dimensions d and thresholds T in order to test its ability to estimate probabilities in case of high-dimensional problems. The results obtained for this toy case are summarized in Table 7.12. The curse of dimensionality strongly restrains the applicability of ADS. Indeed, the number $m = 2^d$ of different partition regions becomes too important when the dimension d increases, and then ADS is not applicable.

Table 7.12 Results of ADS for the polynomial product toy case (for different dimensions and thresholds)

d	$\hat{\mathbb{P}}^{\text{ADS}}$	RB($\hat{\mathbb{P}}^{\text{ADS}}$)	RE($\hat{\mathbb{P}}^{\text{ADS}}$)	**Simulation budget**	ν^{ADS}
5	8.4×10^{-7}	0.4%	3.4%	24,600	4.1×10^4
20	Not affordable	n/a	n/a	n/a	n/a
50	Not affordable	n/a	n/a	n/a	n/a
200	Not affordable	n/a	n/a	n/a	n/a

7.4.4 Conclusion

ADS is a very reliable method for estimating rare event probabilities by combining stratified sampling and directional sampling. With a low simulation budget, ADS reaches a high level of probability accuracy. Nevertheless, this algorithm is not able to cope with high-dimensional systems because applying ADS when $d > 10$ (as an order of magnitude) becomes difficult. When there is no preferential direction, ADS probability estimation is equivalent to a DS probability estimation. The main characteristics of ADS are given in Table 7.13. The application of ADS to the use cases proposed in this book is of interest if their dimension is not too important.

Table 7.13 Summary of ADS for rare event probability estimation

Criteria	**ADS characteristics**
Rarity of $\mathbb{P}_f$	Low influence on the simulation budget
Simulation budget	From 100 to 10^4 samples in most cases
Type of function $\phi(\cdot)$	Regularity and differentiability assumptions
Multimodality of f restricted to Ω_f	Well adapted, resulting from the use of stratification
Dimension of $\mathbf{X}$	Not applicable when $d > 10$
Probability estimate error	Estimation with retrials
Difficulty of implementation	Medium to hard; tuning of the cone stratification

7.5 Geometrical methods

Geometrical methods consist in approximating the failure region Ω_f with some different shapes such as polyhedron (Maire, 1999) or ellipsoid. The failure domain can also be evaluated with radial exploration (Tahim & Spence, 1980) or exploration by orthogonal search (Ogrodzki & Styblinski, 1980). These techniques can be very efficient with a low number of samples, but some knowledge of $\mathbf{x}$ and Ω_f is required to be applicable. These algorithms are notably very sensitive to their initialization. Moreover, as with FORM/SORM, there is no control of the error in these techniques; thus, we never really know whether the failure region has been well approximated. Geometrical methods for estimating rare event probabilities are used in practice in only very specific cases.

References

Ayyub, B., and Kwan-Ling, L. (1989). Structural reliability assessment using Latin hypercube sampling. In *Proceedings of the International Conference on Structural Safety and Reliability, ICOSSAR'89* (p. 174–184). San Francisco, USA.

Bjerager, P. (1988). Probability integration by directional simulation. *Journal of Engineering Mechanics*, *114*(8), 1288–1302.

Bjerager, R. (1991). Methods for structural reliability computation. In C. F. (Ed.), *Reliability problems: General principles and applications in mechanics of solid and structures* (p. 89–136). New York, USA: Springer Verlag.

Breitung, K. (1984). Asymptotic approximation for multinormal integrals. *Journal of Engineering Mechanics*, *110*(3), 357–366.

Cancela, H., and El Khadiri, M. (1995). A recursive variance-reduction algorithm for estimating communication-network reliability. *Reliability, IEEE Transactions on*, *44*(4), 595–602.

Cancela, H., and El Khadiri, M. (2003). The recursive variance-reduction simulation algorithm for network reliability evaluation. *Reliability, IEEE Transactions on*, *52*(2), 207–212.

Cochran, W. G. (1977). *Sampling techniques*. New York, USA: John Wiley & Sons.

Daniels, H. E. (1954). Saddlepoint approximations in statistics. *The Annals of Mathematical Statistics*, *25*(4), 631–650.

de Rocquigny, E., Devictor, N., and Tarantola, S. (2008). *Uncertainty in industrial practice: A guide to quantitative uncertainty management*. Chichester, UK: John Wiley & Sons.

Der Kiureghian, A., and Dakessian, T. (1998). Multiple design points in first and second-order reliability. *Structural Safety*, *20*(1), 37–49.

Dietlevsen, O., and Madsen, H. (1996). *Structural reliability methods*. New York, USA: John Wiley & Sons.

Du, X., and Sudjianto, A. (2004). First order saddlepoint approximation for reliability analysis. *AIAA Journal*, *42*(6), 1199–1207.

Goutis, C., and Casella, G. (1999). Explaining the saddlepoint approximation. *The American Statistician*, *53*(3), 216–224.

Hasofer, A., and Lind, N. (1974). An exact and invariant first-order reliability format. *Journal of Engineering Mechanics*, *100*, 111–121.

Huang, B., and Du, X. (2006). Uncertainty analysis by dimension reduction integration and saddlepoint approximations. *Transactions of the American Society of Mechanical Engineers*, *128*, 26–33.

Huanh, B., and Du, X. (2008). Probabilistic uncertainty analysis by mean-value first order saddlepoint approximation. *Reliability Engineering and System Safety*, *93*(2), 325–336.

Huzurbazar, S. (1999). Practical saddlepoint approximations. *The American Statistician*, *53*(3), 225–232.

Inman, R. L., Helson, J. C., and Campbell, J. E. (1981). An approach to sensitivity analysis of computer models: Part I—Introduction, input variable selection and preliminary variable assessment. *Journal of Quality Technology*, *13*(3), 174–183.

Jensen, J. L. (1995). *Saddlepoint approximations*. Oxford, UK: Oxford University Press.

Karp, R. M., and Luby, M. G. (1983). *A new Monte Carlo method for estimating the failure probability of an n-component system* (Tech. Rep. No. UCB/CSD-83-117). EECS Department, University of California, Berkeley.

Keqin D., Z., Zegong, and Chuntu, L. (1998). Latin hypercube sampling used in the calculation of the fracture probability. *Reliability Engineering and System Safety*, *59*(2), 239–242.

Keramat, M., and Kielbasa, R. (1997). Worst case efficiency of Latin hypercube sampling Monte-Carlo (LHSMC) yield estimator of electrical circuits. In *Proceedings of the IEEE international symposium on circuits and systems* (p. 1660–1663). Hong Kong.

Keramat, M., and Kielbasa, R. (1998). A study of stratified sampling in variance reduction techniques for parametric yield estimation. *Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on*, *45*(5), 575–583.

Keramat, M., and Kielbasa, R. (1999). Modified Latin hypercube sampling Monte-Carlo (MLHSMC) estimation for average quality index. *Analog Integrated Circuits and Signal Processing*, *19*(1), 87–98.

Koutsourelakis, P. (2004). Reliability of structures in high dimensions. Part II. Theoretical validation. *Probabilistic Engineering Mechanics*, *19*(4), 419–423.

Koutsourelakis, P. S., Pradlwarter, H. J., and Schueller, G. I. (2004). Reliability of structures in high dimensions. Part I. Algorithms and application. *Probabilistic Engineering Mechanics*, *19*, 409–417.

Kumamoto, H., Tanaka, T., and Inoue, K. (1987). A new Monte Carlo method for evaluating system-failure probability. *Reliability, IEEE Transactions on*, *R-36*(1), 63–69.

Lassen, T., and Recho, N. (2006). *Fatigue life analyses of welded structures*. New York, USA: ISTE Wiley.

MacKay, M. D. (1992). Latin hypercube sampling as a tool in uncertainty analysis of computer models. In *Proceedings of the 24th conference on winter simulation* (p. 557–564). New York.

Madsen, H., Krenk, S., and Lind, N. C. (1986). *Methods of structural safety*. Englewood Cliffs, USA: Prentice-Hall.

Maire, F. (1999). Rule-extraction by backpropagation of polyhedra. *Neural Networks*, *12*(4–5), 717–725.

McKay, M. D., Beckman, R. J., and Conover, W. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, *21*, 239–245.

Munoz Zuniga, M., Garnier, J., Remy, E., and Rocquigny, E. de. (2011). Adaptive directional stratification for controlled estimation of the probability of a rare event. *Reliability Engineering and System Safety*, *96*(12), 1691–1712.

Ogrodzki, J. T., and Styblinski, M. A. (1980). Optimal tolerancing, centering and yield optimization by One-Dimensional Orthogonal Search (ODOS) technique. In *Proceedings Eur. Conf. Circ. Theory and Design (ECCTD 80)* (p. 480–485). Warsaw, Poland.

Pei-Ling, L., and Kiureghian, A. D. (1991). Optimization algorithms for structural reliability. *Structural Safety*, *9*(3), 161–177.

Pradlwarter, H., Pellissetti, M., Schenk, C., Schuëller, G., Kreis, A., Fransen, S., et al. (2005). Realistic and efficient reliability estimation for aerospace structures. *Computer Methods in Applied Mechanics and Engineering*, *194*(12–16), 1597–1617.

Rackwitz, R., and Flessler, B. (1978). Structural reliability under combined random load sequences. *Computers and Structures*, *9*(5), 489–494.

Schueller, G. I., Pradlwarter, H. J., and Koutsourelakis, P. S. (2004). A critical appraisal of reliability estimation procedures for high dimensions. *Probabilistic Engineering Mechanics*, *19*, 463–474.

Sudret, B. (2012). Meta-models for structural reliability and uncertainty quantification. In S. Q. K. Phoon M. Beer and S. Pang (Eds.), *Asian-pacific symposium on structural reliability and its applications* (p. 53–76). Singapore, Singapore.

Tahim, K., and Spence, R. (1980). A radial exploration algorithm for the statistical analysis of linear circuits. *Circuits and Systems, IEEE Transactions on*, *27*(5), 421–425.

Yan-Gang, Z., and Tetsuro, O. (1999). A general procedure for first/second-order reliability method (FORM/SORM). *Structural Safety*, *21*(2), 95–112.

Zhang, P., Breitkopf, P., Knopf-Lenoir, C., and Zhang, W. (2011). Diffuse response surface model based on moving Latin hypercube patterns for reliability-based design optimization of ultrahigh strength steel NC milling parameters. *Structural and Multidisciplinary Optimization*, *44*(5), 613–628.

Methods for high-dimensional and computationally intensive models

8

M. Balesdent, L. Brevault, S. Lacaze, S. Missoum, J. Morio

Complex simulation codes such as the ones used in aerospace are often computationally expensive and involve a large number of variables. These features significantly hamper the estimation of rare event probabilities. To reduce the computational burden, an analysis of the most important variables of the problem can be performed before applying rare event estimation methods. Another way to reduce this burden is to build a surrogate model of the computationally costly simulation code and to perform the probability estimation on this metamodel. In this chapter, we first review the main techniques used in sensitivity analysis and then describe several surrogate models that are efficient in the probability estimation context.

8.1 Sensitivity analysis

Sensitivity analysis (SA) is the study of how to apportion the variation in the model output, qualitatively or quantitatively to variations in the model inputs (Saltelli, Tarantola, & Campolongo, 2000). Two types of SA can be distinguished: local (around a baseline) and global (over the entire input variable domain of variation). The use of global SA methods allows to characterize the influence of the input variables over the entire input space and to filter out the variables with negligible effects on the output in order to decrease the computational burden and enhance the efficiency of the elected rare event methods. The screening of the most important effects helps the decision maker to determine which variables should be considered.

Different categories of SA methods exist: the variance decomposition methods (Sobol, 1990), ANalysis Of VAriance (ANOVA) (Lamboni, Monod, & Makowski, 2011), differential analysis (Morris method (Morris, 1991)), derivative-based SA (Sobol & Kucherenko, 2009), and linear relationship measures (correlation coefficients (CC), partial correlation coefficients (PCC), standardized regression coefficients (SRC) (Iooss, 2011)). In this chapter, we briefly describe the main methods. A complete review of SA methods can be found in Helton, Johnson, Sallaberry, & Storlie (2006) and Iooss & Lemaître (2015).

The computer-based model that is analyzed is represented as a black box function ϕ with d inputs $\mathbf{X} = (X_1, X_2, \ldots, X_d)^T$ and an output vector $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_m)^T$ related by

$$\mathbf{Y} = \phi(\mathbf{X}) = \phi(X_1, X_2, \ldots, X_d) \tag{8.1}$$

Estimation of Rare Event Probabilities in Complex Aerospace and Other Systems
http://dx.doi.org/10.1016/B978-0-08-100091-5.00008-3

Here we consider a scalar output Y for the sake of clarity, but all the derivations can be generalized for an output vector.

In the first part of this section, we highlight the methods involving importance measures. We start with the most general one: the Sobol method. Then we detail SA methods that can be applied under specific assumptions to the model: ANOVA by design of experiment (DoE), SRC, and PCC. In the second part of this section, we detail screening SA methods (one variable at a time (OAT), Morris method (Morris, 1991)).

8.1.1 Importance measure-based methods

8.1.1.1 Decomposition of the variance

The variance decomposition methods for SA consist of a decomposition of the variance of the output into a sum of contributions resulting from the inputs and their interactions (ANOVA (Lamboni et al., 2011)). Two types of ANOVA can be distinguished: the functional ANOVA based on Sobol approach and the ANOVA by DoE (Lamboni et al., 2011). The functional ANOVA method does not require any hypothesis on the form of the model except that the inputs are independent and $\mathbb{E}(\phi^2(\mathbf{X})) < \infty$. Sobol (1990) has demonstrated the following unique functional decomposition for the function ϕ:

$$\phi(\mathbf{X}) = \phi_0 + \sum_{j=1}^{d} \phi_j(X_j) + \sum_{i<j}^{d} \phi_{ij}(X_i, X_j) + \cdots + \phi_{1\ldots d}(X_1, \ldots, X_d) \tag{8.2}$$

where $\phi_0 = \mathbb{E}(\phi(\mathbf{X})) = \int_\Omega \phi(\mathbf{x}) f(\mathbf{x})\, \mathrm{d}\mathbf{x}$. Ω is the d-dimensional cube $[0,1]^d$ where the inputs are defined and assumed to have a uniform distribution $U_{[0,1]}$ represented by the pdf f. Moreover,

$$\begin{aligned} \phi_j(X_j) &= \mathbb{E}\left(\phi(\mathbf{X})|X_j\right) - \phi_0, \\ \phi_{ij}(X_i, X_j) &= \mathbb{E}(\phi(\mathbf{X})|X_i, X_j) - \mathbb{E}(\phi(\mathbf{X})|X_i) \\ &\quad - \mathbb{E}(\phi(\mathbf{X})|X_j) + \phi_0 \end{aligned}$$

and $\phi_{1\ldots d}(X_1, \ldots, X_d)$ is defined as the difference between $\phi(\mathbf{X})$ and the sum of all the increasing dimension functions such that Equation (8.2) is verified.

Furthermore, each function of the decomposition verifies $\forall l \in \{1, \ldots, s\}$; $\forall \{j_1, \ldots, j_s\} \subseteq \{1, \ldots, d\}$ (Sobol, 1990)

$$\int_\Omega \phi_{j_1,\ldots,j_s}(x_{j_1}, \ldots, x_{j_s})\, \mathrm{d}x_{j_l} = 0 \tag{8.3}$$

The orthogonality of the Sobol decomposition functions can be proved from Equation (8.2) (Sobol, 1990).

Sobol indices

Based on the functional decomposition in Equation (8.2), Sobol introduces the Sobol indices (Sobol, 2001) to quantify the partition of the output variance (Sobol, 1990). With the decomposition of the function $\phi(\cdot)$ into the sum of functions of increasing dimensions (as explained in the previous paragraph) and by using the decomposition of the variance (Sobol, 1990), it follows that

$$\mathbb{V}(Y) = \sum_{j=1}^{d} \mathbb{V}_j(Y) + \sum_{i<j}^{d} \mathbb{V}_{ij}(Y) + \cdots + \mathbb{V}_{123\ldots d}(Y) \tag{8.4}$$

with $\mathbb{V}$ the variance and $\mathbb{V}_j(Y)$ as defined in Equations (8.7) and (8.8). The variability of the output Y from all the input variables except X_j is analyzed by fixing the input variable X_j at a value x_j:

$$\mathbb{V}(Y|X_j = x_j) = \mathbb{E}(Y^2|X_j = x_j) - \mathbb{E}(Y|X_j = x_j)^2 \tag{8.5}$$

To account for all the possible x_j, the expectation of the conditional variance is considered.

Given the total variance

$$\mathbb{V}(Y) = \mathbb{V}[\mathbb{E}(Y|X_j)] + \mathbb{E}[\mathbb{V}(Y|X_j)] \tag{8.6}$$

the value $\mathbb{V}[\mathbb{E}(Y|X_j)]$ can be used for SA. It increases as the importance of the variable X_j with respect to the variance of Y increases. To have a normalized quantity, the first-order Sobol index S_j for the input variable X_j and second-order Sobol index S_{ij} for the interaction between X_i and X_j are defined by

$$S_j = \frac{\mathbb{V}[\mathbb{E}(Y|X_j)]}{\mathbb{V}(Y)} = \frac{\mathbb{V}_j}{\mathbb{V}(Y)} \tag{8.7}$$

$$S_{ij} = \frac{\mathbb{V}[\mathbb{E}(Y|X_i, X_j)] - \mathbb{V}_i - \mathbb{V}_j}{V(Y)} = \frac{\mathbb{V}_{ij}}{\mathbb{V}(Y)} \tag{8.8}$$

The first-order Sobol index quantifies the part of variance of Y from X_j, referred as the *main effect*. The second-order Sobol indices allow to measure the importance of the interaction between two input variables X_i and X_j. The same principle can be used to derive the Sobol indices of order 3, 4, and so on. The total Sobol indices ST_j are the sum of all the Sobol indices relative to X_j

$$\mathrm{ST}_j = \sum_{j \sharp i} S_i \tag{8.9}$$

where $j\sharp i$ stands for all the $S_{i1,\ldots,id}$ terms that include the index j. For instance, ST_1 includes $S_1, S_{12}, \ldots, S_{1k}$, $S_{123}, \ldots, S_{123\ldots d}$. Total Sobol indices measure the part of output variance explained by all the effects in which the input variable j plays a part (the first-order and all the higher orders).

For black box functions, Sobol indices cannot be analytically derived and must be numerically estimated. Several methods can be employed. The crude Monte Carlo method is traditionally used to estimate Sobol indices. Other sampling schemes such as proposed in Jansen (1999) or Fourier amplitude sensitivity test (FAST) (Saltelli, Tarantola, & Chan, 1999) can be performed. However, Sobol index calculations are computationally expensive and require a high number of calls to the studied function. The Sobol method is applicable to all the cases for which variances are finite (linear or nonlinear, monotonic or nonmonotonic functions).

Remark. *Note that methods have recently been proposed for dependent inputs (Caniou, 2012; Chastaing, Gamboa, & Prieur, 2014, 2012; Saltelli et al., 2010) but are not detailed in this book for the sake of conciseness.*

Another approach for SA based on the decomposition of the variance can be used. Instead of considering the functional ANOVA (Equation 8.2), the conditional variances are calculated with a DoE.

ANOVA by design of experiment

ANOVA by DoE differs from the functional ANOVA in that it is based on a discretization of the continuous input variables. Equation (8.4) is discretized according to a chosen DoE into several levels for each input variable. By choosing appropriate levels and DoE, the conditional variances can be approximated.

Each observation from the DoE can be modeled as

$$\begin{aligned} y_{X_{1_{h_1}},X_{2_{h_2}},\ldots,X_{d_{h_d}}} = {} & \eta + \eta_{X_{1_{h_1}}} + \eta_{X_{2_{h_2}}} + \cdots + \eta_{X_{d_{h_d}}} \\ & + \eta_{X_{1_{h_1}}X_{2_{h_2}}} + \cdots + \eta_{X_{1_{h_1}}X_{2_{h_2}}\ldots X_{d_{h_d}}} \end{aligned}$$

with η: the average effect, $\eta_{X_{1_{h_1}}}$: the effect due to the level h_1 of the variable X_1, and $\eta_{X_{1_{h_1}}X_{2_{h_2}}}$: the effect due to the interactions between the level h_1 of the variable X_1 and the level h_2 of the variable X_2. Therefore, if the DoE requires N function calls,

$$\eta = \frac{1}{N}\sum_{j=1}^{d}\sum_{\text{all levels } h_j} y_{X_{1_{h_1}},X_{2_{h_2}},\ldots,X_{d_{h_d}}} = \mathbb{E}\left(y_{X_{1_{h_1}},X_{2_{h_2}},\ldots,X_{d_{h_d}}}\right) \tag{8.10}$$

and for instance

$$\eta_{X_{1_{h_1}}} = \mathbb{E}\left[y_{X_{1_{h_1}},X_{2_{h_2}},\ldots,X_{d_{h_d}}} \middle| X_1 = h_1\right] - \eta \tag{8.11}$$

with the expectation representing the mean of all the observations in the DoE when $X_1 = h_1$. ANOVA by DoE (Lamboni et al., 2011) allows to write the sum of square decomposition:

$$\begin{aligned}\mathrm{SS}\left(\tilde{\mathbf{Y}}\right) = &\mathrm{SS}_{X_1} + \mathrm{SS}_{X_2} + \cdots + \mathrm{SS}_{X_d} \\ &+ \mathrm{SS}_{X_1X_2} + \cdots + \mathrm{SS}_{X_1X_2\ldots X_d}\end{aligned} \tag{8.12}$$

where $\tilde{\mathbf{Y}}$ is the vector of all the observations given by the DoE, SS_{X_j} the quadratic sum characterizing the main effect of the input variable X_j, $\mathrm{SS}_{X_jX_r}$ the quadratic sum characterizing the interaction effect between the input variables X_j and X_r. Equation (8.12) is a discrete version of Equation (8.4).

$\mathrm{SS}(\tilde{\mathbf{Y}})$ measures the total variability in the model output

$$\begin{aligned}\mathrm{SS}\left(\tilde{\mathbf{Y}}\right) &= \sum_{j=1}^{d} \sum_{\text{all levels } h_j} \left(y_{X_{1_{h_1}}, X_{2_{h_2}}, \ldots, X_{d_{h_d}}} - \eta\right)^2, \\ &= N\mathbb{E}\left(y^2_{X_{1_{h_1}}, X_{2_{h_2}}, \ldots, X_{d_{h_d}}}\right) - 2N\eta\mathbb{E}\left(y_{X_{1_{h_1}}, X_{2_{h_2}}, \ldots, X_{d_{h_d}}}\right) \\ &\quad + N\mathbb{E}\left(y_{X_{1_{h_1}}, X_{2_{h_2}}, \ldots, X_{d_{h_d}}}\right)^2, \\ &= N\mathbb{V}(\tilde{\mathbf{Y}})\end{aligned}$$

Moreover, each quadratic sum SS_{X_j} corresponds to the mean over the levels of input variable X_j of the conditional variance of the output $\tilde{\mathbf{Y}}$ evaluated at $X_j = h_j$:

$$\begin{aligned}\mathrm{SS}_{X_j} &= n_j \sum_{\text{levels of } X_j} \left(\mathbb{E}\left[y_{X_{1_{h_1}}, X_{2_{h_2}}, \ldots, X_{d_{h_d}}} \,\middle|\, X_j = h_j\right] - \eta\right)^2, \\ &= N\mathbb{V}(\mathbb{E}(\tilde{\mathbf{Y}}|X_j))\end{aligned}$$

with n_j the level discretization number of the input variable X_j. Based on this decomposition of variance, the sensitivity index "SI" for the input variable X_j is defined by (Lamboni et al., 2011):

$$\mathrm{SI}_{X_j} = \frac{\mathrm{SS}_{X_j}}{\mathrm{SS}(\tilde{\mathbf{Y}})}$$

DoEs such as fractional factorial, Latin square, Latin hypercube sampling, or full factorial designs lead to a decrease of the SA computational cost compared to the Sobol approach (Monod, Naud, & Makowski, 2006). However, the choices of the DoE and the different levels are crucial to accurately approximate the variances.

Other SA methods exist under the assumptions of a linear model such as SRC and PCC and are described in subsequent sections.

8.1.1.2 Standardized regression coefficients

When the function $\phi(\cdot)$ may be approximated as a linear function, sensitivity measures for the model may be computed through linear regression, using standardised

regression coefficients (SRC). A linear model for the dependency of the outputs with respect to the d input variables with N samples ($i = 1, \ldots, N$) is considered:

$$Y_i = a_0 + \sum_{j=1}^{d} a_j X_{ij} + \epsilon_i$$

with a_i the regression coefficients and ϵ_i the errors from the approximation (assumed $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$). Least squares fit between predicted and observed output data is typically used for the determination of the linear model. If we denote Y_i as the observed values, $\overline{Y}$ as the mean, and $\hat{Y}_i$ as the fitted values, then the coefficient of determination is defined by $R^2 = \frac{\sum_{i=1}^{N}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}$. This coefficient allows to determine how well the linear model fits the data. The regression coefficients a_j measure the linear relationship between the inputs and the output. To allow the nondimensionalization of the regression coefficients, it is possible to normalize the coefficients (Ekström & Broed, 2006)

$$\frac{\hat{Y}_i - \overline{Y}}{\hat{t}} = \sum_{j=1}^{d} \frac{a_j \hat{t}_j}{\hat{t}} \frac{X_{ij} - \overline{X}_j}{\hat{t}_j}$$

with $\overline{Y}$ the mean of the output and $\overline{X}_j$ the mean of the input X_j over the N samples. Furthermore,

$$\hat{t} = \left(\sum_{i=1}^{N} \frac{(Y_i - \overline{Y})^2}{N-1} \right)^{1/2}$$

$$\hat{t}_j = \left(\sum_{i=1}^{N} \frac{(X_{ij} - \overline{X}_j)^2}{N-1} \right)^{1/2}$$

$$\mathrm{SRC}_j = \frac{a_j \hat{t}_j}{\hat{t}}$$

If the input variables are independent, SRC is a measure of the importance of the input variables on the variability of the output. Another interpretation comes from the decomposition of the variance for a linear function (by independence of input variables) as

$$\mathbb{V}(Y) = \sum_{i=1}^{d} a_i^2 \mathbb{V}(X_i)$$

with $a_i^2 \mathbb{V}(X_i)$, the part of variance resulting from the input variable X_i. Thus, it is possible to quantify the sensitivity of Y with respect to X_i compared to the part of the variance from X_i on the total variance

$$\mathrm{SRC}_j = a_j \sqrt{\frac{\mathbb{V}(X_j)}{\mathbb{V}(Y)}}$$

It quantifies the part of the variance of the output because of the variance of the variable X_i.

8.1.1.3 Correlation coefficients and partial correlation coefficients

Another SA measure in the case of linear models is given by Pearson's product moment correlation coefficients (Ekström & Broed, 2006). They measure the extent to which two variables can be assumed to have a linear dependency. In our case, we are interested in measuring the dependency between the input variables X_j and the output Y. Considering N observations of Y for different X_j, CC are defined by (Plischke et al., 2009):

$$\begin{aligned}\mathrm{CC}_j = \rho_{X_j Y} &= \frac{\mathrm{cov}(X_j, Y)}{\sqrt{\mathbb{V}(X_j)\mathbb{V}(Y)}}, \\ &= \frac{\sum_{i=1}^{N}(X_{j_i} - \overline{X}_j)(Y_i - \overline{Y})}{\left(\sum_{i=1}^{N}(X_{j_i} - \overline{X}_j)^2\right)^{1/2}\left(\sum_{i=1}^{N}(Y_i - \overline{Y})^2\right)^{1/2}}\end{aligned}$$

with $\mathrm{cov}(X_j, Y)$ the covariance between X_j and Y.

However, CC do not take into account the possible effects that other variables might have. PCC can be calculated to determine the strength of the linear relationship between the two inputs when all linear effects from the other input variables are removed (Iooss, 2011). If we note $S_j = \{X_1, X_2, \ldots, X_{j-1}, X_{j+1}, \ldots, X_d\}$, then the PCC between X_j and Y with S_j fixed is given by

$$\mathrm{PCC}_{j|S_j} = \rho_{X_j Y|S_j} = \frac{\mathrm{cov}(X_j, Y|S_j)}{\sqrt{\mathbb{V}(X_j|S_j)\mathbb{V}(Y|S_j)}}$$

When the input variables are uncorrelated, SRC is equivalent to CC. Note, however, that SRC and PCC measure different quantities. SRC measure the effect on the output of the input variables in terms of a percentage of the output standard deviation. PCC measure how linear is the relationship between one input variable and the output while removing the effect of other input variables.

The quantitative methods for SA quantify the importance of the variability of input variables and their interactions on the variability of the output. However, these methods tend to be computationally intensive. This is particularly true for Sobol method whereas ANOVA by DoE can be an exception if the number of levels chosen is small. When only few calls to the function are possible, screening methods can be employed.

8.1.2 Screening methods

Screening methods complement quantitative methods because the required number of model evaluations is low compared to other SA techniques (Ekström & Broed, 2006). For rare event estimation of complex systems that involve computationally expensive models and a large number of input variables, screening methods could identify the variables that have the strongest effects on the output variability.

8.1.2.1 One variable at a time

One variable at a time (OAT) method is based on the variation of only one variable while the others are kept fixed at a baseline value. If we consider a baseline for the model (the nominal values of the input variables), we perform the OAT analysis by varying one of the inputs in an interval (for instance, $\pm 10\%$) while the other input variables are fixed at the baseline value. The range of the output is analyzed for the set of the d-OAT computations. Another OAT technique consists in computing the partial derivatives of the model function with respect to the input variables. This method is called *local sensitivity analysis* because it depends on the choice of the point where the partial derivatives are calculated (Morris, 1991). The number of model evaluations is of the order of d (Ekström & Broed, 2006). Although these OAT methods require few model evaluations, they are local and do not provide information on the entire range of variation of the input variables.

8.1.2.2 Morris method

The Morris method (Morris, 1991) is based on the repetition of a set of randomized OAT design experiments. The Morris method overcomes the limitation of the local SA by performing partial derivative calculations in different locations of the input variable domain of variation. The method is global because the input variables can vary over their entire domain of definition. The Morris method consists of R repetitions of DoE with sequential OAT variation of the inputs (Iooss, 2011). The first point and the next direction for one experiment are chosen randomly. The Morris method can determine if the effect of the input variable X_j on the output Y is important or negligible, linear or nonlinear, with or without interactions with other input variables X_{-j}. Morris distinguishes three ways an input variable X_j can be important (Morris, 1991) depending on the nature of $\frac{\phi(X_j+\delta_j,X_{-j})-\phi(X_j,X_{-j})}{\delta_j}$, where δ_j is a variation in the input variable:

- If this term is nonnull, then X_j has an influence on the output.
- If this term is nonnull and constant across the DoE samples, X_j therefore has a linear influence on the output and has no interaction with other input variables.
- If this term varies as X_j varies, then X_j affects the nonlinearity of the output with or without interactions.

It is not possible with the Morris indice to discriminate the effects of non linearity and of the interactions with other input variables. The mean of the absolute value of the different partial derivatives is a measure of the sensitivity. The variance is a measure of both the interactions and the nonlinear effects. The main advantage of the Morris

method is the low computational cost, requiring only about one model evaluation for each elementary effect per replication. However, its drawback is that it is not possible to distinguish nonlinearity from interactions that might be essential for the decision maker (Ekström & Broed, 2006).

8.1.3 General remark about SA for rare event probability estimation

We must be careful when applying the SA techniques for reducing the dimension of the input space. Indeed, some variables might have a small impact on the entire domain of variability of $\phi(\cdot)$ but have a large impact on the probability of failure $\mathbb{P}(\phi(\mathbf{X}) > T)$ because the region in the input space relevant to the probability estimation might be very restrictive compared to the definition domain of $\mathbf{X}$. Also, the contribution of these variables to the global sensitivity of $\phi(\cdot)$ might be neglected. The development of dedicated methods to estimate the sensitivity of the rare event probability estimate to the input variables is then necessary and is still an open research issue. We can consult Morio (2011) and Garza & Millwater (2012) for more details.

8.2 Surrogate models for rare event estimation

8.2.1 Introduction

Being able to build an efficient surrogate model which allows to reduce the number of calls to the expensive input–output function $\phi(\cdot)$ while keeping a good accuracy is a key point in rare event probability estimation. A large number of methods have been proposed and compared in recent years. For the sake of conciseness, we only present several methods. For the interested reader, a survey of the different metamodel methods can be found in Sudret (2012). Classical deterministic surrogate models have been derived for the estimation of rare event probability such as polynomials (Bucher & Bourgund, 1990; Das & Zheng, 2000; Gayton, Bourinet, & Lemaire, 2003; Kim & Na, 1997) neural networks or linear approximation such as the Taylor series (Gomes & Awruch, 2004; Hurtado & Alvarez, 2000; Papadrakakis & Lagaros, 2002; Schueremans & Van Gemert, 2005). Polynomial chaos have also been associated with Monte Carlo sampling to estimate failure probabilities (Blatman & Sudret, 2008; Hu & Youn, 2011; J. Li & Xiu, 2010). In order to illustrate the joint use of rare event probability estimators and metamodels, we focus in the following sections on two specific types of surrogate models that have been widely used in such specific applications: support vector machine (SVM) and Kriging (also known as *Gaussian process*) models.

8.2.2 Support vector machines

8.2.2.1 Presentation

An SVM is a machine-learning technique (Vapnik, 2000, 1998) used in different applications such as reliability analysis (Basudhar & Missoum, 2008, 2010) or

classification, and pattern recognition (Shawe-Taylor & Cristianini, 2004; Tou & Gonzalez, 1974). An adaptation of SVM can be derived as a regression tool and is referred to as *support vector machine for regression* (Clarke, Griebsch, & Simpson, 2005). The main characteristic of SVM lies in its ability to define complex decision functions that optimally separate different classes of data samples. In rare event probability estimation (e.g., Bourinet, Deheeger, & Lemaire, 2011; Hurtado, 2004), SVM can be used as a surrogate model of the computationally expensive input–output function $\phi(\cdot)$ around the threshold T. These surrogates are particularly adapted to functions presenting discontinuities and to high-dimensional problems. The purpose of this section is to provide the reader an overview of the SVM main features. For more details, the reader is referred to Cristianini & Shawe-Taylor (2000) and Steinwart & Christmann (2008). In this section, only the classical SVM characteristics are described. Extensions of SVM such as probabilistic SVM (Basudhar & Missoum, 2013; Gao, Gunn, Harris, & Brown, 2002; Platt, 1999) or virtual SVM (Song, Choi, Lee, Zhao, & Lamb, 2013) have been proposed in the literature but are not described in this book for the sake of conciseness.

8.2.2.2 Description

Support vector machines for classification

In its basic form, SVM is a binary classifier. In a reliability context, these classes correspond to the threshold exceedance domain (referred to as the -1 class) and the domain for which the output is underneath the considered threshold (referred to as the $+1$ class). Note that SVM can also be extended to multiclass classification problems (Duan & Keerthi, 2005).

Consider a training set $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$ of p training samples in a d-dimensional space. Each sample is associated with one of the two classes characterized by a value $c_i = \pm 1$. The SVM algorithm finds the boundary (decision function) that optimally separates the training data into the two classes. We present the basic SVM theory through a detailed explanation in the case of a linearly separable data set. We then extend it to the case in which the data are not linearly separable. Notice that the SVM training set should include at least one sample in each class to be built.

Linear decision function In the SVM theory, linear decision function is modeled through the hyperplane defined as

$$\mathbf{w}^T.\mathbf{x} + b = 0$$

where $\mathbf{w}$ is the vector of hyperplane coefficients and b is the bias. It lies "half way" between two hyperplanes that separate the two classes of data. This pair of hyperplanes, referred to as *support hyperplanes*, is required to pass through at least one of the training samples of each class (support vectors) whereas no sample can be found within the margin (Figure 8.1). One of the support hyperplanes consists of the points that satisfy

$$\mathbf{w}^T.\mathbf{x} + b = +1 \tag{8.13}$$

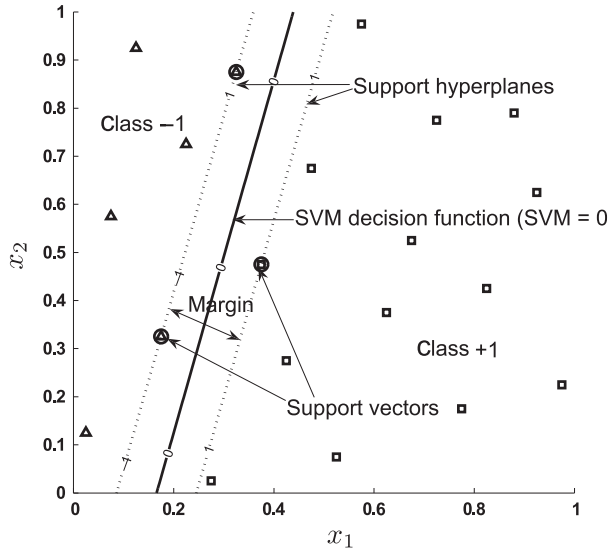

Figure 8.1 Linear decision function separating class +1 (squares) from class −1 (triangles).

The other hyperplane contains the points that follow

$$\mathbf{w}^T.\mathbf{x} + b = -1 \tag{8.14}$$

For separable data, an infinity of decision functions is possible. In order to find the "optimal" decision function, the basic idea is to maximize the "margin" that separates the support hyperplanes. All the points of the class $c = +1$ lead to a positive value of SVM and all the points in the class $c = -1$ are negative. Equations (8.13) and (8.14) and the constraint that no sample can lie between the two aforementioned hyperplanes can be combined in a single global constraint defined as follows

$$c_i(\mathbf{w}^T.\mathbf{x} + b) - 1 \geq 0 \tag{8.15}$$

The width of the SVM margin between the two support hyperplanes is $\frac{2}{\|\mathbf{w}\|}$. Therefore, determining the support hyperplanes (i.e., solving for $\mathbf{w}$ and b) results in the following optimization problem

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{subject to} \quad & c_i(\mathbf{w}^T.\mathbf{x}_i + b) - 1 \geq 0, \quad 1 \leq i \leq p \end{aligned} \tag{8.16}$$

This is a quadratic programming (QP) problem because the objective function is quadratic and the constraints are linear. Problem (8.16) is convex and can be solved

efficiently with available optimization algorithms. As a result, the optimal $\mathbf{w}$, b, and the Lagrange multipliers λ_i at the optimum are obtained. From this, the classification of any test point $\mathbf{x}$ is obtained by the sign of the following function:

$$s(\mathbf{x}, \mathcal{X}) = b + \sum_{i=1}^{p} \lambda_i c_i \mathbf{x}_i^T . \mathbf{x} \tag{8.17}$$

Note that following the Karush Kuhn Tucker conditions (Kuhn & Tucker, 1951), only the Lagrange multipliers associated with the support vectors will be strictly positive whereas the other ones will be equal to zero. In general, the number of support vectors is a small fraction of the total number of samples in $\mathcal{X}$. Equation (8.17) can be rewritten with respect to the number of support vectors "NSV":

$$s(\mathbf{x}, \mathcal{X}) = b + \sum_{i=1}^{\text{NSV}} \lambda_i c_i \mathbf{x}_i^T . \mathbf{x} \tag{8.18}$$

In this case, the data are nonseparable, so optimization problem (8.16) is relaxed by the introduction of slack variables. When the separation function is nonlinear, the approach is generalized through the introduction of a kernel as described in the following.

Nonlinear decision function SVM can be extended to the case of nonlinear decision functions by projecting the original set of variables to a higher dimension space referred to as the *feature space*. In this n-dimension feature space, the new components of a point $\mathbf{x}$ are given by $(\psi_1(\mathbf{x}), \psi_2(\mathbf{x}), \ldots, \psi_n(\mathbf{x}))$ where ψ_i are the features. The specific characteristic of SVM is that the nonlinear decision function is obtained by formulating the linear classification problem in the feature space. The classification is then obtained by the sign of

$$s(\mathbf{x}, \mathcal{X}) = b + \sum_{i=1}^{\text{NSV}} \lambda_i c_i < \boldsymbol{\psi}(\mathbf{x}_i), \boldsymbol{\psi}(\mathbf{x}) > \tag{8.19}$$

where $\boldsymbol{\psi} = (\psi_1(\mathbf{x}), \psi_2(\mathbf{x}), \ldots, \psi_n(\mathbf{x}))$ and $<,>$ is the inner product.

The inner product in Equation (8.19) forms a kernel K, so the decision function is written

$$s(\mathbf{x}, \mathcal{X}) = b + \sum_{i=1}^{\text{NSV}} \lambda_i c_i K(\mathbf{x}_i, \mathbf{x}) \tag{8.20}$$

This mechanism is referred to as the *kernel trick* in the SVM literature. The two most commonly used kernel functions are the polynomial and the Gaussian kernels. Some other kernels can be used such as multilayer perceptions, Fourier series, or

splines (Gunn, 1998). The Gaussian kernel is the most used in the literature and is defined as

$$K(\mathbf{x}_i, \mathbf{x}) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}\|^2}{\sigma^2}\right) \tag{8.21}$$

where σ is the width of the Gaussian kernel.

Support vector machines for regression

The SVM used for support vector regression (SVR) are built with the same approach as SVM but consider the response function value instead of the classes. SVR requires the use of different loss functions (Smola & Schölkopf, 2004). Several loss functions exist (quadratic, Laplace, Huber, ϵ-insensitive). The latter is particularly used in the SVR literature and is employed in this section to explain SVR mechanisms. This function can be written as follows:

$$L_\epsilon(y) = \begin{cases} 0 & \text{for } |\phi(\mathbf{x}) - y| < \epsilon \\ |\phi(\mathbf{x}) - y| - \epsilon & \text{otherwise} \end{cases}$$

with ϵ a tolerance margin. Let $\mathcal{X}$ be the current training set and $\boldsymbol{\phi}_p = [\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_p)]^T$ be the corresponding vector of responses. In case of linear model, the SVR construction results in solving the following optimization problem:

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{subject to} \quad & \mathbf{w}^T.\mathbf{x}_i + b - \phi(\mathbf{x}_i) \leq \epsilon, \quad 1 \leq i \leq p \\ & -\mathbf{w}^T.\mathbf{x}_i - b + \phi(\mathbf{x}_i) \leq \epsilon, \quad 1 \leq i \leq p \end{aligned} \tag{8.22}$$

When the optimization problem is not feasible with respect to the constraints, slack variables can be used to find a solution. For the sake of conciseness, the SVR for nonlinear model is not described in this section. The same kernel trick as the one used for SVM can be employed in order to build nonlinear SVR. For details, one can consult Basak, Pal, & Patranabis (2007). SVR has been applied with rare event techniques such as importance sampling (Dai, Zhang, Wang, & Xue, 2012) or first-order reliability methods (FORM) (Li, Lü, & Yue, 2006).

8.2.2.3 Refinement strategies

To our knowledge, the SVM model in most applications is based on a fixed training set built on the global input space (Li et al., 2006) or in specific zones such as those relevant to the most probable point (e.g., Wang, Yu, & Du, 2012). However, in some cases, adaptive refinement strategies have been proposed to account for the input space zones relevant to rare event estimation. These refinement strategies take into account

the spatial location of the samples in order to determine points to add to the existing training set. In this section, the most used techniques to refine SVM models are briefly described.

Determining the SVM training set by minimizing $(\phi(\mathbf{x}) - T)^k$

This method, proposed in Hurtado & Alvarez (2010), consists in building a series of SVM models in order to approximate the zone $\{\mathbf{x}|\phi(\mathbf{x}) = T\}$. For that purpose, the following sequence of optimization problems is solved

$$\min_{\mathbf{x}}(\phi(\mathbf{x}) - T)^k, \quad k = 2o, \quad o \in \mathbb{N} \tag{8.23}$$

where k is a parameter that takes different decreasing even values (i.e., 10 for the first sequence iteration and then 8, 6, etc.). The optimization problem is solved with a particle swarm optimization (PSO) algorithm. Once this problem is solved, the optimal population found by PSO is used as the training set to build the SVM model. The next optimization problem solving uses the previous optimal PSO population as initialization. The evolution of the k coefficient allows to iteratively improve the accuracy of the SVM model, smoothing the curvature of the optimization problem objective function. At each sequence iteration, an estimation of the rare event probability is performed. The algorithm stops when the probability estimate converges.

Adaptive refinement of the SVM

Hurtado & Alvarez (2003) proposed an adaptive refinement method that consists of the following steps:

1. A transformation of the input space to the standard space is performed using random variable transformation techniques (see Section 2.3.3.4).
2. In the standard space, N aleatory samples are generated according to the input joint distribution.
3. From these samples, l points are selected, corresponding to the ones with the highest distance to the origin of the standard space. These points are used to build an initial SVM model, and $\phi(\cdot)$ is evaluated on these points.
4. In the N-generated samples, if there exist samples between the SVM margins, the closest point to the SVM separation is added to the SVM training set and the SVM model is updated. This step is repeated until no sample is present between the two margins.

This method allows to adaptively reduce the margins around the SVM separation, reducing the uncertainty of the SVM in the relevant zones to rare event estimation. This method has been used with crude Monte Carlo (Hurtado & Alvarez, 2003) and has been extended for the use of sequential simulation techniques as described in the following section.

Subsets by support vector margin algorithm for reliability estimation (^{2}SMART)

The ^{2}SMART method, described in Deheeger & Lemaire (2007) and Bourinet et al. (2011), is dedicated to the use of adaptive importance splitting technique (AST) and consists in defining one SVM model at each adaptive threshold involved in AST. For

each intermediate threshold, an SVM model is built using a three-stage refinement approach (localization, stabilization, and convergence), which allows to accurately represent the regions corresponding to the involved thresholds. At the ith stage of AST, the main steps of ^{2}SMART are:

1. A first set of samples is generated to build an SVM model in the region corresponding to the ith level of AST, say $\hat{\phi}_{T_i}(\cdot)$, and some of these samples are used to determine the current intermediate threshold T_i, using the ρ-quantile level of AST and the SVM model in regression.
2. The SVM $\hat{\phi}_{T_i}(\cdot)$ is refined using a three-stage approach, in an iterative manner, inducing resampling (by Metropolis Hastings algorithm; see Section 2.3.3.3) and clustering the generated samples. For that purpose, three populations of samples of different size are generated and used to refine the current SVM model.
3. The last step consists in evaluating the conditional probability $\mathbb{P}(\hat{\phi}_{T_i}(\mathbf{X}) > T_i)$, corresponding to the current threshold T_i.

Adaptive refinement of SVM using max-min technique

The max-min technique, proposed in Basudhar & Missoum (2010), allows to sequentially add new samples to the current training set in order to refine the SVM model. This technique consists in solving the following optimization problem:

$$\begin{aligned} &\max_{\mathbf{x}} \min_{\mathbf{x}_i \in X} \quad \|\mathbf{x}_i - \mathbf{x}\| \\ &\text{subject to} \quad \hat{\phi}(\mathbf{x}) = T \end{aligned} \tag{8.24}$$

with $\mathbf{x}_i, i \in \{1, \ldots, p\}$ the different samples of the current training set. This method allows to generate a sample located on the approximated iso-value $\hat{\phi}(\mathbf{x}) = T$, which is at the maximal distance of the current training set samples (Figure 8.2). This method has been applied with the CMC method (Basudhar, 2011) but is applicable to any simulation technique. Moreover, this method is not dedicated to SVM and can be applied to refine other surrogate models, such as Kriging. The max-min approach does not take into account the distribution of the input variables to refine the surrogate model in high probability content regions. To overcome this issue, an improvement of the max-min technique has been proposed and is described in the next section.

Improvement of max-min technique: generalized max-min

The generalized max-min technique proposed in Lacaze & Missoum (2014) is based on the max-min strategy but accounts for the pdf of the input random variables as follows:

$$\begin{aligned} &\max_{\mathbf{x}} \min_{\mathbf{x}_i \in X} \quad \|\mathbf{x}_i - \mathbf{x}\| f(\mathbf{x})^{1/d} \\ &\text{subject to} \quad \hat{\phi}(\mathbf{x}) = T \end{aligned} \tag{8.25}$$

with $f(\cdot)$ the joint pdf of the input variables and d the dimension of the input space. The main difference between the max-min and generalized max-min approaches

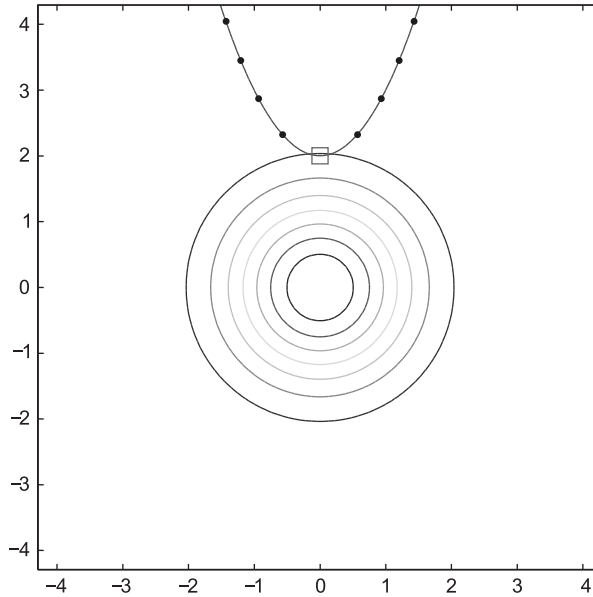

Figure 8.2 Refinement samples by max-min, input density, and curve of iso-value $\phi(\mathbf{x}) = T$.

comes from the weighting by the input variable joint pdf, which enables to refine the surrogate model in relevant regions to rare event probability estimation. Figure 8.3 illustrates the difference between the two sampling approaches. This method has been applied with CMC (Lacaze & Missoum, 2014) or subset simulation (Lacaze & Missoum, 2013). As for the max-min technique, the generalized max-min approach is not dedicated to SVM and can be used to refine other surrogate models.

8.2.3 Kriging

8.2.3.1 Presentation

The Kriging method (Matheron, 1963; Santner, Williams, & Notz, 2003; Sasena, 2002) offers some advantages in rare event probability estimation. Indeed, this surrogate model is based on a Gaussian process that enables to estimate the variance of the prediction error and consequently to define a confidence domain of the surrogate model. This indicator can be directly used to refine the model, that is, to choose new points to evaluate the exact function that allow to improve the accuracy of the model. Kriging has been extensively used with the classical Monte Carlo estimator (Echard, Gayton, & Lemaire, 2011), importance sampling method (Balesdent, Morio, & Marzat, 2013; Dubourg, Deheeger, & Sudret, 2011; Schueremans & Van Gemert, 2005), importance sampling with control variates (Cannamela, Garnier, & Iooss, 2008), or subset simulation (Bect, Ginsbourger, Li, Picheny, & Vazquez, 2012; Li, Bect, & Vazquez, 2012; Vazquez & Bect, 2009). The way to refine the Kriging

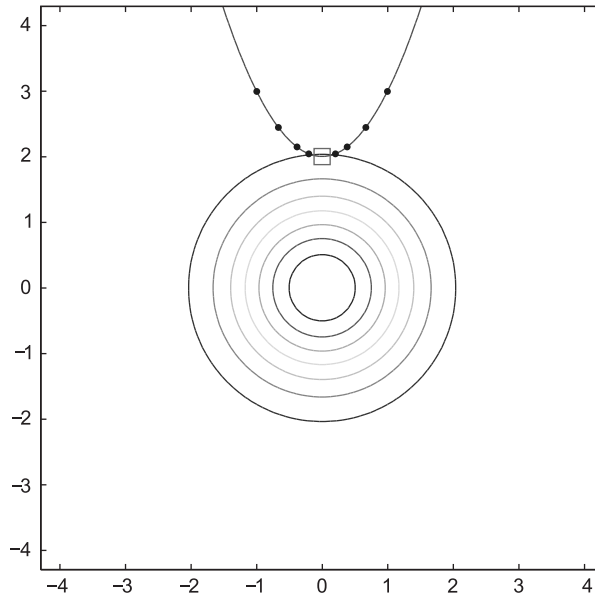

Figure 8.3 Refinement samples by generalized max-min, input density, and curve of iso-value $\phi(\mathbf{x}) = T$.

model is a key point to provide an accurate estimation and different strategies have been proposed to exploit the complete probabilistic description given by the Kriging in order to evaluate the minimal number of points on the real expensive input–output function. The main methods are described in the remainder of this chapter. A numerical comparison of different Kriging-based methods to estimate a probability of failure can be found in Li, Bect, & Vazquez (2010).

8.2.3.2 Description of Kriging

Kriging (Kleijnen, 2009; Matheron, 1963; Sasena, 2002) is a statistical surrogate model that can be used to approximate the input–output function $\phi(\cdot)$ on its input space $\mathbb{X} \subset \mathbb{R}^d$. It requires only a small initial training set, $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$, for which the function values have been computed and stored into $\boldsymbol{\phi}_p = [\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_p)]^T$. The Kriging model consists of a Gaussian process $\Upsilon(\cdot)$ that is expressed for any input vector $\mathbf{x} \in \mathbb{X}$ as

$$\Upsilon(\mathbf{x}) = m(\mathbf{x}) + \zeta(\mathbf{x}) \tag{8.26}$$

where the mean function $m(\cdot)$ is an optional regression model estimated from available data (polynomial in $\mathbf{x}$, for example), and $\zeta(\cdot)$ is a zero-mean Gaussian process with covariance function $\mathrm{cov}(\cdot, \cdot)$. Because the actual covariance function is unknown in practice, it can be modeled as

$$\operatorname{cov}\left(\zeta\left(\mathbf{x}_i\right), \zeta\left(\mathbf{x}_j\right)\right) = \sigma_\zeta^2 \operatorname{Corr}\left(\mathbf{x}_i, \mathbf{x}_j\right) \tag{8.27}$$

where σ_ζ^2 is the process variance and $\operatorname{Corr}(\cdot,\cdot)$ is a parametric correlation function. A classical choice for this correlation function is

$$\operatorname{Corr}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\sum_{l=1}^{d} \theta_l |x_{il} - x_{jl}|^{p_l}\right) \tag{8.28}$$

where the parameters $0 < p_l \leq 2$ reflect the smoothness of the interpolation (2 is the smoothest and corresponds to the Gaussian correlation function) whereas the θ_l is scale factor that can be estimated, for example, by maximum likelihood (Sasena, 2002). Other correlation functions exist in literature and the choice of the correlation functions can have a large impact on the prediction accuracy (Xie, Nelson, & Staum, 2010).

Kriging provides an optimal unbiased linear predictor at any $\mathbf{x} \in \mathbb{X}$ as

$$\hat{\phi}(\mathbf{x}, \mathcal{X}) = m(\mathbf{x}) + \mathbf{r}\left(\mathbf{x}, \mathcal{X}\right)^T \mathbf{R}^{-1}(\mathcal{X}) \left(\boldsymbol{\phi}_p(\mathcal{X}) - \mathbf{m}_p(\mathcal{X})\right) \tag{8.29}$$

where

$$\begin{cases} \mathbf{R}_{|ij}(\mathcal{X}) = \operatorname{Corr}\left(\mathbf{x}_i, \mathbf{x}_j\right) \\ \mathbf{r}\left(\mathbf{x}, \mathcal{X}\right) = \left[\operatorname{Corr}\left(\mathbf{x}, \mathbf{x}_1\right), \ldots, \operatorname{Corr}\left(\mathbf{x}, \mathbf{x}_p\right)\right]^T \\ \mathbf{m}_p(\mathcal{X}) = \left[m\left(\mathbf{x}_1\right), \ldots, m\left(\mathbf{x}_p\right)\right]^T \end{cases} \tag{8.30}$$

Moreover, the use of Gaussian processes makes it possible to compute confidence intervals for the prediction Equation (8.29) through the variance

$$\sigma^2(\mathbf{x}) = \sigma_\zeta^2 \left(1 - \mathbf{r}(\mathbf{x}, \mathcal{X})^T \mathbf{R}^{-1}(\mathcal{X}) \mathbf{r}(\mathbf{x}, \mathcal{X})\right) \tag{8.31}$$

An illustration of Kriging interpolation and corresponding confidence interval is given in Figure 8.4.

Remark. *Because the Kriging model directly involves a spatial correlation matrix between the training set samples and its inversion, this surrogate model suffers from the curse of dimensionality and can be intractable when the dimension of the input space and the number of training set samples increases. Moreover, this surrogate can present difficulties in modeling a discontinuous function.*

8.2.3.3 Refinement strategies

In case of rare event probability estimation, the surrogate model must be accurate in the zones of relevance, that is, in the vicinity of the threshold T and in the high probability content regions. The use of the exact function ϕ and its surrogate $\hat{\phi}$ in the probability calculation will lead to the same result if $\forall \mathbf{x} \in \mathbb{R}^d, \mathbf{1}_{\phi(\mathbf{x})>T} = \mathbf{1}_{\hat{\phi}(\mathbf{x},X)>T}$

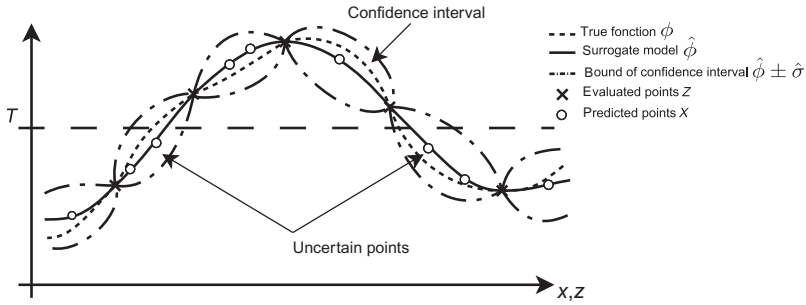

Figure 8.4 Kriging model and corresponding confidence interval.

(Figure 8.4). In other words, the surrogate model might not be representative of the exact function outside the zones of interest as it does not take part of the probability estimation. From the initial training set $\mathcal{X}$, the Kriging properties (i.e., Gaussian process, estimation of the predicted error variance) are valuable in determining the additional samples that must be evaluated on $\phi(\cdot)$ to refine its surrogate model. Different refinement strategies have been developed in the literature and are briefly described here. Two categories can be distinguished in the methods: the direct and the one-step look-ahead methods. The first ones use directly the Kriging model to determine the sample to be added to the training set whereas the latter ones estimate the influence of the training set candidate sample on the updated surrogate model (i.e., Kriging model built from the candidate point in addition to past training set samples). The refinement-stopping criteria used in the different methods are based on the Kriging prediction error in order to evaluate the accuracy of the surrogate model and its impact on the rare event probability estimate.

Direct methods

Active learning reliability method combining Kriging and probability estimation method This method (Echard, 2012) determines a new sample point $\mathbf{x}$ to add to the training set $\mathcal{X}$ by solving the following optimization problem:

$$\max_{\mathbf{x}} \left[1 - \Phi_{0,1} \left(\frac{|T - \hat{\phi}(\mathbf{x}, \mathcal{X})|}{\sigma(\mathbf{x}, \mathcal{X})} \right) \right] \tag{8.32}$$

where $\Phi_{0,1}(\cdot)$ is the cdf of the standard Gaussian distribution. The used criterion generates a sample for which the Kriging prediction is closed to the threshold (numerator of Equation 8.32) and which presents a high prediction error (denominator of Equation 8.32). Due to the monotonicity of the involved cdf, the optimization problem (Equation 8.32) is equivalent to:

$$\min_{\mathbf{x}} \frac{|T - \hat{\phi}(\mathbf{x}, \mathcal{X})|}{\sigma(\mathbf{x}, \mathcal{X})} \tag{8.33}$$

This criterion has been coupled with CMC (Echard et al., 2011), importance sampling (Echard, Gayton, Lemaire, & Relun, 2013), and AST (Echard, 2012). In practice, the optimization problem is not solved, and given a sample set $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ provided by CMC, IS, or AST, the new sample that will be added to the training set is determined by

$$\mathbf{X} = \underset{\mathbf{X}_1,\ldots,\mathbf{X}_N}{\operatorname{argmin}} \left\{ \frac{|T - \hat{\phi}(\mathbf{X}_1, \mathcal{X})|}{\sigma(\mathbf{X}_1, \mathcal{X})}, \ldots, \frac{|T - \hat{\phi}(\mathbf{X}_N, \mathcal{X})|}{\sigma(\mathbf{X}_N, \mathcal{X})} \right\}$$

Efficient global reliability analysis The efficient global reliability analysis (EGRA) criterion, initially proposed in Ranjan, Bingham, & Michailidis (2008) and Bichon, Eldred, Swiler, Mahadevan, & McFarland (2008) and generalized in Bect et al. (2012), relies on an optimization problem involving an integral in the vicinity of the threshold in order to find the point to add to the training set. The corresponding optimization problem is the following:

$$\max_{\mathbf{x}} \mathbb{E}_{\Upsilon} \left(\mathbf{1}_{(k.\sigma(\mathbf{x},\mathcal{X}))^{\delta} - |T - \upsilon(\mathbf{x},\mathcal{X})|^{\delta} \geq 0} \right) \tag{8.34}$$

where υ denotes a realization of the Gaussian process Υ (Equation 8.26) and $\mathbb{E}_{\Upsilon}(\cdot)$ stands for the expectation of the Kriging model. Depending of the value of δ, this criterion is equivalent to the one proposed in Bichon et al. (2008) ($\delta = 1$) and the one described in Ranjan et al. (2008) ($\delta = 2$). EGRA can be associated with simulation techniques for rare event estimation and has been for instance applied with importance sampling (Bichon et al., 2008).

Margin probability function In Oakley (2004), the authors propose a statistic based on the so-called *margin probability* to refine the Kriging model. Let us define the following sets:

$$\hat{\mathbb{X}}^{-1}_{1-\alpha}(\mathbb{X}) = \left\{ \mathbf{x} \in \mathbb{X} \colon \hat{\phi}(\mathbf{x}, \mathcal{X}) \leq T - k_{1-\alpha}\sigma^2(\mathbf{x}, \mathcal{X}) \right\} \tag{8.35}$$

$$\hat{\mathbb{X}}^{+1}_{1-\alpha}(\mathbb{X}) = \left\{ \mathbf{x} \in \mathbb{X} \colon \hat{\phi}(\mathbf{x}, \mathcal{X}) \leq T + k_{1-\alpha}\sigma^2(\mathbf{x}, \mathcal{X}) \right\} \tag{8.36}$$

with $1 - \alpha$ the confidence level of the Kriging model (e.g., a confidence interval $1 - \alpha$ of 95% corresponds to $k_{1-\alpha} \approx 1.96$). The margin probability is defined by

$$\begin{aligned} \mathbb{MP}(\mathbf{x}) &= \mathbb{P}\left[\hat{\phi}(\mathbf{x}, \mathcal{X}) \in \left(\widehat{\mathbb{X}}^{+1}_{1-\alpha}(\mathbb{X}) \setminus \widehat{\mathbb{X}}^{-1}_{1-\alpha}(\mathbb{X}) \right) \right], \\ &= \Phi_{0,1}\left(\frac{T + k_{1-\alpha}\sigma(\mathbf{x}, \mathcal{X}) - \hat{\phi}(\mathbf{x}, \mathcal{X})}{\sigma(\mathbf{x}, \mathcal{X})} \right) \\ &\quad - \Phi_{0,1}\left(\frac{T - k_{1-\alpha}\sigma(\mathbf{x}, \mathcal{X}) - \hat{\phi}(\mathbf{x}, \mathcal{X})}{\sigma(\mathbf{x}, \mathcal{X})} \right) \end{aligned} \tag{8.37}$$

The new point that is added to the training set is determined with

$$\max_{\mathbf{x}} \mathbb{MP}(\mathbf{x})$$

This criterion has been used in Picheny, Ginsbourger, Roustant, Haftka, & Kim (2010) to develop the target integrated mean square error (TIMSE) as described in the "Target integrated mean square error" section.

Generalized max-min From our knowledge, the generalized max-min (Lacaze & Missoum, 2014) (see the "Improvement of max-min technique: generalized max-min" section) has not been applied for Kriging refinement purposes. However, this method is not dedicated to SVM and can be used to sample points in the zones of interest regarding to the probability estimation.

One-step look-ahead methods

The one-step look-ahead methods differ from the direct methods in using an estimation of the effect of the updated kriging model (i.e., based on the training set augmented with the new candidate sample) to refine the surrogate model. For numerical tractability and to estimate the influence of an added sample in the training set, these methods consider that the variance of the Gaussian process σ_ζ^2 in Equation (8.31), the kernel parameters, and the regression model do not change, and only the correlation matrix is updated.

Target integrated mean square error The TIMSE proposed in Picheny et al. (2010) extends the margin probability (see the "Margin probability function" section) to define a criterion that is an integration over the uncertain space in the vicinity of the threshold. To define the vicinity of the threshold, a weighting function is introduced

$$W(\mathbf{x}', \{\mathbf{x}, \mathcal{X}\}) = \frac{1}{\sigma(\mathbf{x}', \{\mathbf{x}, \mathcal{X}\})\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\hat{\phi}(\mathbf{x}', \{\mathbf{x}, \mathcal{X}\}) - T}{\sigma(\mathbf{x}', \{\mathbf{x}, \mathcal{X}\})}\right)^2\right)$$

with $\mathbf{x}' \in \mathbb{X}$, $\mathbf{x}$ the candidate sample to add to the training set $\mathcal{X}$ and $\{\mathbf{x}, \mathcal{X}\}$ the increased training set.

The determination of the new sample results from solving an optimization problem involving an expected value calculation of the mean square error of the Kriging model weighted by $W(\cdot)$:

$$\min_{\mathbf{x}} \int_{\mathbb{X}} \sigma\left(\mathbf{x}', \{\mathbf{x}, \mathcal{X}\}\right)^2 W(\mathbf{x}', \{\mathbf{x}, \mathcal{X}\}) f(\mathbf{x}')\, \mathrm{d}\mathbf{x}'$$

This criterion minimizes the global uncertainty over the uncertain space in the vicinity of the threshold, taking into account the influence of the candidate sample and the initial pdf of the input variables. From our knowledge, this criterion has been coupled

with subset simulation (Picheny et al., 2010) but is compatible with all the simulation techniques used for rare event estimation.

Reduction of misclassification uncertainty for rare event simulation techniques This method is described in Balesdent et al. (2013) and consists in analyzing the different samples generated by the used probability estimation techniques. This method is used to find the sample point to add to the current training set in order to best reduce the uncertainty of the generated samples that can be misclassified because of the use of Kriging model. Let $\mathbb{X}_N = \{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ be a sample set provided by the rare event simulation technique. Using the notations of Equations (8.35) and (8.36), let $\tilde{\mathbb{X}}_N = \widehat{\mathbb{X}}^{+1}_{1-\alpha}(\mathbb{X}_N) \setminus \widehat{\mathbb{X}}^{-1}_{1-\alpha}(\mathbb{X}_N)$ be the uncertain samples of $\mathbb{X}_N$ regarding the threshold exceedance, that is, the samples that can be misclassified because of the Kriging model uncertainty. The new sample to add to the current training set is obtained by solving the following optimization problem:

$$\max_{\mathbf{x}} \sum_{i=1}^{\mathrm{Card}(\tilde{\mathbb{X}}_N)} \left(\sigma\left(\tilde{\mathbf{X}}_i, \mathcal{X}\right) - \sigma\left(\tilde{\mathbf{X}}_i, \{\mathbf{x}, \mathcal{X}\}\right)\right) \tag{8.38}$$

with $\mathrm{Card}(\tilde{\mathbb{X}}_N)$ the number of samples present in the uncertain sample set $\tilde{\mathbb{X}}_N$. This criterion allows to refine the Kriging model with the sample that has the highest influence on the total standard deviation of the uncertain sample set. This method has been applied in Balesdent et al. (2013) with parametric and nonparametric importance sampling but is also applicable with all the simulation techniques used to estimate a rare event probability.

Stepwise uncertainty reduction The stepwise uncertainty reduction (SUR) (Bect et al., 2012) determines the new sample $\mathbf{x}$ to add to the current training set $\mathcal{X}$ by solving an optimization problem. Given the two following expressions

$$\tau(\mathbf{x}, \mathcal{X}) = 1 - \Phi_{0,1}\left(\frac{|\upsilon(\mathbf{x}, \mathcal{X}) - T|}{\sigma(\mathbf{x}, \mathcal{X})}\right)$$
$$\nu(\mathbf{x}, \mathcal{X}) = \tau(\mathbf{x}, \mathcal{X})(1 - \tau(\mathbf{x}, \mathcal{X}))$$

four SUR criteria to minimize have been introduced

$$J_1^{\mathrm{SUR}}(\mathbf{x}, \mathcal{X}) = \mathbb{E}_\Upsilon\left[\left(\int \sqrt{\tau(\mathbf{x}', \{\mathbf{x}, \mathcal{X}\})} f(\mathbf{x}')\, \mathrm{d}\mathbf{x}'\right)^2\right] \tag{8.39}$$

$$J_2^{\mathrm{SUR}}(\mathbf{x}, \mathcal{X}) = \mathbb{E}_\Upsilon\left[\left(\int \sqrt{\nu(\mathbf{x}', \{\mathbf{x}, \mathcal{X}\})} f(\mathbf{x}')\, \mathrm{d}\mathbf{x}'\right)^2\right] \tag{8.40}$$

$$J_3^{\mathrm{SUR}}(\mathbf{x}, \mathcal{X}) = \mathbb{E}_\Upsilon\left[\int \tau(\mathbf{x}', \{\mathbf{x}, \mathcal{X}\}) f(\mathbf{x}')\, \mathrm{d}\mathbf{x}'\right] \tag{8.41}$$

$$J_4^{\mathrm{SUR}}(\mathbf{x}, \mathcal{X}) = \mathbb{E}_\Upsilon\left[\int \nu(\mathbf{x}', \{\mathbf{x}, \mathcal{X}\}) f(\mathbf{x}')\, \mathrm{d}\mathbf{x}'\right] \tag{8.42}$$

with $\mathbb{E}_{\Upsilon}(\cdot)$ the expectation of the Kriging model. These criteria involve a double integration, one over the uncertain space and one over the Gaussian process space. This method has been applied with crude Monte Carlo (Bect et al., 2012) and subset simulation (Li et al., 2012) but is applicable with all the simulation techniques used to estimate rare event probability.

K-means clustering strategy for Kriging refinement

The K-means clustering strategy (Dubourg, 2011; Dubourg et al., 2011; Hartigan & Wong, 1979) is applicable with all the previous criteria. This method is not based on an optimization method to refine the Kriging model but consists of two steps: a sampling step and a clustering step in order to provide a set of new samples to add to the current training set. From the chosen refinement criterion, this method generates new samples using a Monte Carlo Markov chain (MCMC) method (see Chapter 2). The criterion is used to create the target distribution employed in the proposal/refusal step of MCMC. Then a K-means clustering method (MacQueen, 1967) allows to select the samples relevant to the Kriging refinement. These cluster centroids are obtained by solving an optimization problem involving the $\mathcal{L}_2$ distance with respect to the MCMC samples. Then these centroids are added to the training set. This method has been associated with subset sampling (Dubourg et al., 2011) and importance sampling (Dubourg, Sudret, & Deheeger, 2013).

8.2.4 Conclusion

In this section are described two of the most used surrogate models that can be associated with rare event probability estimation techniques. In our experience, Kriging is very efficient for modeling relatively smooth input–output functions in small to moderate dimensions (e.g., a few dozen). When the dimension on the input space increases or $\phi(\cdot)$ presents discontinuities, the use of the SVM model instead of Kriging is recommended. The characteristics of the surrogate model for rare event estimation are summarized in Table 8.1.

Table 8.1 Summary on surrogate model for rare event probability estimation

Criteria	Estimation characteristics
Rarity of $\mathbb{P}_f$	Often not adapted to very rare events ($\mathbb{P}_f < 10^{-6}$)
Simulation budget	From 10 to 10^3 samples in most cases
Type of function $\phi(\cdot)$	Regularity assumptions (for Kriging)
Multimodality of f restricted to Ω_f	Particularly adapted to multimodal failure region
Dimension of $\mathbf{X}$	Depends on the type of surrogate model
Probability estimate error	Estimation with retrials
Difficulty	High; computation of complex sample criteria

References

Balesdent, M., Morio, J., and Marzat, J. (2013). Kriging-based adaptive importance sampling algorithms for rare event estimation. *Structural Safety*, *13*, 1–10.

Basak, D., Pal, S., and Patranabis, D. C. (2007). Support vector regression. *Neural Information Processing-Letters and Reviews*, *11*(10), 203–224.

Basudhar, A. (2011). Computational optimal design and uncertainty quantification of complex systems using explicit decision boundaries. Unpublished doctoral dissertation, University of Arizona.

Basudhar, A., and Missoum, S. (2008). Adaptive explicit decision functions for probabilistic design and optimization using support vector machines. *Computers & Structures*, *86*(19–20), 1904–1917.

Basudhar, A., and Missoum, S. (2010). An improved adaptive sampling scheme for the construction of explicit boundaries. *Structural and Multidisciplinary Optimization*, *42*(4), 517–529.

Basudhar, A., and Missoum, S. (2013). Reliability assessment using probabilistic support vector machines. *International Journal of Reliability and Safety*, *7*(2), 156–173.

Bect, J., Ginsbourger, D., Li, L., Picheny, V., and Vazquez, E. (2012). Sequential design of computer experiments for the estimation of a probability of failure. *Statistics and Computing*, *22*(3), 773–793.

Bichon, B. J., Eldred, M. S., Swiler, L. P., Mahadevan, S., and McFarland, J. M. (2008). Efficient global reliability analysis for nonlinear implicit performance functions. *AIAA Journal*, *46*(10), 2459–2468.

Blatman, G., and Sudret, B. (2008). Adaptive sparse polynomial chaos expansions-application to structural reliability. In *Proceedings of the 4th international ASRANet colloquium*. Athens, Greece.

Bourinet, J.-M., Deheeger, F., and Lemaire, M. (2011). Assessing small failure probabilities by combined subset simulation and support vector machines. *Structural Safety*, *33*(6), 343–353.

Bucher, C., and Bourgund, U. (1990). A fast and efficient response surface approach for structural reliability problems. *Structural Safety*, *7*(1), 57–66.

Caniou, Y. (2012). Global sensitivity analysis for nested and multiscale modelling. Unpublished doctoral dissertation, Université Blaise Pascal-Clermont-Ferrand II.

Cannamela, C., Garnier, J., and Iooss, B. (2008). Controlled stratification for quantile estimation. *Annals of Applied Statistics*, *2*(4), 1554–1580.

Chastaing, G., Gamboa, F., and Prieur, C. (2014). Generalized Sobol sensitivity indices for dependent variables: Numerical methods. *Journal of Statistical Computation and Simulation*, *85*(7), 1306–1333.

Chastaing, G., Gamboa, F., and Prieur, C. (2012). Generalized Hoeffding-Sobol decomposition for dependent variables-application to sensitivity analysis. *Electronic Journal of Statistics*, *6*, 2420–2448.

Clarke, S. M., Griebsch, J. H., and Simpson, T. W. (2005). Analysis of support vector regression for approximation of complex engineering analyses. *Journal of Mechanical Design*, *127*(6), 1077–1087.

Cristianini, N., and Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. New York, NY: Cambridge University Press.

Dai, H., Zhang, H., Wang, W., and Xue, G. (2012). Structural reliability assessment by local approximation of limit state functions using adaptive Markov chain simulation and support vector regression. *Computer-Aided Civil and Infrastructure Engineering*, *27*(9), 676–686.

Das, P., and Zheng, Y. (2000). Cumulative formation of response surface and its use in reliability analysis. *Probabilistic Engineering Mechanics*, *15*(4), 309–315.

Deheeger, F., and Lemaire, M. (2007). Support vector machine for efficient subset simulations: 2SMART method. In *10th international conference on application of statistics and probability in civil engineering, proceedings and monographs in engineering, water and earth sciences* (pp. 259–260).

Duan, K.-B., and Keerthi, S. S. (2005). Which is the best multiclass SVM method? An empirical study. In *Multiple classifier systems* (pp. 278–285). Berlin: Springer.

Dubourg, V. (2011). Adaptive surrogate models for reliability analysis and reliability-based design optimization. Unpublished doctoral dissertation, Université Blaise Pascal.

Dubourg, V., Deheeger, E., and Sudret, B. (2011). Metamodel-based importance sampling for the simulation of rare events. In M. Faber, J. Kohler, and K. Nishilima (Eds.), *Proceedings of the 11th international conference of statistics and probability in civil engineering (ICASP2011)*. Zurich, Switzerland.

Dubourg, V., Sudret, B., and Deheeger, F. (2013). Metamodel-based importance sampling for structural reliability analysis. *Probabilistic Engineering Mechanics*, *33*, 47–57.

Echard, B. (2012). Kriging-based reliability assessment of structures submitted to fatigue. Unpublished doctoral dissertation, Université Blaise Pascal.

Echard, B., Gayton, N., and Lemaire, M. (2011). AK-MCS : An active learning reliability method combining Kriging and Monte Carlo simulation. *Structural Safety*, *33*, 145–154.

Echard, B., Gayton, N., Lemaire, M., and Relun, N. (2013). A combined importance sampling and kriging reliability method for small failure probabilities with time-demanding numerical models. *Reliability Engineering & System Safety*, *111*, 232–240.

Ekström, P.-A., and Broed, R. (2006). *Sensitivity analysis methods and a biosphere test case implemented in EIKOS* (Tech. Rep.). Posiva Working Report, Helsinki (Finland).

Gao, J. B., Gunn, S. R., Harris, C. J., and Brown, M. (2002). A probabilistic framework for SVM regression and error bar estimation. *Machine Learning*, *46*(1–3), 71–89.

Garza, J., and Millwater, H. (2012). Sensitivity of probability-of-failure estimates with respect to probability of detection curve parameters. *International Journal of Pressure Vessels and Piping*, *92*, 84–95.

Gayton, N., Bourinet, J., and Lemaire, M. (2003). CQ2RS: A new statistical approach to the response surface method for reliability analysis. *Structural Safety*, *25*(1), 99–121.

Gomes, H., and Awruch, A. (2004). Comparison of response surface and neural network with other methods for structural reliability analysis. *Structural Safety*, *26*, 49–67.

Gunn, S. R. (1998). *Support vector machines for classification and regression*. *ISIS technical report*, *14*.

Hartigan, J. A., and Wong, M. A. (1979). Algorithm AS 136: A *k*-means clustering algorithm. *Applied Statistics*, *28*, 100–108.

Helton, J. C., Johnson, J. D., Sallaberry, C. J., and Storlie, C. B. (2006). Survey of sampling-based methods for uncertainty and sensitivity analysis. *Reliability Engineering & System Safety*, *91*(10), 1175–1209.

Hu, C., and Youn, B. (2011). Adaptive-sparse polynomial chaos expansion for reliability analysis and design of complex engineering systems. *Structural and Multidisciplinary Optimization*, *43*(3), 419–442.

Hurtado, J. E. (2004). An examination of methods for approximating implicit limit state functions from the viewpoint of statistical learning theory. *Structural Safety*, *26*(3), 271–293.

Hurtado, J. E., and Alvarez, D. A. (2000). Reliability assessment of structural systems using neural networks. In *Proceedings of European congress on computational methods in applied sciences and engineering, ECCOMAS.*

Hurtado, J. E., and Alvarez, D. A. (2003). Classification approach for reliability analysis with stochastic finite-element modeling. *Journal of Structural Engineering*, *129*(8), 1141–1149.

Hurtado, J. E., and Alvarez, D. A. (2010). An optimization method for learning statistical classifiers in structural reliability. *Probabilistic Engineering Mechanics*, *25*(1), 26–34.

Iooss, B. (2011). Revue sur l'analyse de sensibilité globale de modèles numériques (in French). *Journal de la Société Française de Statistique*, *152*(1), 3–25.

Iooss, B., and Lemaître, P. (2015). A review on global sensitivity analysis methods. In C. Meloni and G. Dellino (Eds.), *Uncertainty management in simulation-optimization of complex systems: Algorithms and applications.* Berlin: Springer.

Jansen, M. J. (1999). Analysis of variance designs for model output. *Computer Physics Communications*, *117*(1), 35–43.

Kim, S.-H., and Na, S.-W. (1997). Response surface method using vector projected sampling points. *Structural Safety*, *19*(1), 3–19.

Kleijnen, J. P. C. (2009). Kriging metamodeling in simulation: A review. *European Journal of Operational Research*, *192*(3), 707–716.

Kuhn, H. W., and Tucker, A. W. (1951). Nonlinear programming. In J. Neyman (Ed.), *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability* (pp. 481–492). Berkeley, California: University of California Press.

Lacaze, S., and Missoum, S. (2013). Reliability-based design optimization using kriging and support vector machines. In *11th international conference on structural safety & reliability*. New York, NY, USA. January 2013.

Lacaze, S., and Missoum, S. (2014). A generalized "max-min" sample for surrogate update. *Structural and Multidisciplinary Optimization*, *49*(4), 683–687.

Lamboni, M., Monod, H., and Makowski, D. (2011). Multivariate sensitivity analysis to measure global contribution of input factors in dynamic models. *Reliability Engineering & System Safety*, *96*(4), 450–459.

Li, H.-S., Lü, Z.-Z., and Yue, Z.-F. (2006). Support vector machine for structural reliability analysis. *Applied Mathematics and Mechanics*, *27*, 1295–1303.

Li, J., and Xiu, D. (2010). Evaluation of failure probability via surrogate models. *Journal of Computational Physics*, *229*, 8966–8980.

Li, L., Bect, J., and Vazquez, E. (2010). A numerical comparison of two sequential kriging-based algorithms to estimate a probability of failure. In *Uncertainty in computer model conference.* Sheffield, UK. July 12–14, 2010.

Li, L., Bect, J., and Vazquez, E. (2012). Bayesian subset simulation: A kriging-based subset simulation algorithm for the estimation of small probabilities of failure. In *Proceedings of PSAM 11 and ESREL 2012*. Helsinki, Finland. June 25–29, 2012.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 281–297).

Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, *58*(8), 1246.

Monod, H., Naud, C., and Makowski, D. (2006). Uncertainty and sensitivity analysis for crop models. In D. Wallach, D. Makowski, and J. Jones (Eds.), *Working with dynamic crop models* (pp. 55–100). Amsterdam, Boston, Heidelberg: Academic Press-Elsevier.

Morio, J. (2011). Influence of input PDF parameters of a model on a failure probability estimation. *Simulation Modelling Practice and Theory*, *19*(10), 2244–2255.

Morris, M. D. (1991). Factorial sampling plans for preliminary computational experiments. *Technometrics*, *33*(2), 161–174.

Oakley, J. (2004). Estimating percentiles of uncertain computer code outputs. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *53*(1), 83–93.

Papadrakakis, M., and Lagaros, N. D. (2002). Reliability-based structural optimization using neural networks and Monte Carlo simulation. *Computer Methods in Applied Mechanics and Engineering*, *191*(32), 3491–3507.

Picheny, V., Ginsbourger, D., Roustant, O., Haftka, R. T., and Kim, N.-H. (2010). Adaptive designs of experiments for accurate approximation of a target region. *Journal of Mechanical Design*, *132*(7), 071008.

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in large margin classifiers* (pp. 61–74). Cambridge, MA: MIT Press.

Plischke, E., Röhlig, K.-J., Badea, A., Lavín, R. B., Ekström, P.-A., and Hotzel, S. (2009). *Sensitivity analysis benchmark based on the use of analytic and synthetic PA cases milestone (N: M2. 1. D. 11). European commission report*. PAMINA project.

Ranjan, P., Bingham, D., and Michailidis, G. (2008). Sequential experiment design for contour estimation from complex computer codes. *Technometrics*, *50*(4), 527–541.

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., and Tarantola, S. (2010). Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Computer Physics Communications*, *181*(2), 259–270.

Saltelli, A., Tarantola, S., and Campolongo, F. (2000). Sensitivity analysis as an ingredient of modeling. *Statistical Science*, *15*(4), 377–395.

Saltelli, A., Tarantola, S., and Chan, K.-S. (1999). A quantitative model-independent method for global sensitivity analysis of model output. *Technometrics*, *41*(1), 39–56.

Santner, T., Williams, B., and Notz, N. (2003). *The design and analysis of computer experiments*. New York, NY: Springer-Verlag.

Sasena, M. (2002). Flexibility and efficiency enhancements for constrained global design optimization with kriging approximation. Unpublished doctoral dissertation, University of Michigan.

Schueremans, L., and Van Gemert, D. (2005). Use of kriging as meta-model in simulation procedures for structural reliability. In *9th international conference on structural safety and reliability*. Rome (pp. 2483–2490).

Shawe-Taylor, J., and Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge: Cambridge University Press.

Smola, A. J., and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*(3), 199–222.

Sobol, I. M. (1990). On sensitivity estimation for nonlinear mathematical models. *Matematicheskoe Modelirovanie*, *2*(1), 112–118.

Sobol, I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation*, *55*(1–3), 271–280.

Sobol, I. M., and Kucherenko, S. (2009). Derivative based global sensitivity measures and their link with global sensitivity indices. *Mathematics and Computers in Simulation*, *79*(10), 3009–3017.

Song, H., Choi, K. K., Lee, I., Zhao, L., and Lamb, D. (2013). Adaptive virtual support vector machine for reliability analysis of high-dimensional problems. *Structural and Multidisciplinary Optimization*, *47*(4), 479–491.

Steinwart, I., and Christmann, A. (2008). *Support vector machines*. New York, NY: Springer.

Sudret, B. (2012). Meta-models for structural reliability and uncertainty quantification. In S. Q. K. Phoon, M. Beer, and S. Pang (Eds.), *Asian-pacific symposium on structural reliability and its applications*. Singapore, Singapore (pp. 53–76).

Tou, J. T., and Gonzalez, R. C. (1974). *Pattern recognition principles*. London, UK: Addison Wesley.

Vapnik, V. N. (2000). *The nature of statistical learning theory. Statistics for engineering and information science*. Berlin: Springer-Verlag.

Vapnik, V. N. (1998). *Statistical learning theory* (Vol. 2). New York, USA: Wiley.

Vazquez, E., and Bect, J. (2009). A sequential Bayesian algorithm to estimate a probability of failure. In *15th IFAC, symposium on system identification (SYSID'09).* 5 pp. Saint-Malo, France. July 6–8.

Wang, Y., Yu, X., and Du, X. (2012). Reliability analysis with SVM and gradient information at MPP. In *Proceedings of the seventh China-Japan-Korea joint symposium on optimization of structural and mechanical systems.* Huanshang, China, June.

Xie, W., Nelson, B., and Staum, J. (2010). The influence of correlation functions on stochastic kriging metamodels. In *Simulation conference (WSC), proceedings of the 2010 winter* (pp. 1067–1078).

Special developments for time-variant systems

9

D. Jacquemart, J. Morio, F. Le Gland, M. Balesdent

In this chapter, we analyze the different methods for estimating the probability of a rare event in dynamical systems modeled by Markov chains. These algorithms are relatively similar in principle to the ones defined for static input-output systems. Some special developments must nevertheless be considered for an efficient applicability.

9.1 General notations

Let us recall the notation of Chapter 3 and consider a continuous time Markov process $(\mathbf{X}_t)_{0\leq t\leq S}$ in state space $\mathbb{R}^d$ with S a stopping time. Its initial law at time 0 is defined by $\pi_0(\mathbf{x}_0)$, and its transition kernels are denoted $\pi_{s,t}(\mathbf{x}, \mathbf{x}')$. In this chapter, we describe different algorithms that propose a statistic that can be efficient for estimating the probability $\mathbb{P}(\mathcal{R})$ with $\mathcal{R}$ defined by

$$\mathcal{R} = \{\mathbf{X}_t, \text{ with } \varphi(\mathbf{X}_t) \geq T \text{ and } 0 \leq t \leq S\}$$

where T is a given threshold and $\varphi(\cdot)$ is a real valued function. Let us also consider the random variable Z defined by

$$Z = \sup_{0\leq t\leq S} \varphi(\mathbf{X}_t) \tag{9.1}$$

We can remark that Z does not depend on time t and is consequently a static variable. The probability $\mathbb{P}(\mathcal{R})$ is thus given by

$$\mathbb{P}(\mathcal{R}) = \mathbb{P}(Z > T)$$

Because it is not possible to directly simulate $(\mathbf{X}_t)_{0\leq t\leq S}$ in practice, discrete time approach can also be considered (see Section 3.2) with $t = j\Delta_t$ and $S = c\Delta_t$. A discrete time Markov chain $(\mathbf{X}_j)_{0\leq j\leq c}$ with initial law $\pi_0(\mathbf{x}_0)$ at time 0 and transition kernels $\pi_j(\mathbf{x}_j|\mathbf{x}_{j-1})$ is simulated. The following describes the principle of the different methods for rare event probability estimation with a continuous or a discrete formalism without any loss of generality.

9.2 Toy case

The different algorithms proposed for probability estimation on dynamical systems are compared in a general toy case, the standard one-dimensional Brownian bridge. It consists in a time nonhomogeneous process that solves the following stochastic differential equation:

Estimation of Rare Event Probabilities in Complex Aerospace and Other Systems
http://dx.doi.org/10.1016/B978-0-08-100091-5.00009-5

$$\mathrm{d}X_t = \frac{X_t}{t-1}\,\mathrm{d}t + \mathrm{d}W_t, \quad \text{with } X_0 = 0$$

defined for $t \in [0, 1]$. The term W_t is a standard Brownian motion process, that is, W_t follows a normal distribution with mean 0 and variance t at time t. The target probability can be expressed with the function $\varphi(\cdot)$ that corresponds here to the identity function. It can be shown that $X_1 = 0$ almost surely and that the probability of interest has the following theoretical expression (Borodin & Salminen, 2002),

$$\mathbb{P}\left(\sup_{0\le t\le 1}(X_t) > T\right) = \mathbb{P}(Z > T) = \exp\left(-2T^2\right)$$

With a threshold T equal to 2, the theoretical probability of exceedance is 3.35×10^{-4}. The probability of exceedance with $T = 3$ is equal to 1.52×10^{-8}. The Euler scheme (Kroese, Taimre, & Botev, 2011) leads to the following discretization for the simulation of the Brownian bridge process:

$$X_{j+1} = X_j + \frac{c}{jc-1}X_j + \sqrt{c}W_j$$

where W_j has a standard normal distribution. In this particular simple case, the transition probabilities of the Brownian bridge have an explicit expression and it is not necessary to use a numerical approximation scheme to sample the trajectories. In the different simulations proposed in the following sections, c is set to $c = 10^{-3}$.

9.3 Crude Monte Carlo

9.3.1 Principle

A usual way to estimate the probability $\mathbb{P}(\mathcal{R})$ is to consider crude Monte Carlo (CMC) method (Robert & Casella, 2005; Sobol, 1994). In practice, we can generate N iid as $(\mathbf{X}_t)_{0\le t\le S}$ Markov processes $(\mathbf{X}_t^{(1)})_{0\le t\le S}, (\mathbf{X}_t^{(2)})_{0\le t\le S}, \ldots, (\mathbf{X}_t^{(N)})_{0\le t\le S}$. One then estimates N samples $Z_1, \ldots, Z_N$ of the variable Z defined in Equation (9.1). A sample Z_i is thus given by

$$Z_i = \sup_{0\le t\le S} \varphi(\mathbf{X}_t^{(i)})$$

The probability $\mathbb{P}(\mathcal{R})$ is then estimated with

$$\hat{\mathbb{P}}_{\mathrm{CMC}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{Z_i > T}$$

ALGORITHM 16 Crude Monte Carlo simulations with Markov chains for probability estimation

Input: The discrete Markov chain $(\mathbf{X}_j)_{0\leq j\leq c}$ with initial distribution $\pi_0(\mathbf{x}_0)$ and transition kernels $\pi_j(\mathbf{x}_j|\mathbf{x}_{j-1})$, the number of samples N, the function $\varphi(\cdot)$, the threshold T, and the stopping time c.

Output: The probability estimate $\hat{\mathbb{P}}^{\text{CMC}}$.

1 Sample $(\mathbf{X}_j^{(1)})_{0\leq j\leq c}, (\mathbf{X}_j^{(2)})_{0\leq j\leq c}, \ldots, (\mathbf{X}_j^{(N)})_{0\leq j\leq c}$ with initial distribution $\pi_0(\mathbf{x}_0)$ and transition kernels $\pi_j(\mathbf{x}_j|\mathbf{x}_{j-1})$.

2 Estimate $Z_i = \sup_{0\leq j\leq c} \varphi(\mathbf{X}_j^{(i)})$.

3 Estimate $\hat{\mathbb{P}}^{\text{CMC}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{Z_i>T}$.

4 **return** $\hat{\mathbb{P}}^{\text{CMC}}$.

The CMC method for probability estimation with Markov processes is described in Algorithm 16. This estimator has exactly the same statistical properties as the CMC estimate defined in Section 5.1 for the static input–output function. CMC quantile estimate can also be derived in the same way as in Section 5.1.

9.3.2 Application to a toy case

Brownian bridge

CMC algorithm is applied to the Brownian bridge toy case. An illustration of 10 CMC sample paths is given in Figure 9.1. The probability estimates obtained with CMC for this toy case are provided in Table 9.1 for two different thresholds. The CMC estimate needs a significant number of samples to be accurate for rare event probability estimation. The slight bias in the CMC estimation results from the discretization of the Markov process.

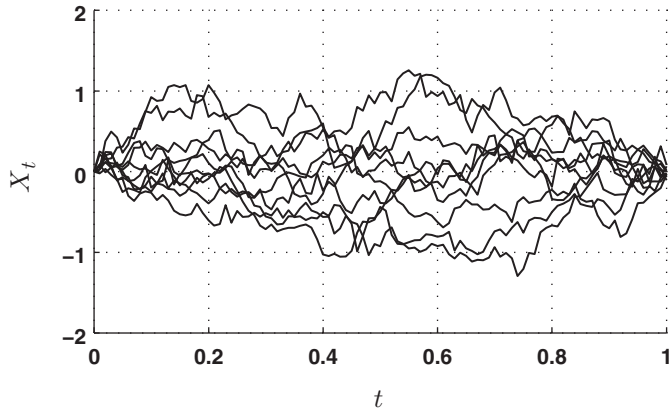

Figure 9.1 Some trials of CMC Brownian bridge sample paths.

Table 9.1 Results of CMC for the Brownian bridge

T	$\hat{\mathbb{P}}^{\text{CMC}}$	RB ($\hat{\mathbb{P}}^{\text{CMC}}$)	RE ($\hat{\mathbb{P}}^{\text{CMC}}$)	**Simulation budget**
2	2.4×10^{-4}	-15%	62%	10^4
2	2.4×10^{-4}	-15%	5%	10^6
3	0	n/a	n/a	10^4
3	0	n/a	n/a	10^6

Table 9.2 Summary of CMC rare event probability estimation for Markov processes

Criteria	**CMC characteristics**
Rarity of $\mathbb{P}(R)$	Not adapted. Slow convergence of CMC
Simulation budget	$\frac{10^2}{\mathbb{P}(R)}$ samples required to estimate $\mathbb{P}(R)$ with a 10% error
Type of process $\mathbf{X}_t$	No regularity condition on $\mathbf{X}_t$
Dimension of $\mathbf{X}_t$	No influence on the CMC convergence
Probability estimate error	Analytical formula depending on $\mathbb{P}(R)$ and N
Difficulty of implementation	None

9.3.3 Conclusion

The implementation of CMC on Markov processes is simple in most cases. However, as with input–output systems, CMC requires a high simulation budget to estimate a rare event probability with a decent relative error. This simulation budget is often not affordable in realistic applications. The main characteristics of CMC are summarized in Table 9.2. CMC is still seen as a reference due to its simple implementation and thus, its application to the use case is of interest.

9.4 Importance sampling

9.4.1 Principle

In the same way as described in Section 5.3 for static input–output function, the idea of importance sampling (IS) (Glynn & Iglehart, 1989; Juneja & Shahabuddin, 2001), also called *sequential importance sampling* when applied to Markov chains, is to replace the original probability distribution of the process by an auxiliary measure. Although the optimal change of density is not *a priori* Markovian, the IS auxiliary measure remains Markovian for convenience (Sandmann, 2005). An estimate of $\mathbb{P}(\mathcal{R})$ is obtained by generating Markov chains $(\mathbf{X}_j^{(i)})_{0 \leq j \leq c}$ for $i = 1, \ldots, N$ with respect to new transition kernels $\tilde{\pi}_j(\mathbf{x}_j | \mathbf{x}_{j-1})$ for $j = 1, \ldots, c$ and initial law $\tilde{\pi}_0(\mathbf{x}_0)$. Then we compute the IS probability estimate with

$$\hat{\mathbb{P}}^{\mathrm{IS}} = \frac{\sum_{i=1}^{N} \omega_i \mathbf{1}_{Z_i > T}}{\sum_{i=1}^{N} \omega_i}$$

with

$$Z_i = \sup_{0 \le j \le c} \varphi(\mathbf{X}_j^{(i)})$$

and

$$\omega_i = \frac{\pi_0(\mathbf{x}_0^{(i)}) \prod_{j=1}^{c} \pi_j(\mathbf{x}_j^{(i)} | \mathbf{x}_{j-1}^{(i)})}{\tilde{\pi}_0(\mathbf{x}_0^{(i)}) \prod_{j=1}^{c} \tilde{\pi}_j(\mathbf{x}_j^{(i)} | \mathbf{x}_{j-1}^{(i)})}$$

The main difficulty of IS is to determine valuable sampling densities $\tilde{\pi}_j$ that can reduce the CMC variance. For a general Markov process, there is, to our knowledge, no general method to determine an efficient IS auxiliary measure because of the very broad variety of instances that can be involved. Moreover, in practice, IS can lead to a degeneracy problem when only a few of the sample paths $(\mathbf{X}_j^{(i)})_{0 \le j \le c}$ will have significant weights ω_i, and all the other sample paths will have very small weights. One common way to deal with degeneracy in particle filtering is resampling, but it is not adapted to rare event simulation. In the case of continuous state space when the value of c in the relation $c\Delta_t = S$ is not too large (e.g., say a few dozen), it is nevertheless possible (Johansen, 2006; Johansen, Del Moral, & Doucet, 2006) to determine potentially efficient sampling transition kernels using the sequential Monte Carlo sampler framework. Tuning these kernels requires the user to have a significant experience with the studied system. Two examples are provided in Johansen (2006). The IS algorithm is detailed in Algorithm 17.

ALGORITHM 17 Importance sampling algorithm with Markov processes for probability estimation

Input: The discrete time Markov process $(\mathbf{X}_j)_{0 \le j \le c}$ with initial distribution $\pi_0(\mathbf{x}_0)$ and transition kernels $\pi_j(\mathbf{x}_j | \mathbf{x}_{j-1})$, the IS auxiliary initial transition distribution $\tilde{\pi}_0(\mathbf{x}_0)$ and kernels $\tilde{\pi}_j(\mathbf{x}_j | \mathbf{x}_{j-1})$, the number of samples N, the function $\varphi(\cdot)$, the threshold T, and the stopping time c.

Output: The probability estimate $\hat{\mathbb{P}}^{\mathrm{IS}}$.

1 Sample $(\mathbf{X}_j^{(1)})_{0 \le j \le c}$, $(\mathbf{X}_j^{(2)})_{0 \le j \le c}$, $\ldots$, $(\mathbf{X}_j^{(N)})_{0 \le j \le c}$ with initial distribution $\tilde{\pi}_0(x_0)$ and transition kernels $\tilde{\pi}_j(\mathbf{x}_j | \mathbf{x}_{j-1})$.

2 Estimate $Z_i = \sup_{0 \le j \le c} \varphi(\mathbf{X}_j^{(i)})$.

3 Estimate $\hat{\mathbb{P}}^{\mathrm{IS}} = \frac{\sum_{i=1}^{N} \omega_i \mathbf{1}_{Z_i > T}}{\sum_{i=1}^{N} \omega_i}$ with $\omega_i = \frac{\pi_0(\mathbf{x}_0^{(i)}) \prod_{j=1}^{c} \pi_j(\mathbf{x}_j^{(i)} | \mathbf{x}_{j-1}^{(i)})}{\tilde{\pi}_0(\mathbf{x}_0^{(i)}) \prod_{j=1}^{c} \tilde{\pi}_j(\mathbf{x}_j^{(i)} | \mathbf{x}_{j-1}^{(i)})}$.

4 **return** $\hat{\mathbb{P}}^{\mathrm{IS}}$.

9.4.2 Application to a toy case

Brownian bridge

To apply IS on the Brownian bridge, the proposed auxiliary IS density W'_j at time j is a Gaussian distribution with mean b and variance 1. The considered model of IS auxiliary density is simple because b does not depend on j. We have *a priori* no idea of the potential efficiency of this parametric type of density on the Brownian bridge. If $b = 0$, then IS corresponds to the CMC case. The difficulty is that there is no rule on how to tune the parameter b. IS is applied to the Brownian bridge for different settings of b to evaluate the efficiency of this algorithm. The different results are presented in Table 9.3. IS enables to significantly decrease the variance of the estimation when b is well chosen. The choice of an efficient value for b is complex and is driven by variance minimization. This task could require a large number of samples. Similar to CMC, a part of the bias in the IS estimation results from the discretization of the Markov process. Ten IS sample paths for $b = 0.5$ are proposed in Figure 9.2.

9.4.3 Conclusion

IS is theoretically an interesting approach for estimating rare event probabilities on a continuous Markov process. Nevertheless, the practical determination of the $c + 1$ densities $\tilde{\pi}_j$, $j = 0, \ldots, c$ is an open issue. Because no guide to their tuning exists,

Table 9.3 Results of IS for the Brownian bridge

T	b	$\hat{\mathbb{P}}^{\text{IS}}$	RB ($\hat{\mathbb{P}}^{\text{IS}}$)	RE ($\hat{\mathbb{P}}^{\text{IS}}$)	**Simulation budget**	ν^{IS}
2	0.01	3.2×10^{-4}	−5%	112%	10^3	n/a
2	0.1	2.9×10^{-4}	−13%	8%	10^3	501
2	0.2	7.6×10^{-5}	−77%	125%	10^3	n/a
3	0.2	1.26×10^{-8}	−17%	10%	10^3	6×10^6

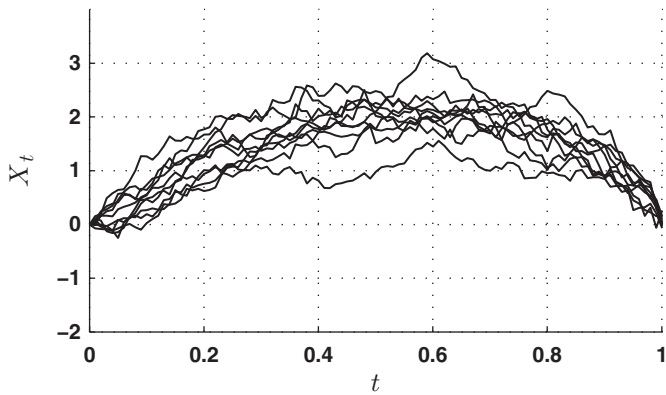

Figure 9.2 Some trials of IS Brownian bridge sample paths with $b = 0.5$.

Table 9.4 Summary of IS rare event probability estimation for Markov processes

Criteria	IS characteristics
Rarity of $\mathbb{P}(R)$	Increase the difficulty to find valuable $\tilde{\pi}_j, j = 0, \ldots, c$
Simulation budget	At least 10^3 samples
Type of process $\mathbf{X}_t$	No regularity condition on $\mathbf{X}_t$
Dimension of $\mathbf{X}_t$	Increase the difficulty to find valuable $\tilde{\pi}_j, j = 0, \ldots, c$
Probability estimate error	Estimated with retrials
Difficulty of implementation	Low to medium; difficult to find valuable $\tilde{\pi}_j, j = 0, \ldots, c$

the user must conduct a series of experiments in order to determine potential efficient auxiliary densities for the estimation of a given probability.

The main characteristics of IS are presented in Table 9.4. Thus, the application of IS on realistic test cases with Markov processes is more complex than on input–output functions but is nevertheless of interest for comparison.

9.5 Importance splitting

9.5.1 Principle

The idea of importance splitting for Markov processes is to decompose the sought probability in a product of conditional probabilities that can be estimated with a reasonable computation time. A similar approach has been proposed for static input–output function (see Section 5.4). Splitting methods on dynamical systems have been compared in Garvels & Kroese (1998) and L'Ecuyer, Demers, & Tuffin (2006). Some developments can also be found in Cérou, Del Moral, Le Gland, & Lezaud (2005) and Cérou, Del Moral, Le Gland, & Lezaud (2006). In these papers, a unified framework is described and rigorous proofs of convergence are given.

The splitting method is notably adapted when we consider processes that have continuous trajectories or at least right continuous and left limited trajectories. The principle of splitting is to determine a sequence of decreasing supersets of $\mathbf{B}$:

$$\mathbb{R}^d = \mathbf{B}_0 \supset \mathbf{B}_1 \supset \cdots \supset \mathbf{B}_{m-1} \supset \mathbf{B}_m = \mathbf{B}$$

Let us define $S_k = \inf\{t \geq 0 \text{ with } \mathbf{X}_t \in \mathbf{B}_k\}$ for $k = 1, \ldots, m$, and $S_{\mathbf{B}} = \inf\{t \geq 0, \text{ with } \mathbf{X}_t \in \mathbf{B}\}$. We consequently have $\mathbb{P}(S_{\mathbf{B}} \leq S) = \mathbb{P}(\mathcal{R})$. A Bayes' formula gives the following decomposition product:

$$\mathbb{P}(\mathcal{R}) = \prod_{k=1}^{m} \mathbb{P}(\mathbf{B}_k | \mathbf{B}_{k-1})$$

The previous equation is equivalent to

$$\mathbb{P}(S_{\mathbf{B}} \leq S) = \mathbb{P}(S_1 \leq S) \times \prod_{k=2}^{m} \mathbb{P}(S_k \leq S | S_{k-1} \leq S)$$

The importance splitting principle consists in estimating separately each probability $\mathbb{P}(S_k \leq S | S_{k-1} \leq S)$ with accuracy for a small sample size. The supersets of $\mathbf{B}$ are often defined by using the function $\varphi(\cdot)$ in the following way

$$\mathbf{B}_k = \{\mathbf{x}, \text{ with } \varphi(\mathbf{x}) \geq T_k\}, \quad k = 1, \ldots, m$$

for a suitable sequence of real numbers $T_0 < T_1 < \cdots < T_m = T$. The variable S_k and its associated probability can then be rewritten with

$$S_k = \inf\{t \geq 0, \text{ with } \varphi(\mathbf{X}_t) > T_k\}$$

and thus

$$\mathbb{P}(S_B \leq S) = \mathbb{P}\left(\max_{0 \leq t \leq S} \varphi(\mathbf{X}_t) > T\right)$$

which leads to

$$\mathbb{P}(S_k \leq S | S_{k-1} \leq S) = \mathbb{P}\left(\max_{0 \leq t \leq S} \varphi(\mathbf{X}_t) > T_k | \max_{0 \leq t \leq S} \varphi(\mathbf{X}_t) > T_{k-1}\right)$$

In practice, the goal of the splitting technique is to estimate the conditional probabilities $\mathbb{P}(S_k \leq S | S_{k-1} \leq S)$. They are chosen large enough to be estimated by Monte Carlo sampling with accuracy. In a first stage, N samples of Markov process $\mathbf{X}_t$ are generated until time $\min(S_1, S)$. If I_1 is the number of trajectories that have reached $\mathbf{B}_1$, then $\mathbb{P}(S_1 \leq S)$ is estimated by I_1/N. For stage $k \geq 1$, N starting points are randomly and uniformly chosen among the I_k crossing points between the subset $\mathbf{B}_k$ and the previously sampled trajectories. N paths of the process $\mathbf{X}_t$ are sampled from these crossing points according to the Markov dynamic of the process until time $\min(S_{k+1}, S)$. Among those trajectories, I_{k+1} reaches the set $\mathbf{B}_{k+1}$, and $\mathbb{P}(S_{k+1} \leq S | S_k \leq S)$ is estimated by I_{k+1}/N.

The main parameters that must be set in splitting are the thresholds $T_1, T_2, \ldots, T_{m-1}$. This choice is very important because it strongly influences the variance of the estimated probability. Indeed, if a threshold is misplaced, the probability of reaching it starting from the previous threshold could be very small and splitting becomes inefficient. Because there is no analytical formula that links the thresholds to conditional probabilities, it is often complicated to set the threshold values *a priori*. To circumvent this problem, the algorithm presented in Cérou & Guyader (2007)

proposes an adaptive choice of the thresholds. To estimate the thresholds, the authors consider quantile estimation in the same way as that can be done for static input–output function. For that purpose, a general adaptive importance splitting algorithm (GAISA) has been proposed from Cérou & Guyader (2007). It is decomposed in two stages. The first one estimates intermediate thresholds with $\tilde{N}$ trajectories, and the second one determines the starting points at level $\mathbf{B}_k$. We can indeed sample some new Markov processes from the entrance distribution at level $\mathbf{B}_k$ and determine their maxima before time S. If we define the threshold as T_{k+1} the $(1-\rho)$-quantile of these maxima, then we have:

$$\mathbb{P}(S_{k+1} \leq S | S_k \leq S) \approx \rho$$

The GAISA is detailed in Algorithm 21.

GAISA can also be used to estimate α-quantiles of the variable Z. The application principle is relatively similar to the one proposed with adaptive splitting in Section 5.4. Indeed, at each iteration k, instead of testing $T_k < T$, we compare the conditional probability product $\prod_{l=1}^{k} \hat{p}_l < \alpha$, and then the estimate $\hat{q}_\alpha^{\text{GAISA}}$ is the $\frac{\alpha}{\prod_{l=1}^{k} \hat{p}_l}$-quantile of the samples $\max_{S_k^{(j_i)} \leq t \leq S} \varphi(\mathbf{X}_t^{(i)})$, $i = 1, \ldots, N$ in the last GAISA iteration.

9.5.2 Application to a toy case

Brownian bridge

The GAISA is applied to the Brownian bridge. From our experience of the algorithm, the sample number $\tilde{N}$ is set to $\tilde{N} = 0.1 \times N$. There is indeed no need to use a high number of trajectories to estimate the intermediate thresholds. The choice of ρ is a trade-off between computation time and variance. Here the quantile parameter ρ is equal to 0.5 because it seems to give the best probability estimators. The different results are presented in Table 9.5. GAISA is a very efficient algorithm for this toy case with a low relative variance. In the same way as in the previous estimation methods, the bias in the GAISA estimation results from the discretization of the Markov process. Ten GAISA sample paths for $N = 10$ and $T = 3$ are given in Figure 9.3.

Table 9.5 Results of GAISA for the Brownian bridge

T	N	$\hat{\mathbb{P}}^{\text{GAISA}}$	$\text{RB}(\hat{\mathbb{P}}^{\text{GAISA}})$	$\text{RE}(\hat{\mathbb{P}}^{\text{GAISA}})$	**Simulation budget**	ν^{GAISA}
2	10^3	2.9×10^{-4}	-16	15	4440	31
2	10^4	2.9×10^{-4}	-13	6	4.5×10^4	21
3	10^3	1.2×10^{-8}	-20	39	8975	4.8×10^4
3	10^4	1.2×10^{-8}	-18	14	8.9×10^4	3.7×10^4

ALGORITHM 18 GAISA for probability estimation

Input: The continuous time Markov process $(\mathbf{X}_t)_{0\le t\le S}$ with initial distribution $\pi_0(\mathbf{x}_0)$ and transition kernels $\pi_{s,t}(\mathbf{x}, \mathbf{x}')$, the number $\tilde{N}$ of trajectories to estimate a threshold, the number N of trajectories to estimate crossing points, a constant $\rho \in (0, 1)$, the function $\varphi(\cdot)$, the threshold T, and the stopping time S.

Output: $\hat{\mathbb{P}}^{\text{GAISA}}$

Set $T_0 = 0$, $J_0 = \{1, 2, \ldots, N\}$.
for $i = 1, \ldots, N$ **do**
  Set $S_0^{(i)} = 0$ and sample $\mathbf{X}_0'^{(i)}$ independently from law π_0.
  Set $e_0^{(i)} = \mathbf{X}_0'^{(i)}$.
$k = 0$.
while $T_k < T$ **do**
  for $i = 1, \ldots, \tilde{N}$ **do**
    Choose randomly and uniformly a subscript $j_i \in J_k$.
    Sample a path $\mathbf{X}_t^{(i)}$ starting from $e_k^{(j_i)}$ at time $S_k^{(j_i)}$ and until final time S.
    Set $Z_i = \max_{S_k^{(j_i)} \le t \le S} \varphi(\mathbf{X}_t^{(i)})$.
  Estimate the threshold $T_{k+1} = Z_{([(1-\rho)\tilde{N}]+1)}$.
  for $i = 1, \ldots, N$ **do**
    Choose randomly and uniformly a subscript $j_i \in J_k$.
    Sample a path $\mathbf{X}_t'^{(i)}$ starting from $e_k^{(j_i)}$ at time $S_k^{(j_i)}$ and until time $\min(S_{k+1}^{(i)}, S)$ where $S_{k+1}^{(i)} = \inf\{S_k^{(j_i)} \le t \le S, \text{ with } \varphi(\mathbf{X}_t'^{(i)}) \ge T_{k+1}\}$.
    Set $e_{k+1}^{(i)} = \mathbf{X}'^{(i)}_{\min(S_{k+1}^{(i)}, S)}$.
  Set $J_{k+1} = \{i, (\min(S_{k+1}^{(i)}, S)) < S\}$ and $I_{k+1} = |J_{k+1}|$.
  Set $\hat{p}_{k+1} = \dfrac{I_{k+1}}{N}$.
  $k \leftarrow k + 1$.
Among the last sampled paths $t \mapsto \varphi(\mathbf{X}_t'^{(1)}), \ldots, t \mapsto \varphi(\mathbf{X}_t'^{(N)})$, a proportion r reaches S before final time T.
Estimate $\hat{\mathbb{P}}^{\text{GAISA}} = \prod_{l=1}^{k-1} \hat{p}_l \times r$.
return $\hat{\mathbb{P}}^{\text{GAISA}}$.

9.5.3 Conclusion

Importance splitting is a very efficient method for estimating rare event probabilities on continuous Markov chains with a high improvement of CMC performances. The tuning of its parameters does not significantly influence the estimator variance. The main characteristics of importance splitting are presented in Table 9.6. The use of splitting on realistic test cases is mandatory.

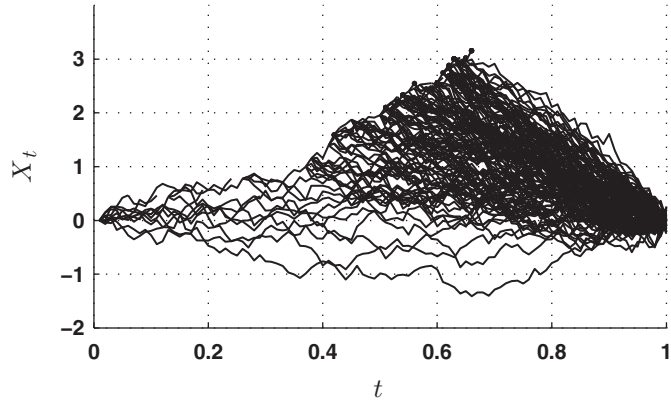

Figure 9.3 Set of GAISA sample trajectories on the Brownian bridge.

Table 9.6 Summary of GAISA rare event probability estimate for Markov processes

Criteria	GAISA characteristics
Rarity of $\mathbb{P}(R)$	Increase efficiency relatively to CMC
Simulation budget	At least 10^3 samples
Type of process $\mathbf{X}_t$	$\mathbf{X}_t$ must be right continuous with left limits
Dimension of $\mathbf{X}_t$	Still efficient on high dimensional systems
Probability estimate error	Estimated with retrials
Difficulty of implementation	Medium to high because of the algorithm complexity

9.6 Weighted importance resampling

9.6.1 Principle

IS algorithms could increase the occurrence of rare events, but in most cases, the studied dynamical system is so complicated that it is impossible to determine properly an efficient auxiliary distribution. An alternative to IS is weighted importance resampling (WIR) algorithm, also called *interacting particle system*, presented in Del Moral & Garnier (2005). It provides a method for rare event probability estimation in the form $\mathbb{P}(\mathbf{X}_c \in \mathbf{B}) = \mathbb{P}(\varphi(\mathbf{X}_c) > T)$, that is, for some events that occur at terminal time based on large deviation theory (LDT) considerations.

This formalism is not directly adapted to the estimation of $\mathbb{P}(\mathcal{R})$. Nevertheless, we can rewrite the estimation of the probability

$$\mathbb{P}\left(\varphi(\mathbf{X}_j) > T, \text{ for some } 0 \leq j \leq c\right) = \mathbb{P}\left(\max_{0\leq j\leq c} \varphi(\mathbf{X}_j) > T\right)$$

so that it fits the WIR framework. To apply WIR to the estimation of $\mathbb{P}(\mathcal{R})$, the idea is to introduce the process of maxima M_j of the Markov chain $\mathbf{X}_j$ before time j,

$$M_j = \max_{l=0,\ldots,j} \varphi(\mathbf{X}_l) = \max(M_{j-1}, \varphi(\mathbf{X}_j))$$

However, the sequence M_j is not a Markov one and, thus, it is necessary to consider the process $\mathbf{Y}_j$

$$\mathbf{Y}_j = (\mathbf{X}_j, M_j) \tag{9.2}$$

which is a Markovian sequence. It is straightforward to check why $\mathbf{Y}$ is a Markov chain. Indeed, one has

$$\mathbf{Y}_{j+1} = (\mathbf{X}_{j+1}, \max(M_j, \varphi(\mathbf{X}_{j+1})))$$

Because the conditional law of $\mathbf{X}_{j+1}$ given the past depends only on the present state $\mathbf{X}_j$, the preceding formulation shows that the conditional law of $\mathbf{Y}_{j+1}$ given the past depends only on the present state $\mathbf{Y}_j$. WIR can thus be applied without any restriction to estimate $\mathbb{P}(\mathcal{R})$ with Markov process $\mathbf{Y}_j$.

The WIR algorithm consists of a set of N paths $(\mathbf{X}_j^{(i)})_{1\leq j\leq c}$, $i = 1, \ldots, N$. The initial generation is a set of a N-sample $\mathbf{X}_0^{(1)}, \ldots, \mathbf{X}_0^{(N)}$ independently and identically sampled from the initial distribution of the chain π_0. The trajectories are updated from j to $j+1$ to give an advantage to the ones that can potentially reach the rare event $\mathbf{B}$. WIR is performed in two steps. First, the selection stage consists in choosing with replacement the trajectories according to an empirical weighted measure built with a function G_j that depends on $\mathbf{X}_{0:j}^{(1)}, \ldots, \mathbf{X}_{0:j}^{(N)}$. The notation $\mathbf{X}_{0:j}$ is used for the vector $(\mathbf{X}_0, \mathbf{X}_1, \ldots, \mathbf{X}_j)^T$. The function G_j must be strictly positive. Trajectories that are more likely to reach the rare set are multiplied, and the others are removed. Second, the mutation stage consists in applying the Markov transition kernel π_j to the trajectory evolution. The complete method is presented in Algorithm 12.

The WIR probability estimate is unbiased. The paper (Del Moral & Garnier, 2005) also proposes to use one of the two following weighted functions G_j:

$$G_j^{\beta}(\mathbf{X}_{0:j}) = \exp(\beta\varphi(\mathbf{X}_j))$$

and

$$G_j^{\alpha}(\mathbf{X}_{0:j}) = \exp(\alpha(\varphi(\mathbf{X}_j) - \varphi(\mathbf{X}_{j-1}))) \tag{9.3}$$

for some strictly positive parameters β and α. In Equation (9.3), we must set $\mathbf{X}_{-1} = \mathbf{x}_0$ where $\mathbf{x}_0$ is any point of $\mathbb{R}^d$. According to Del Moral & Garnier (2005), the use of weight function G_j^{α} gives better results. Also, using these functions does not require remembering the entire trajectories $\mathbf{X}_{0:j}^{(i)}$ but only $\mathbf{X}_j^{(i)}$ or $(\mathbf{X}_{j-1}^{(i)}, \mathbf{X}_j^{(i)})$ at each step.

WIR tuning is nevertheless needed when using the functions G_j^{β} or G_j^{α}. In general, for two different probability estimations on same Markov chains, Algorithm 12 is not

ALGORITHM 19 The WIR algorithm

Input: The discrete time Markov process $(\mathbf{X}_j)_{0\leq j\leq c}$ with initial distribution $\pi_0(\mathbf{x}_0)$ and transition kernels $\pi_j(\mathbf{x}_j|\mathbf{x}_{j-1})$, the sample size N, the strictly positive weight functions G_j, and the function $\varphi(\cdot)$

Output: $\hat{\mathbb{P}}^{\text{WIR}}$

for $i = 1, \ldots, N$ **do**
- Sample independently $\mathbf{X}_0^{(i)}$ from law π_0.
- Initialize the weights: $W_0^{(i)} = 1$.
- **for** $j = 0, \ldots, c-1$ **do**
  - Compute the normalizing constant $\eta_j = \frac{1}{N}\sum_{i=1}^{N} G_j(\mathbf{X}_{0:j}^{(i)})$.
  - Choose independently N paths according to the measure $\mu_j(\mathrm{d}\mathbf{x}_{0:j}) = \frac{1}{N\eta_j}\sum_{i=1}^{N} G_j(\mathbf{X}_{0:j}^{(i)})\delta_{\mathbf{X}_{0:j}^{(i)}}(\mathrm{d}\mathbf{x}_{0:j})$,
  - and rearrange the weights accordingly. The selected paths are noted $\tilde{\mathbf{X}}_{0:j}^{(i)}$, $i = 1, \ldots, N$ and the corresponding weights $\tilde{W}_j^{(i)}$.
- **for** $i = 1, \ldots, N$ **do**
  - The chain evolves under π_{j+1}: $\tilde{\mathbf{X}}_j^{(i)} \xrightarrow{\pi_{j+1}} \mathbf{X}_{j+1}^{(i)}$.
  - Set $\mathbf{X}_{0:j+1}^{(i)} = (\tilde{\mathbf{X}}_{0:j}^{(i)}, \mathbf{X}_{j+1}^{(i)})$.
  - Set $W_{j+1}^{(i)} = \tilde{W}_j^{(i)} \times \left(G_j(\tilde{\mathbf{X}}_{0:j})\right)^{-1}$.

Estimate $\hat{\mathbb{P}}^{WIR} = \prod_{j=0}^{c-1}\eta_j \times \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\varphi((\mathbf{X}_n^{(i)})>T)} \times \left(W_n^{(i)}\right)$.

efficient for the same values of α (or β). Although asymptotic variance expression can be computed (Del Moral & Garnier, 2005), optimal values of these parameters depend on the unknown probability to estimate.

With the WIR algorithm, only a fraction of paths reaches the rare set, whereas most of the trajectories sampled with efficient importance transition kernels get through it. As a result, the smaller the probability, the larger the number of needed trajectories. WIR main advantage is that the transition from X_j to X_{j+1} does not need to be changed. To our knowledge, WIR has not been derived to estimate quantiles.

9.6.2 Application to a toy case

Brownian bridge

WIR is applied to the Brownian bridge for two values of simulation budget and selection parameter α. Typical WIR sample paths are illustrated in Figure 9.4. In the same way as before, the discretization of the Markov process implies a bias in the WIR probability estimation. The different results are presented in Table 9.7. A fine tuning of α is needed to reduce the estimation variance (Morio, Jacquemart, Balesdent, &

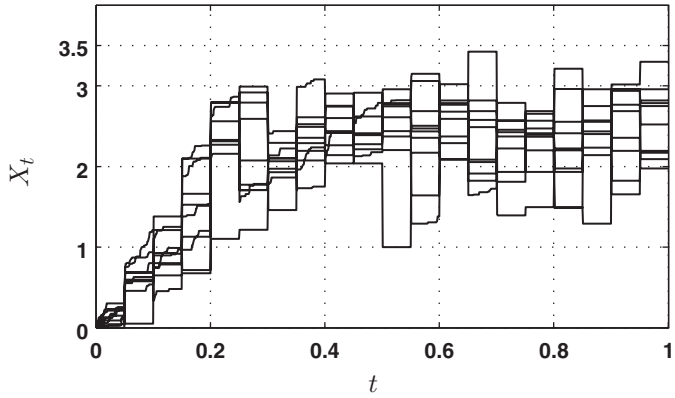

Figure 9.4 WIR typical sample trajectories on the Brownian bridge.

Table 9.7 Results of WIR for the Brownian bridge

T	α	$\hat{\mathbb{P}}^{\text{WIR}}$	RB($\hat{\mathbb{P}}^{\text{WIR}}$)	RE($\hat{\mathbb{P}}^{\text{WIR}}$)	**Simulation budget**	ν^{WIR}
2	6	2.9×10^{-4}	-14	33	10^3	27
2	6	2.9×10^{-4}	-14	10	10^4	30
2	10	2.9×10^{-4}	-16	76	10^3	5
2	10	2.9×10^{-4}	-14	31	10^4	3
3	6	1.1×10^{-8}	-26	234	10^3	1.2×10^{-4}
3	6	1.3×10^{-8}	-17	86	10^4	8.8×10^3
3	10	1.1×10^{-8}	-30	77	10^3	1.1×10^5
3	10	1.3×10^{-8}	-12	35	10^4	5.3×10^4

Marzat, 2013). The performances of WIR are relatively equivalent to those obtained with GAISA in the previous section.

9.6.3 Conclusion

WIR is a valuable alternative to GAISA for the estimation of rare event probabilities. There is no need to determine an auxiliary sampling density unlike IS but a selection parameter must be tuned. The main characteristics of WIR are presented in Table 9.8. The use of WIR on realistic test cases would be interesting.

9.7 Extreme value theory

9.7.1 Principle

The principle of extreme value theory (EVT) is described in Section 6.1 for input–output functions. Their application is the same for Markov processes. We assume that a set of N iid as $(\mathbf{X}_t)_{0\leq t\leq S}$ Markov processes $(\mathbf{X}_t^{(1)})_{0\leq t\leq S}$, $(\mathbf{X}_t^{(2)})_{0\leq t\leq S}$, ...,

Table 9.8 Summary of WIR rare event probability estimation for Markov processes

Criteria	WIR characteristics
Rarity of $\mathbb{P}(R)$	Increase efficiency relatively to CMC
Simulation budget	At least 10^3 samples
Type of process $\mathbf{X}_t$	$\mathbf{X}_t$ must be right continuous with left limits
Dimension of $\mathbf{X}_t$	Still efficient on high dimensional systems
Probability estimate error	Estimated with retrials
Difficulty of implementation	Medium to high because of the algorithm complexity

$(\mathbf{X}_t^{(N)})_{0\leq t\leq S}$ are available. If we consider the variable Z defined in Equation (9.1), we can then obtain the samples $Z_1, \ldots, Z_N$ such that

$$Z_i = \sup_{0\leq t\leq S} \varphi(\mathbf{X}_t^{(i)})$$

The peak over threshold (POT) approach can be then applied to the samples $Z_1, \ldots, Z_N$ in the same way as described in Section 6.1 for probability or quantile estimation. The whole procedure is given in Algorithm 20.

9.7.2 Application to a toy case

Brownian bridge

The POT approach is applied to the Brownian bridge case. The parameter $\hat{u}$ is estimated with mean excess plot, and the parameters $(\hat{\xi}, \hat{\beta}(\hat{u}))$ of the generalized Pareto distribution (GPD) are evaluated with maximum likelihood. The results obtained with POT are presented in Table 9.9. The conclusions are exactly the same as for input–output functions. CMC probability estimation is improved with the use of POT but we must be careful if it is applied on too rare events.

ALGORITHM 20 Peak over threshold technique for probability estimation with Markov processes

Input: The samples $(\mathbf{X}_t^{(1)})_{0\leq t\leq S}, (\mathbf{X}_t^{(2)})_{0\leq t\leq S}, \ldots, (\mathbf{X}_t^{(N)})_{0\leq t\leq S}$, the threshold T, and the stopping time S.

Output: The probability estimate $\hat{\mathbb{P}}^{\text{POT}}$.

1 Estimate $Z_i = \sup_{0\leq t\leq S} \varphi(\mathbf{X}_t^{(i)})$, $i = 1, \ldots, N$.

2 Estimate $\hat{u}$ with Hill plot or the mean excess plot from $Z_1, \ldots, Z_N$.

3 Estimate $(\hat{\xi}, \hat{\beta}(\hat{u}))$ by maximum likelihood with samples $\{Z_i, Z_i > \hat{u}\}$, $i = 1, \ldots, N$.

4 Estimate $\hat{\mathbb{P}}^{\text{POT}} = \left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{Z_i>\hat{u}}\right) \times \left(1 - H_{\xi,\beta(\hat{u})}(T - \hat{u})\right)$.

5 **return** $\hat{\mathbb{P}}^{\text{POT}}$

Table 9.9 Results of POT for the Brownian bridge

T	$\hat{\mathbb{P}}^{\text{POT}}$	RB ($\hat{\mathbb{P}}^{\text{POT}}$)	RE ($\hat{\mathbb{P}}^{\text{POT}}$)	**Simulation budget**	ν^{POT}
2	3.2×10^{-4}	−4%	45%	1000	15
3	4.7×10^{-5}	320,000%	160%	1000	n/a
2	3.2×10^{-4}	−5%	12%	10,000	21
3	1.2×10^{-6}	7900%	352%	10,000	n/a

9.7.3 Conclusion

POT is the only method that can be applied when only a fixed set of Markov processes is available. Nevertheless, the use of POT for probability estimation of too rare events is not advised even if this estimation is theoretically possible. The main characteristics of POT are presented in Table 9.10. The application of POT on realistic cases is advised only when resampling is impossible.

Table 9.10 Summary of POT rare event probability estimation for Markov processes

Criteria	**POT characteristics**
Rarity of $\mathbb{P}(R)$	Possible bias with a weak simulation budget
Simulation budget	Can be theoretically applied with any sample size
Type of process $\mathbf{X}_t$	No regularity condition on $\mathbf{X}_t$
Dimension of $\mathbf{X}_t$	No influence
Probability estimate error	Estimation with bootstrap samples
Difficulty of implementation	Low; approximate estimation of $\hat{u}$ with graphical methods

References

Borodin, A. N., and Salminen, P. (2002). *Handbook of Brownian motion—Facts and formulae*. Basel, Boston, Berlin: Birkhüser Verlag.

Cérou, F., Del Moral, P., Le Gland, F., and Lezaud, P. (2005). Limit theorems for the multilevel splitting algorithm in the simulation of rare events. In *Winter simulation conference proceedings* (pp. 682–691). Orlando, USA.

Cérou, F., Del Moral, P., Le Gland, F., and Lezaud, P. (2006). Genetic genealogical models in rare event analysis. *ALEA: Latin American Journal of Probability and Mathematical Statistics*, *1*, 1–19.

Cérou, F., and Guyader, A. (2007). Adaptive multilevel splitting for rare event analysis. *Stochastic Analysis and Applications*, *25*(2), 417–443.

Del Moral, P., and Garnier, J. (2005). Genealogical particle analysis of rare events. *The Annals of Applied Probability*, *15*(4), 2496–2534.

Garvels, M., and Kroese, D. (1998). A comparison of RESTART implementations. In *Winter simulation conference proceedings* (Vol. 1, pp. 601–608). Washington, USA.

Glynn, P., and Iglehart, D. (1989). Importance sampling for stochastic simulations. *Management Science*, *35*(11), 1367–1392.

Johansen, A. (2006). Some non-standard sequential Monte Carlo methods and their applications. Unpublished doctoral dissertation, University of Cambridge.

Johansen, A., Del Moral, P., and Doucet, A. (2006). Sequential Monte Carlo samplers for rare events. In *6th international workshop on rare event simulation* (pp. 256–267). October. Bamberg

Juneja, S., and Shahabuddin, P. (2001). Splitting-based importance-sampling algorithm for fast simulation of Markov reliability models with general repair-policies. *IEEE Transactions on Reliability*, *50*(3), 235–245.

Kroese, D. P., Taimre, T., and Botev, Z. I. (2011). *Handbook of Monte Carlo methods*. Hoboken, NJ: John Wiley and Sons, Inc.

L'Ecuyer, P., Demers, V., and Tuffin, B. (2006). Splitting for rare-event simulation. In *Winter simulation conference proceedings* (pp. 137–148). Monterey, USA.

Morio, J., Jacquemart, D., Balesdent, M., and Marzat, J. (2013). Optimisation of interacting particle systems for rare event estimation. *Computational Statistics & Data Analysis*, *66*, 117–128.

Robert, C., and Casella, G. (2005). Monte Carlo statistical methods. New York, NY: Springer.

Sandmann, W. (2005). On optimal importance sampling for discrete-time Markov chains. In *Second international conference on the quantitative evaluation of systems* (pp. 230–239).

Sobol, I. M. (1994). *A primer for the Monte Carlo method.* Boca Raton, FL: CRC Press.

This page intentionally left blank

Part Three

Benchmark of the Methods to Aerospace Problems

This page intentionally left blank

Estimation of launch vehicle stage fallout zone

10

L. Brevault, M. Balesdent, J. Morio

10.1 Principle

The estimation of launch vehicle fallout safety zone is a crucial problem in aerospace because a mistake in the estimation potentially involves dramatic repercussions on the population and the environment. Figure 10.1 illustrates some boosters of NASA Space shuttle floating in the Atlantic Ocean. For that purpose, an efficient estimation of the probability that a launch vehicle stage (or boosters) falls at a farther distance than a given safety limit is strategic for the qualification of such vehicles.

In this chapter, we consider a hypothetical small launch vehicle (∼150 tons) that lifts off from the European spaceport at Kourou (French Guyana) and aims to deliver a payload in a polar orbit (altitude of 700 km). This launch vehicle is composed of three solid propulsion stages. The first and second stages fall back into the Atlantic Ocean (Figure 10.2), and the third stage is launched into the same orbit as the payload. We consider in this chapter the fallback zone of the second stage. The second-stage separation occurs at an altitude of approximately 164 km and with a velocity of 3.7 km s^{-1}. Examples of impact points generated from 500 trials and associated joint pdf estimated with kde are illustrated in Figure 10.3.

10.2 Simulation description

This case can be modeled as an input–output function $\phi(\cdot)$ with the launch vehicle stage fallback phase initial conditions and several launch vehicle characteristics as inputs. The input–output function is the integration of the launch vehicle stage fallback trajectory. The output is the distance between the estimated launcher stage fallback position and the predicted one. A launch vehicle stage fallback trajectory simulator developed at Onera is used to estimate the impact points of the second stage.

The components of the input vector are:

- The stage separation altitude X_1 and velocity X_2 perturbations. The stage separation can also vary depending on weather conditions during the atmospheric flight.
- The flight path X_3 and azimuth X_4 angles at the stage separation (two inputs). These angles characterize the orientation of the stage with respect to the velocity vector at the stage separation and thus influence the ballistic fallback phase and consequently the impact position.

Estimation of Rare Event Probabilities in Complex Aerospace and Other Systems
http://dx.doi.org/10.1016/B978-0-08-100091-5.00010-1

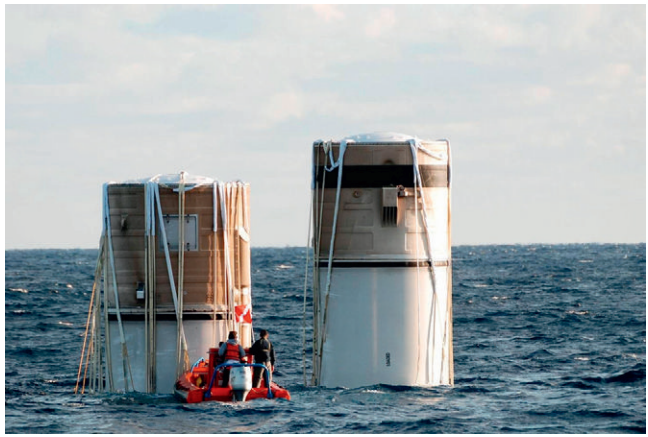

Figure 10.1 Boosters of NASA Space shuttle floating in the Atlantic Ocean © NASA.

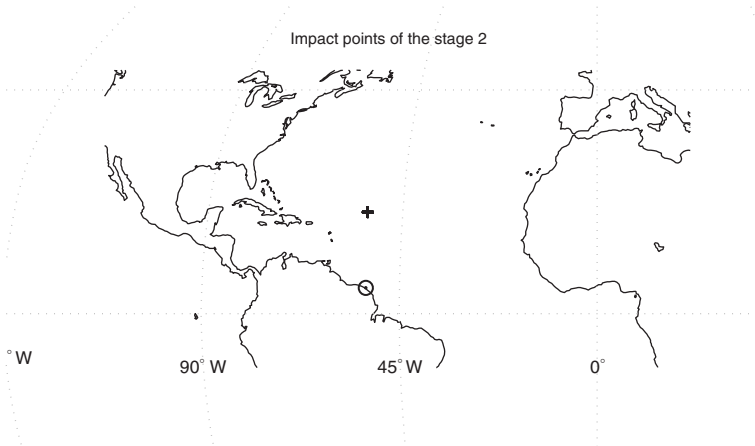

Figure 10.2 Launch vehicle lift off site (circle) and second-stage impact points (plus).

- The launch vehicle mass X_5. The mass of the different parts of the launch vehicle is also slightly random during the fallback because the propellant might not be totally burned during the powered flight.
- The weather variations X_6. These variations during the fallback can influence the impact position.

For the sake of clarity, we suppose here that all the random input variables are independent and follow standard Gaussian laws and that the input–output code includes the transformation that allows to switch from the standard space to the

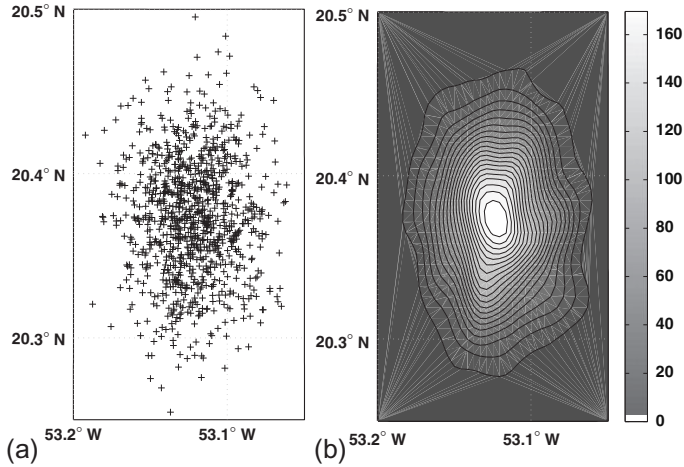

Figure 10.3 Second-stage impact points obtained from 500 random trials and associated joint pdf (obtained using kde). (a) Impact points. (b) Joint pdf.

physical space in which the launch vehicle mass, altitude, velocity, flight path angle, azimuth angle, and weather variations evolve.

10.3 Analysis of the input space

Let us consider a distance threshold of 12 km between the predicted and the simulated impact points. 10^6 samples have been generated with Crude Monte Carlo (CMC) and show that the target probability is on the order of 5×10^{-3} with a 1% relative error. The marginal distributions (obtained with histograms) of the input variables conditional to the threshold exceedance are given in Figure 10.4. As illustrated in this figure, the velocity and flight path angle distributions that lead to the considered event present two symmetric modes. Consequently, approaches based on the use of parametric pdf, such as cross entropy (CE), must be finely tuned in order to converge.

10.4 Estimation results

A huge CMC of 10^8 samples is performed with a threshold corresponding to 20 km (Figure 10.5). The probability estimate is 9.9×10^{-7} with a relative standard error of 10%. We will consider it as a reference probability for this case in the following discussion. The different available algorithms for rare event probability estimation are applied 50 times in this case in order to obtain statistics on their probability estimate. Their performances are analyzed in the following subsections.

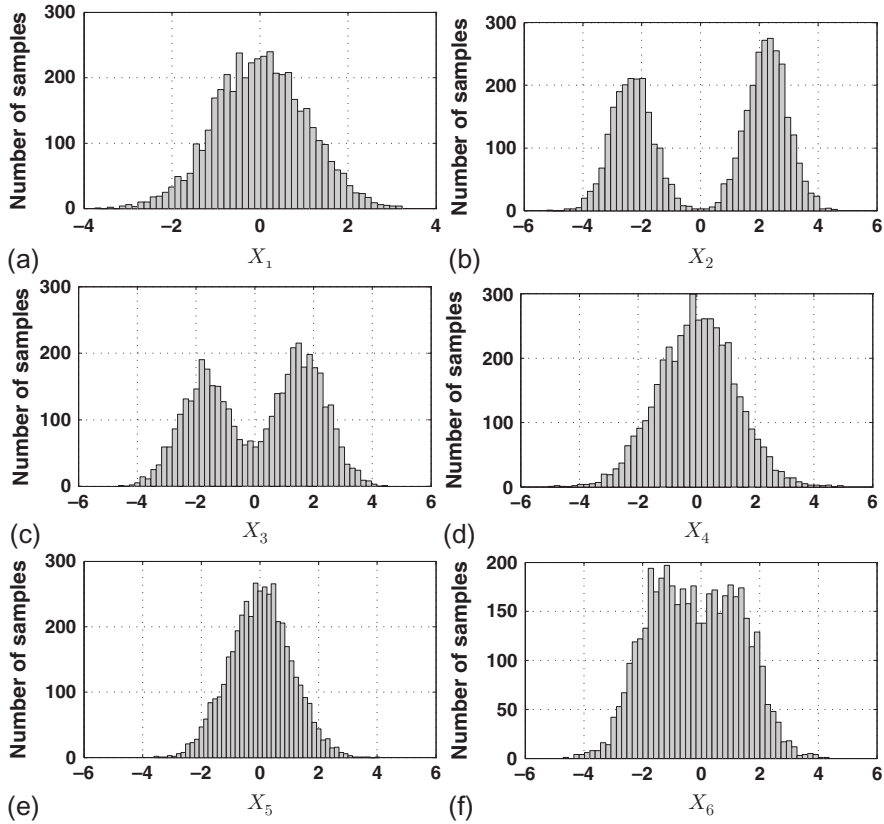

Figure 10.4 Input variable histograms (in the standard space) of samples that lead to a threshold exceedance (obtained with $T = 12$ km and CMC with 10^6 samples). (a) Altitude at stage separation: X_1. (b) Velocity at stage separation: X_2. (c) Flight path angle at stage separation: X_3. (d) Azimuth angle at stage separation: X_4. (e) Stage mass variation: X_5. (f) Weather variation: X_6.

10.4.1 Adaptive splitting technique

With the notations defined in Section 5.4, the adaptive splitting technique (AST) parameter tuning that gives the best results for this test case involves:

- Number of samples per iteration N: 1000.
- Value of the quantile parameter (ρ) used to define the intermediary thresholds: 0.7.
- Number of applications of the Markovian kernels during the sample generation step: 2.

The AST probability estimation results are given in Table 10.1. AST closely converges to the reference probability with a relative bias of 11% and a relative error of 32%, which is quite large for this application. However, the efficiency with respect to CMC is relatively important (398). To obtain the same results in terms of relative standard error with CMC, the simulation budget must be increased by a factor 398. The evolution of the generated output samples using AST is illustrated in Figure 10.6 for

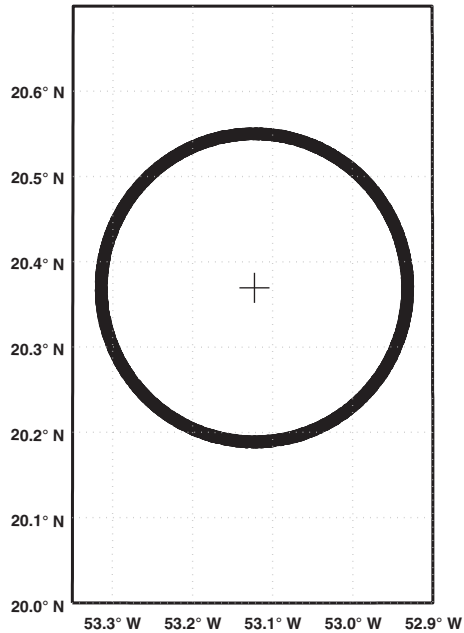

Figure 10.5 Predicted launch vehicle stage impact (+) and probability threshold.

Table 10.1 Results obtained with AST for the launcher stage fallout zone test case

$\hat{\mathbb{P}}^{\text{AST}}$	RB ($\hat{\mathbb{P}}^{\text{AST}}$)	RE ($\hat{\mathbb{P}}^{\text{AST}}$)	**Simulation budget**	ν^{AST}
1.1×10^{-6}	11%	32%	25,000	398

several iterations. In this test case and with the parameter tuning described previously, AST converges in 12 iterations.

10.4.2 Importance sampling

Because the failure space is totally unknown, only adaptive importance sampling techniques (i.e., CE and nonparametric adaptive importance sampling (NAIS)) are applied to the launch vehicle fallout test case.

10.4.2.1 Nonparametric adaptive importance sampling

The following parameters are used as parameter tunings of NAIS:

- Number of samples per iteration N: 1000.
- Value of the quantile parameter (ρ) used to define the intermediary thresholds: 0.8.
- Type of kernels used as the auxiliary pdf: Gaussian and Laplace.

NAIS is applied to the launch vehicle fallout zone test case with success (see Table 10.2). Indeed, with a very restricted budget of 5000 samples, this algorithm

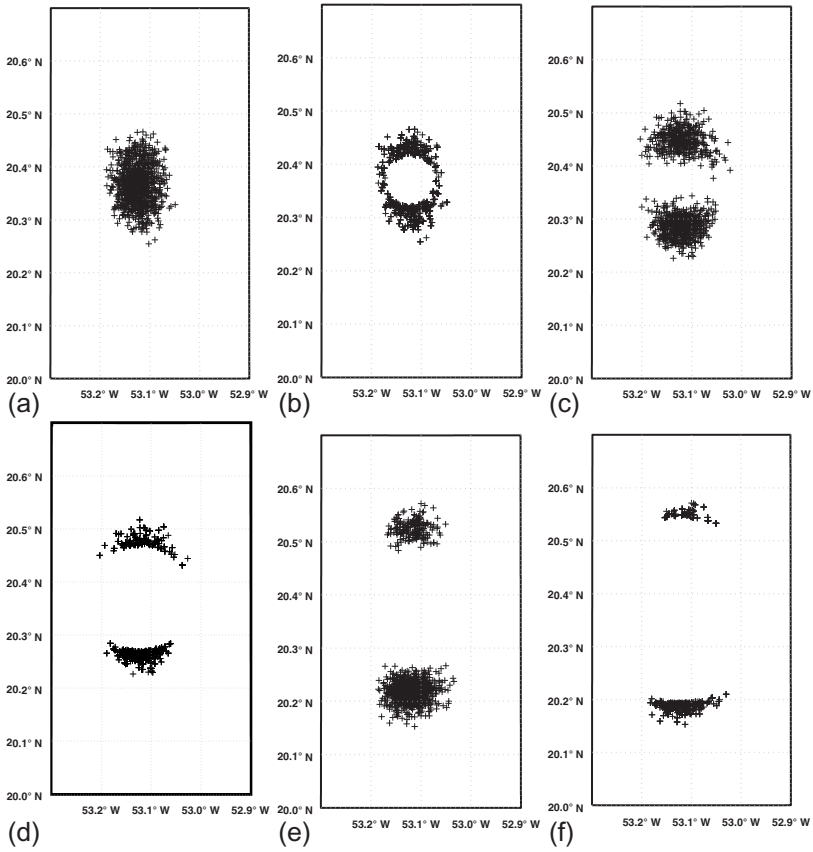

Figure 10.6 Impact positions corresponding to AST-generated samples. (a) Iteration 1—initial samples. (b) Iteration 1—selection of samples. (c) Iteration 4—resampling. (d) Iteration 4—selection of samples. (e) Last iteration—resampling. (f) Last iteration—selection of samples.

Table 10.2 Results obtained with NAIS for the launcher stage fallout zone test case

Kernel	$\hat{\mathbb{P}}^{\text{NAIS}}$	RB ($\hat{\mathbb{P}}^{\text{NAIS}}$)	RE ($\hat{\mathbb{P}}^{\text{NAIS}}$)	**Simulation budget**	ν^{NAIS}
Gaussian	9.95×10^{-7}	0.5%	19%	5000	5102
Laplace	1.06×10^{-6}	7%	27%	5000	2834

allows to estimate the probability of interest with very low relative bias (0.5% with Gaussian kernel and 7% with Laplace kernel) and relative error (19% with Gaussian kernel and 27% with Laplace kernels, respectively), which results in a very high efficiency relative to CMC (5102 with Gaussian kernels and 2834 with Laplace kernels, respectively). The efficiency of the NAIS estimator can be explained by

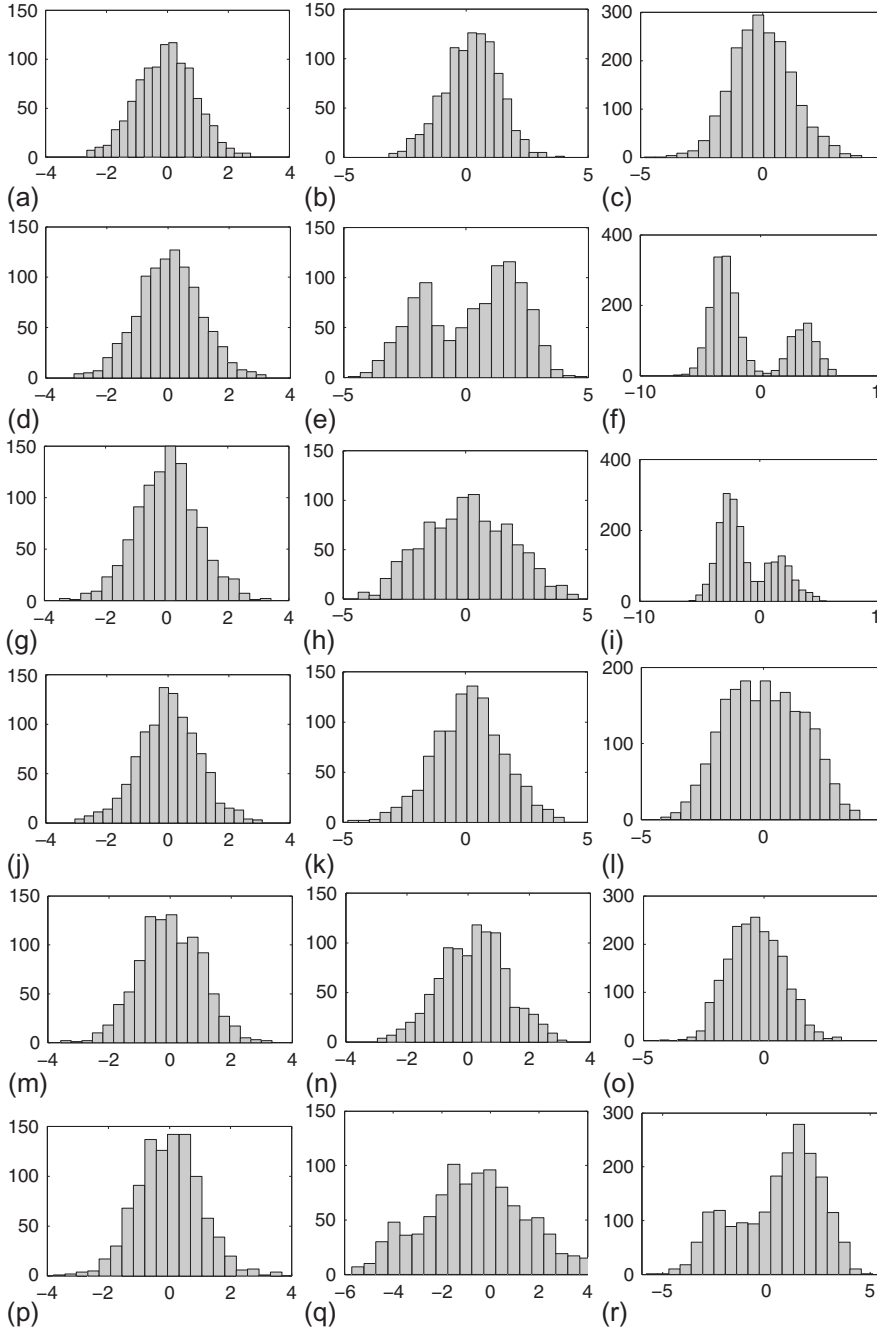

Figure 10.7 Histograms of NAIS-generated samples over the iterations (with Gaussian kernels). (a) X_1—iteration 1. (b) X_1—iteration 3. (c) X_1—iteration 5. (d) X_2—iteration 1. (e) X_2—iteration 3. (f) X_2—iteration 5. (g) X_3—iteration 1. (h) X_3—iteration 3. (i) X_3—iteration 5. (j) X_4—iteration 1. (k) X_4—iteration 3. (l) X_4—iteration 5. (m) X_5—iteration 1. (n) X_5—iteration 3. (o) X_5—iteration 5. (p) X_6—iteration 1. (q) X_6—iteration 3. (r) X_6—iteration 5.

the fact that the dimension of the input space is not very important (the dimension is only 6) and is compatible with the use of kernel-based auxiliary pdf. Moreover, because the optimal auxiliary pdf is multimodal (see Figure 10.7), NAIS estimates the probability with accuracy whereas classical CE fails.

10.4.2.2 Cross-entropy optimization

CE is applied to this test case with the following parameter tuning:

- Number of samples per iteration N : 1000.
- Value of the quantile parameter (ρ) used to define the intermediary thresholds : 0.95.
- Type of auxiliary pdf: Gaussian and Laplace. The center and bandwidth of the auxiliary pdf are optimized.

As illustrated in Figure 10.8, CE succeeds in catching only the most dominant mode of the optimal auxiliary pdf with Gaussian kernels. This is the result of the static parameterization of the auxiliary pdf with CE, which in this case is not compatible with multimodal distributions. This effect is less sensitive with Laplace distribution because its pdf tail is heavier than the tail of Gaussian pdf (Table 10.3). Thus, CE with Laplace pdf is able to successfully estimate a valuable probability, but its efficiency is still lower than NAIS efficiency.

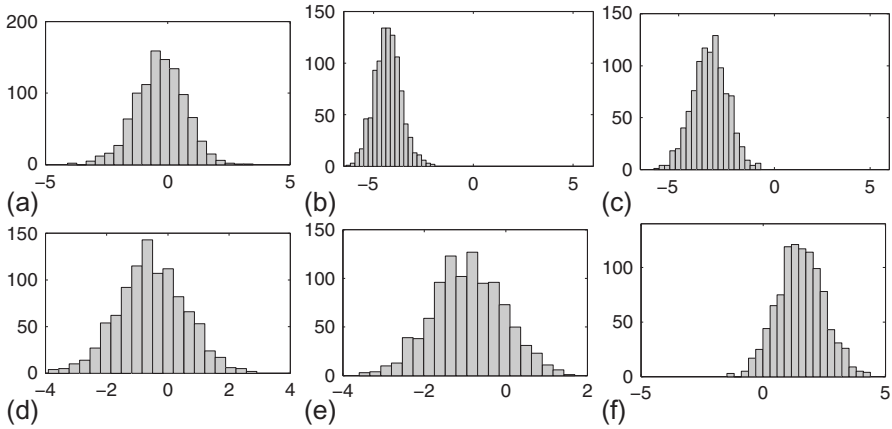

Figure 10.8 Histograms of CE-generated samples at the last iteration (with Gaussian auxiliary pdf). (a) X_1. (b) X_2. (c) X_3. (d) X_4. (e) X_5. (f) X_6.

Table 10.3 Results obtained with CE using Gaussian and Laplace auxiliary pdf of which the center and the bandwidth are optimized

Aux. pdf	$\hat{\mathbb{P}}^{\mathrm{CE}}$	RB ($\hat{\mathbb{P}}^{\mathrm{CE}}$)	RE ($\hat{\mathbb{P}}^{\mathrm{CE}}$)	**Simulation budget**	ν^{CE}
Gaussian	6.0×10^{-7}	−39%	54%	8000	423
Laplace	9.1×10^{-7}	−8%	59%	8000	363

10.4.3 Directional sampling

Directional sampling (DS) is applied to the launch vehicle stage fallout zone test case for different numbers of generated directions with the following tunings:

- Number of directions: 10, 100, 1000, and 10,000.
- Repartition of points on the unit hypersphere: random points uniformly distributed.

The results obtained with DS for different numbers of directions are given in Table 10.4. The efficiency of DS in terms of relative error particularly depends on the number of generated directions. Thus, a compromise has to be established between the simulation budget (i.e., number of directions) and the accuracy of the method. Nevertheless, for a sufficient number of directions (1000), the efficiency of DS is analogous to AST in terms of relative error and simulation budget.

Table 10.4 Results obtained with DS for the launcher stage fallout zone test case

Number of directions	$\hat{\mathbb{P}}^{\mathrm{DS}}$	RB ($\hat{\mathbb{P}}^{\mathrm{DS}}$)	RE ($\hat{\mathbb{P}}^{\mathrm{DS}}$)	**Simulation budget**	ν^{DS}
10	7.3×10^{-7}	−26%	399%	222	286
100	9.9×10^{-7}	0.4%	96%	2537	432
1000	9.4×10^{-7}	−6%	32%	26,761	373
1000	1.0×10^{-6}	4%	9%	282,650	438

10.4.4 Adaptive directional sampling

Adaptive directional sampling (ADS) is applied to this test case with the following parameter tuning:

- Same number of samples for all the ADS stages: $\gamma_1(N) = \gamma_2(N) = 0.5$.
- Number m of cones in the stratification of variable $\mathbf{A}$ is set to $2^d = 64$.

ADS is well adapted to this rare event probability estimation because it is able to catch the different multiple failure regions as illustrated in Table 10.5. The efficiency of ADS is of the same order as NAIS and is thus particularly accurate with a small simulation budget.

Table 10.5 Results obtained with ADS for the launcher stage fallout zone test case

$\hat{\mathbb{P}}^{\mathrm{ADS}}$	RB ($\hat{\mathbb{P}}^{\mathrm{ADS}}$)	RE ($\hat{\mathbb{P}}^{\mathrm{ADS}}$)	**Simulation budget**	ν^{ADS}
1.0×10^{-7}	7%	27%	3200	4300

10.4.5 FORM/SORM

The most probable point found by first-order reliability methods (FORM) and second-order reliability methods (SORM) is given in Table 10.6, and its impact position is illustrated in Figure 10.9. This point does not correspond to the maximum of importance sampling auxiliary pdf, which explains the relatively inaccurate results obtained by FORM and SORM. Because the failure space is not connected, FORM/SORM is not adapted to this case and does not succeed in estimating a valuable probability (Table 10.7).

Table 10.6 Coordinates of the design point

x_1	x_2	x_3	x_4	x_5	x_6
0.45	3.69	2.54	0.30	0.47	−2.34

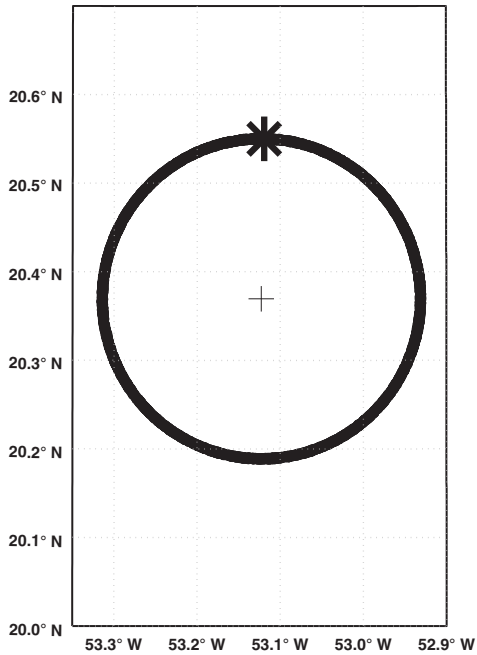

Figure 10.9 Impact position relative to the most probable point (∗) and probability threshold.

Table 10.7 Results obtained with FORM and SORM for the launcher stage fallout zone test case

$\hat{\mathbb{P}}^{\text{FORM}}$	RB ($\hat{\mathbb{P}}^{\text{FORM}}$)	**Simulation budget**
1.6×10^{-7}	-84%	137
$\hat{\mathbb{P}}^{\text{SORM}}$	RB ($\hat{\mathbb{P}}^{\text{SORM}}$)	**Simulation budget**
2.37×10^{-7}	-76%	267

10.4.6 Line sampling

The important direction of line sampling (LS) is determined through the design point given by FORM. Because FORM is not effective in this case, LS is also inadequate whatever the simulation budget (Table 10.8). LS enables to sample around only one of the failure modes and thus a biased probability estimation is obtained.

Table 10.8 Results obtained with LS for the launcher stage fallout zone test case

$\hat{\mathbb{P}}^{\text{LS}}$	RB ($\hat{\mathbb{P}}^{\text{LS}}$)	RE ($\hat{\mathbb{P}}^{\text{LS}}$)	**Simulation budget**	ν^{LS}
2.3×10^{-7}	-75%	19%	200	n/a
2.3×10^{-7}	-77%	6%	2000	n/a

10.5 Conclusion

The main characteristic of this use case that concerns a launcher fallout zone probability estimation is the multimodality of the failure region. Consequently, algorithms that may only be applied when there is only one main failure mode such as FORM/SORM or LS are thus not efficient. The most relevant algorithms for this test case are NAIS and ADS, which are particularly efficient because the input dimension is low. AST, DS, and CE succeed in estimating the rare event probability but with a lower efficiency than NAIS and ADS. Because there are several failure modes, CE with unimodal densities is not completely adapted. This also explains why ADS performs better than DS. The performance of AST can be considered disappointing relative to the other methods. This is principally caused by the low dimension of this test case and by the relatively high required simulation budget to obtain an AST probability estimation.

This page intentionally left blank

Estimation of collision probability between space debris and satellites

11

J. Morio, J.C. Dolado Pérez, C. Vergé, R. Pastel, M. Balesdent

11.1 Principle

On February 10th, 2009, active commercial satellite Iridium-33 and out-of-order Russian satellite Cosmos-2251 collided (Kelso, 2009). The impact produced more than 2000 trackable debris. Most debris can destroy any satellite whether in use or not that it might encounter. Space debris can in general be from rocket stages, nonoperational satellites, and fragments from explosions or collisions as illustrated in Figure 11.1. The safest practice for satellites that encounter space debris is to avoid collision. Avoidance maneuvers are efficient in reducing the collision probability between two orbiting objects; nevertheless, they consume fuel, reducing the operational lifetime of the satellite and disturb its operational mission. Consequently, teams responsible for satellite safety must determine a satellite's operational mission before determining actual collision avoidance maneuver and try to combine, whenever possible, planned station-keeping maneuvers with collision avoidance maneuvers. Avoidance maneuvers are based on the estimated collision probability among other parameters.

In this test case, we consider two spatial objects (a debris object and a satellite) orbiting around an Earth-centered inertial reference frame. The orbital motion of the spatial objects is simulated using a simplified deterministic dynamical model that can be modeled as an input–output function. Their geometry is assumed spherical (i.e., the objects have a high tumbling motion when compared with their orbital period) and we assume that we perfectly know the radius of the sphere and the mass of the objects. We wonder about the relative position of the satellite and the debris and ask whether the distance between the two objects could be smaller than a conflict distance T during the given time span I.

11.2 Simulation description

To model the orbital motion of both space objects, we consider a general perturbation approach in which the original equation of motion is replaced with an analytical approximation that captures the essential character of the motion over some limited time interval, which also enables the determination of analytical integration of the

Estimation of Rare Event Probabilities in Complex Aerospace and Other Systems
http://dx.doi.org/10.1016/B978-0-08-100091-5.00011-3

Figure 11.1 Rocket body explosions © ESA.

equations. SGP4 model (Miura, 2009) is used to propagate the trajectories of debris and satellite according to the time. At time t, the spatial objects will be represented by their six-dimensional state vectors $\vec{s}_1(t)$ and $\vec{s}_2(t)$, that is, their three-dimensional position vectors $\vec{r}_1(t)$ and $\vec{r}_2(t)$ and their three-dimensional speed vectors $\vec{v}_1(t)$ and $\vec{v}_2(t)$ such that $\vec{s}_i = (\vec{r}_i, \vec{v}_i)$. The initial conditions for our example are defined in terms of two-line elements (TLE), similar to those provided by North American Aerospace Defense Command (NORAD), as the SGP4 model used for the orbital propagation of the considered objects. The initial condition value is denoted $\vec{s}_i^m$ at a given time t_i^m. SGP4 is used to propagate the orbit of both space objects through time, denoted by a scalar continuous function ν such that

$$\forall i \in \{1,2\}, \quad \forall t \in I, \quad \vec{s}_i(t) = \nu(\vec{s}_i^m, t_i^m, t),$$
$$\delta = \min_{t \in I}\{\|\vec{r}_2 - \vec{r}_1\|(t)\}$$

The function of time $t \in I \mapsto \|\vec{r}_2 - \vec{r}_1\|(t)$ makes δ available through numerical optimization in a deterministic approach. Figure 11.2 presents the corresponding debris and satellite trajectories in an inertial Earth-centered reference frame and their relative distance as a function of time. In fact, the position and velocity of space objects are estimated from more or less imprecise measurements. Whereas the measurement means used for satellites (e.g., GPS, laser) result in a reasonable orbital accuracy (e.g., several tens of meters), the measurement means used for debris and uncooperative space objects (e.g., mainly radar and telescopes) could result in quite imprecise orbits (e.g., several hundred of meters or few kilometers). This lack of accuracy depends on a high number of factors. TLE sum up this information and feed

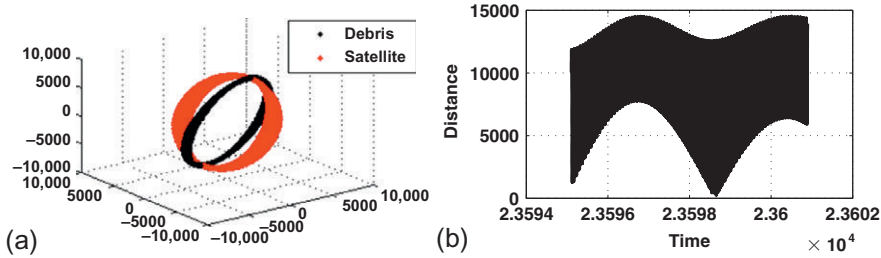

Figure 11.2 Debris and satellite trajectories and their relative distance (in km) in function of time (in days). (a) Trajectories (scale in km). (b) Relative distance.

the models with the couple $(\vec{s}_i^m, t_i^m)$ for $i = 1, 2$, but to cope with their uncertainty, we have added iid Gaussian noises to the model inputs $\vec{s}_i^m$.

This case can be modeled as an input–output function where:

- The input $\mathbf{X}$ is a 12-dimensional standard multivariate normal random vector that corresponds to the satellite and debris measurement errors on their position and speed. For simplicity, the drag coefficient and the atmospheric effects are considered as perfectly known; thus, we will not consider these parameters as uncertain.
- The input–output function $\phi(\cdot)$ enables to propagate the debris and satellite trajectories with the SGP4 model during I. The input–output code includes the transformation that allows to switch from the standard space of the input to the physical space in which the satellite and debris position and speed evolve.
- The output Y is the minimum distance between the debris and the satellite during I.

In this test case, the quantity of interest is the probability that $\mathbb{P}(\phi(\mathbf{X}) < T) = \mathbb{P}(Y < T)$.

11.3 Analysis of the input space

Let us consider a distance threshold of $T = 50\,\text{m}$. 10^6 samples have been generated with Crude Monte Carlo (CMC) and show that the target probability is in the order of 7×10^{-3} with a 2% relative error. The marginal distributions (obtained with histograms) of the input variables conditional to the threshold nonexceedance are given in Figure 11.3. Except for X_1, the marginal distributions that lead to the rare event strongly correspond to the initial marginal distributions. The different correlations between the input components are the most significant contributors to the reach of the rare event due to high nonlinearity of the function $\phi(\cdot)$. This feature makes the considered rare event probability difficult to estimate in this case.

11.4 Estimation results

A CMC estimation with 10^7 samples has been performed and enables to estimate $\mathbb{P}(Y < T)$ with $T = 10$ m. The CMC probability estimate is 4.2×10^{-5} with a 5%

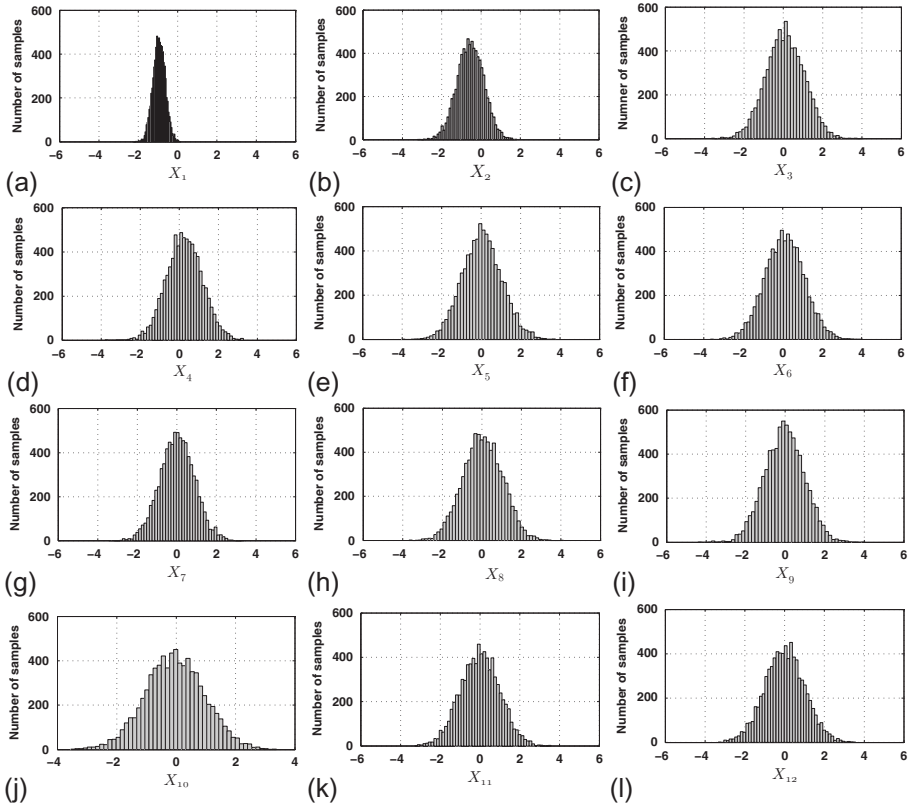

Figure 11.3 Input variable histograms (in the standard space) of samples that lead to a threshold nonexceedance (obtained with $T = 50$ m and CMC with 10^6 samples). (a) X_1. (b) X_2. (c) X_3. (d) X_4. (e) X_5. (f) X_6. (g) X_7. (h) X_8. (i) X_9. (j) X_{10}. (k) X_{11}. (l) X_{12}.

relative error. This value corresponds to a reference probability for the estimation of the different algorithm bias. All the different available algorithms for rare event probability estimation are applied 50 times in order to obtain statistics on their probability estimate. Their performances are analyzed in the following subsections.

11.4.1 Adaptive splitting technique

With the notations defined in Section 5.4, the adaptive splitting technique (AST) parameter tuning that give the best results for this test case are:

- Number of samples per iteration N: 1000.
- Value of the quantile parameter (ρ) used to define the intermediary thresholds: 0.7.
- Number of applications of the Markovian kernels during the sample generation step: 2.

The AST probability estimation results are given in Table 11.1. AST is applied to the debris-satellite collision case with success. This algorithm allows to estimate the

Table 11.1 Results obtained with AST for the debris-satellite collision case

$\hat{\mathbb{P}}^{\text{AST}}$	RB ($\hat{\mathbb{P}}^{\text{AST}}$)	RE ($\hat{\mathbb{P}}^{\text{AST}}$)	**Simulation budget**	ν^{AST}
5.2×10^{-5}	20%	22%	19,000	25

probability of interest with a valuable efficiency. AST converges in eight iterations to the target probability with a relative bias of 20% and a relative error of 22%.

11.4.2 Importance sampling

Only adaptive importance sampling techniques (i.e., CE and nonparametric adaptive importance sampling (NAIS)) can be applied to this case because the failure space is not known *a priori*.

11.4.2.1 Nonparametric adaptive importance sampling

Whatever the parameter tuning used for NAIS, this algorithm has not been able to converge in 100 iterations with a simulation budget of 10^5 samples. The dimension of the input space is too large for a potentially efficient application of NAIS in this case.

11.4.2.2 Cross-entropy optimization

Cross entropy (CE) is applied to this debris-satellite case with the following parameter tuning:

- Number of samples per iteration N : 1000.
- Value of the quantile parameter (ρ) used to define the intermediary thresholds: 0.95.
- Types of auxiliary pdf: Gaussian and Laplace. The center and bandwidth of the auxiliary pdf are optimized.

The probability estimates obtained with CE are indicated in Table 11.2. CE converges but estimates the target probability with a very significant bias. The strong nonconnectivity of the failure space explains why CE is not adapted to this case and is not able to capture its different modes.

Table 11.2 Results obtained with CE using Gaussian and Laplace auxiliary pdfs of which the center and the bandwidth are optimized

Aux. pdf	$\hat{\mathbb{P}}^{\text{CE}}$	RB ($\hat{\mathbb{P}}^{\text{CE}}$)	RE ($\hat{\mathbb{P}}^{\text{CE}}$)	**Simulation budget**	ν^{CE}
Gaussian	2.5×10^{-10}	-100%	203%	4000	n/a
Laplace	6.2×10^{-9}	-99%	183%	8000	n/a

11.4.3 Directional sampling

Directional sampling (DS) is applied to the debris-satellite case for different numbers of generated directions, with the following parameters:

- Number of directions: 10, 100, 1000, and 10,000.
- Repartition of points on the unit hypersphere: random points uniformly distributed.

Table 11.3 presents the different results obtained with DS. The DS probability estimates are biased positively because the assumptions used in DS are not valid. With the notations of Section 7.3, we notice that $1 - F_{R^2}(r_j^2)$ is not a good estimate of $\mathbb{P}(\phi(R\mathbf{A}) < T|\mathbf{A} = \mathbf{A}_j)$ because there exists $r'_j > r_j$ such that $\mathbf{1}_{\phi(r'_j\mathbf{A}_j)<T} = 0$. The lower the difference $r'j - r_j$, the worse is the approximation $1 - F_{R^2}(r_j^2)$.

Table 11.3 Results obtained with DS for the debris-satellite case

Number of directions	$\hat{\mathbb{P}}^{\mathrm{DS}}$	RB ($\hat{\mathbb{P}}^{\mathrm{DS}}$)	RE ($\hat{\mathbb{P}}^{\mathrm{DS}}$)	Simulation budget	ν^{DS}
10	8.9×10^{-3}	20,977%	309%	234	n/a
100	4.5×10^{-3}	10,635%	106%	2422	n/a
1000	5.2×10^{-3}	12,387%	30%	23,451	n/a
10,000	4.9×10^{-3}	12,434%	14%	282,650	n/a

11.4.4 Adaptive directional sampling

Adaptive directional sampling (ADS) requires a simulation budget that is too important to be applied to this debris-satellite case. Indeed, the number m of cones in the stratification of variable $\mathbf{A}$ is set to $2^{12} = 4096$. With $N = 20$ samples per cone, ADS needs 163,840 samples for one probability estimation. The dimension of the input is too important for an efficient application of ADS. Moreover, in the same way as DS, the probability approximation with the chi-squared distribution is not valid.

11.4.5 FORM/SORM

First-order reliability method (FORM) and second-order reliability method (SORM) do not succeed to estimate a valuable debris-satellite collision probability. Indeed, the

Table 11.4 Results obtained with FORM and SORM for the debris-satellite case

$\hat{\mathbb{P}}^{\mathrm{FORM}}$	RB ($\hat{\mathbb{P}}^{\mathrm{FORM}}$)	Simulation budget
0.13	285,330%	186
$\hat{\mathbb{P}}^{\mathrm{SORM}}$	RB ($\hat{\mathbb{P}}^{\mathrm{SORM}}$)	Simulation budget
6.32×10^{-3}	13,595%	90

FORM/SORM estimates are positively biased as summarized in Table 11.4, because the failure space is disconnected and its limit state surface is strongly nonlinear.

11.4.6 Line sampling

In the same way as FORM/SORM, LS is not efficient to estimate this debris-satellite collision probability (see Table 11.5). The important direction of LS is determined through the design point given by FORM, which is not effective on this case. Moreover, the approximation of Equation (7.5) is not true since there is not one unique intersection between the input failure region and the chosen sampling direction.

Table 11.5 Results obtained with LS for the debris-satellite case

$\hat{\mathbb{P}}^{\mathrm{LS}}$	RB ($\hat{\mathbb{P}}^{\mathrm{LS}}$)	RE ($\hat{\mathbb{P}}^{\mathrm{LS}}$)	**Simulation budget**	ν^{LS}
3.9×10^{-3}	9171%	25%	200	n/a
3.7×10^{-3}	8612%	6%	2000	n/a

11.5 Conclusion

AST is the only algorithm that is able to estimate with accuracy the target probability of this debris-satellite case, which is a very complex because the failure space is disconnected and relatively high-dimensional. Moreover, the reach of the rare event results mainly from specific combinations of input components. Most of the approximations made in rare event probability estimation techniques such as DS, ADS, FORM/SORM, and LS are not valid, and these algorithms result in biased estimates. Moreover, finding an efficient IS auxiliary distribution is also too complex with NAIS or CE because of the input dimension and the disconnected failure space. AST does not require any hypothesis and is well adapted to the cases where $d > 10$. The efficiency of AST compared to CMC is medium (about 25 for the chosen probability estimate), but this latter will increase if the target probability decreases.

References

Kelso, T. (2009). Analysis of the Iridium 33-Cosmos 2251 collision. In S. Ryan, The Maui Economic Development Board (Ed.), *Proceedings of the advanced Maui optical and space surveillance technologies conference* (p. E3). Wailea, USA.

Miura, N. Z. (2009). Comparison and design of Simplified General Perturbation Models (SGP4) and code for NASA Johnson Space Center, Orbital Debris Program Office. Unpublished doctoral dissertation, Faculty of California Polytechnic State University.

This page intentionally left blank

Analysis of extreme aircraft wake vortex circulations

12

J. Morio, I. De Visscher, M. Duponcheel, G. Winckelmans, D. Jacquemart, M. Balesdent

12.1 Principle

As a consequence of its take-off, an aircraft generates a complex turbulent wake emanating from the wing and horizontal tail plane. This wake rolls up to form a pair of counter-rotating vortices in the far field (Figure 12.1) and lasting for several minutes after the aircraft has flown by an object. Those vortices whose initial circulation and lateral spacing depend on the aircraft characteristics are transported and decay depending on the environmental conditions (wind, turbulence, temperature stratification, and ground proximity). The initial circulation of the vortices, Γ_0, is related to the aircraft weight, flight speed, wing span, and wing loading through $\Gamma_0 = \frac{W}{\rho V s b}$ with W the aircraft weight, V its flight speed, b its span, ρ the air density, and s the spacing factor defined as the ratio between vortex lateral spacing and aircraft span, which depends on the wing loading. A following aircraft's encounter with such a vortex pair can be hazardous because of the induced rolling moment and downwash velocity. The related wake hazard is at the origin of the separations to be applied between a landing or departing aircraft. Nevertheless, the increased air traffic in recent years makes these separation standards a limiting factor for the capacity of the busiest airports. A solution to airport congestion is thus to reduce some of the separation minima while at least maintaining the current safety level. For that purpose, a deeper understanding of the behavior of wake vortices and their interactions with a following aircraft is needed. This requires characterizing the potential encounter of wake vortex circulation, including the extreme cases with very low decay, as leading to the highest possible circulation at the time of encounter.

12.2 Simulation description

A wake vortex prediction tool, called the *deterministic wake vortex model* (DVM), has been developed at Université Catholique de Louvain (UCL) to predict the behavior of aircraft wake vortices. DVM uses several simplified physics-based models (described in De Visscher, Bricteux, & Winckelmans, 2013; De Visscher, Lonfils, & Winckelmans, 2013; De Visscher et al., 2010) to forecast in real time the transport and circulation decay of wake vortices depending on the aircraft that generates them and on the environmental conditions (meteorological conditions and ground

Estimation of Rare Event Probabilities in Complex Aerospace and Other Systems
http://dx.doi.org/10.1016/B978-0-08-100091-5.00012-5

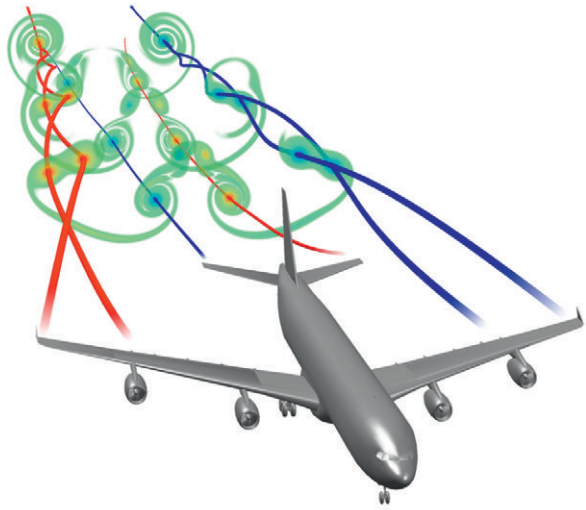

Figure 12.1 Illustration of the wake roll-up behind an aircraft (Simulation by G. Daeninck, UCL).

proximity). From a more practical point of view, DVM consists of a scalar input–output function. The uncertain input parameters, considered for this study, are the aircraft weight, the vortex spacing factor, and the lateral wind. The simulation code also considers other parameters for the aircraft's characteristics (wingspan, flight angles, etc.) and meteorological conditions (e.g., headwind, atmospheric turbulence, thermal stratification), but they are assumed to be set in the considered scenario. It is to be noted that the dimensionless coefficients of the simplified physical models used in DVM can also be varied, but those variations are not considered here. The output analyzed here is the total wake vortex circulation that characterizes the vortex strength at a given time after the aircraft has flown by.

In the chosen scenario, the aircraft has the following characteristics:

- Its wingspan is 79.75 m.
- Its weight is uncertain and follows a uniform law on the interval $U_{[0.75\times\text{MLW},0.95\times\text{MLW}]}$, with maximum landing weight (MLW); the aircraft MLW is set to 386,000 kg.
- The aircraft flies at an altitude of 80 m, thus close to ground.
- The aircraft flight speed is 70 m s^{-1}.

The spacing factor of the wake vortex follows a uniform law on $U_{[0.7,0.9]}$. Only the crosswind is considered. Its speed at 10 m high is described by a normal density of mean value 1.75 m s^{-1} and standard deviation 0.3 m s^{-1}. The whole wind profile follows a logarithmic layer wind profile ("Manual on Low-level Wind Shear", 2005).

In this chapter, we assume that only some crude Monte Carlo (CMC) samples of the total wake vortex circulation in a given scenario are available and no new CMC sample

can be then obtained. The only possible method to obtain a probability estimate of a high threshold exceedance in this case is the extreme value theory (EVT). It is indeed able to cope with a fixed set of samples.

12.3 Estimation results

A set of 10,000 iid samples $Y_1, Y_2, \ldots, Y_N$ of the total wake vortex circulation is available. The sample set corresponds to CMC samples of the left wake vortex circulation 60 s after its generation. The resulting histogram is presented in Figure 12.2. From this set of 10,000 samples, we have used a bootstrap procedure, 100 sets of $N = 100, 1000, 5000$ samples. The EVT—and, more precisely, the peak over threshold (POT) approach—is applied to these sets of samples. The parameter $\hat{u}$ of the generalized Pareto distribution (GPD) is estimated using the mean excess plot that is defined by $(u, \mathbb{E}(Y - u | Y > u))$. Figure 12.3 presents the mean excess plot on the samples that lead to the threshold $\hat{u} = 655\ \mathrm{m^2\,s^{-1}}$. The parameters $(\hat{\xi}, \hat{\beta}(\hat{u}))$ of the GPD are then evaluated with maximum likelihood. The different results obtained with POT are presented in Table 12.1.

Because the different probabilities that are considered are not known, their true value is replaced by their estimate obtained with the 10,000 samples to characterize RE ($\hat{\mathbb{P}}^{\mathrm{POT}}$) and RB ($\hat{\mathbb{P}}^{\mathrm{POT}}$). Then the relative deviation RE ($\hat{\mathbb{P}}^{\mathrm{POT}}$) and bias RB ($\hat{\mathbb{P}}^{\mathrm{POT}}$) of the probability are estimated using the bootstrap procedure. This approach is also applied for the CMC probability estimation results.

Different sample sizes are simulated in order to analyze the size influence on the accuracy of the probability estimate. For comparison, CMC probability estimates are also given in Table 12.2. EVT enables to decrease the relative error of CMC

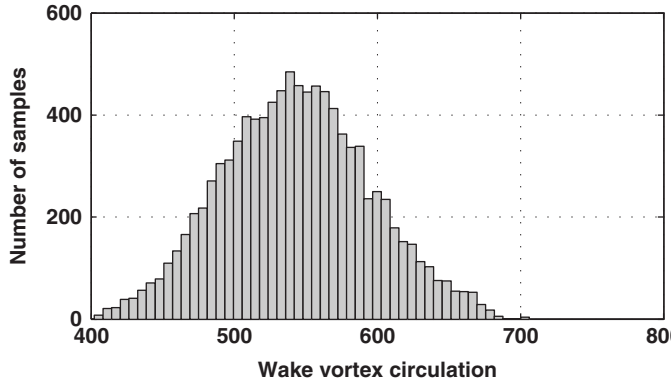

Figure 12.2 Histogram of CMC samples $Y_1, Y_2, \ldots, Y_N$. The wake vortex circulation is in $\mathrm{m^2\,s^{-1}}$.

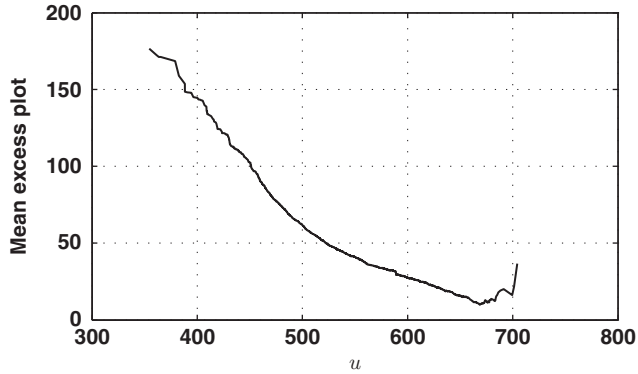

Figure 12.3 Mean excess plot of samples $Y_1, Y_2, \ldots, Y_N$. u is in $\mathrm{m}^2\,\mathrm{s}^{-1}$.

Table 12.1 Results of POT for the wake vortex test case

T	$\hat{\mathbb{P}}^{\mathrm{POT}}$	RB ($\hat{\mathbb{P}}^{\mathrm{POT}}$)	RE ($\hat{\mathbb{P}}^{\mathrm{POT}}$)	**Simulation budget**	ν^{POT}
650	2.9×10^{-2}	7%	18%	100	10.3
675	9.4×10^{-3}	135%	81%	100	n/a
700	8.7×10^{-4}	210%	450%	100	n/a
725	8.9×10^{-4}	2019%	603%	100	n/a
650	2.8×10^{-2}	8%	9%	1000	4.1
675	5.3×10^{-3}	33%	35%	1000	1.9
700	3.2×10^{-4}	16%	138%	1000	1.7
725	2.1×10^{-4}	402%	434%	1000	n/a
650	2.9×10^{-2}	5%	4%	5000	4.1
675	4.2×10^{-3}	6%	20%	5000	1.3
700	3.2×10^{-4}	15%	61%	5000	1.7
725	2.8×10^{-5}	−33%	128%	5000	4.4

estimations for a given simulation budget. Nevertheless, this estimation appears to be biased when the available simulation budget is too low.

12.4 Conclusion

EVT proposes a valuable probability estimate on a fixed set of N samples when the target probability does not exceed approximately $\frac{10^{-2}}{N}$ in practice. For rare events, a bias can appear in the estimate, and we must be cautious about its validity. The use of

Table 12.2 Results of CMC for the wake vortex test case

T	$\hat{\mathbb{P}}^{\mathrm{CMC}}$	RB ($\hat{\mathbb{P}}^{\mathrm{CMC}}$)	RE ($\hat{\mathbb{P}}^{\mathrm{CMC}}$)	**Simulation budget**
650	3.4×10^{-2}	14%	58%	100
675	3.8×10^{-3}	−6%	153%	100
700	2.0×10^{-4}	−29%	703%	100
725	1.0×10^{-4}	92%	1022%	100
650	2.7×10^{-2}	−7%	18%	1000
675	3.8×10^{-3}	−5%	55%	1000
700	1.8×10^{-4}	36%	241%	1000
725	3.0×10^{-5}	−29%	571%	1000
650	3.0×10^{-2}	4%	9%	5000
675	3.9×10^{-3}	−3%	21%	5000
700	2.7×10^{-4}	−4%	80%	5000
725	5.2×10^{-5}	19%	175%	5000

EVT on this typical fixed-sample set scenario is nevertheless mandatory to improve the CMC estimation method even if one has to be careful on the tunings of EVT parameters.

References

De Visscher, I., Bricteux, L., and Winckelmans, G. (2013). Aircraft vortices in stably stratified and weakly turbulent atmospheres: Simulation and modeling. *AIAA Journal*, *51*(3), 551–566.

De Visscher, I., Lonfils, T., and Winckelmans, G. (2013). Fast-time modeling of ground effects on wake vortex transport and decay. *Journal of Aircraft*, *50*(5), 1514–1525.

De Visscher, I., Winckelmans, G., Lonfils, T., Bricteux, L., Duponcheel, M., and Bourgeois, N. (2010). The WAKE4D simulation platform for predicting aircraft wake vortex transport and decay: Description and examples of application. In *AIAA atmospheric and space environments conference.* August. Toronto, Ontario, Canada. Paper AIAA 2010-7994.

Manual on low-level wind shear [Computer software manual]. (2005). (Doc 9817- AN/449)

This page intentionally left blank

Estimation of conflict probability between aircraft

13

J. Morio, D. Jacquemart, M. Balesdent

13.1 Principle

A flight plan consists of a sequence of waypoints and speeds along the trajectory. It ensures that the aircraft can safely reach its destination and complies with air traffic control requirements in order to minimize the risk of midair collision. Deterministic trajectory of an aircraft is governed by flight mechanics equations. However, the large number of involved parameters, particularly because of the wind and of the tracking, navigation, and control error, makes the trajectory prediction inexact, and random patterns must be considered. An illustration of aircraft conflict is shown in Figure 13.1.

An aircraft trajectory is well modeled by a stochastic process in continuous time (Prandini, Hu, Lygeros, & Sastry, 2000). A first approach is to include randomness to the flight mechanics equations describing the aircraft motion (Blom, Bakker, Everdij, & Park, 2003). Another method (Hu, Lygeros, Prandini, & Sastry, 1999; Paielli & Erzberger, 1997, 1999) consists in superimposing a stochastic component to the nominal aircraft motion. This chapter uses the latter idea.

More precisely, we assume that the three-dimensional position of an aircraft at time t is given by the following drifted Brownian motion,

$$\mathrm{d}\mathbf{X}_t = \mathbf{v}\,\mathrm{d}t + \boldsymbol{\sigma}_t\,\mathrm{d}\mathbf{W}_t \tag{13.1}$$

where $\mathbf{v}$ is a three-dimensional speed vector, $\boldsymbol{\sigma}_t$ is a correlation matrix, and $\mathbf{W}_t$ is a standard three-dimensional Brownian motion. The initial position is given by some $\mathbf{X}_0$. The cross-correlation between the along track and cross-track error is small enough to be considered null. The vertical position error can be negligible when compared to the along track and cross-track position error (Paielli & Erzberger, 1997). If we assume that the aircraft has a constant speed $v = \|\mathbf{v}\|$ in a coordinate system in which the abscissa axis is parallel to the aircraft trajectory, we have $\mathbf{X}_t = (X_t^{(a)}, X_t^{(c)}, X_t^{(\mathrm{v})})^T$ where

$$\begin{cases} \mathrm{d}X_t^{(a)} = v\,\mathrm{d}t + g^a(t)\,\mathrm{d}W_t^{(a)} \\ \mathrm{d}X_t^{(c)} = g^c(t)\,\mathrm{d}W_t^{(c)} \\ \mathrm{d}X_t^{(\mathrm{v})} = 0 \end{cases} \tag{13.2}$$

Estimation of Rare Event Probabilities in Complex Aerospace and Other Systems
http://dx.doi.org/10.1016/B978-0-08-100091-5.00013-7

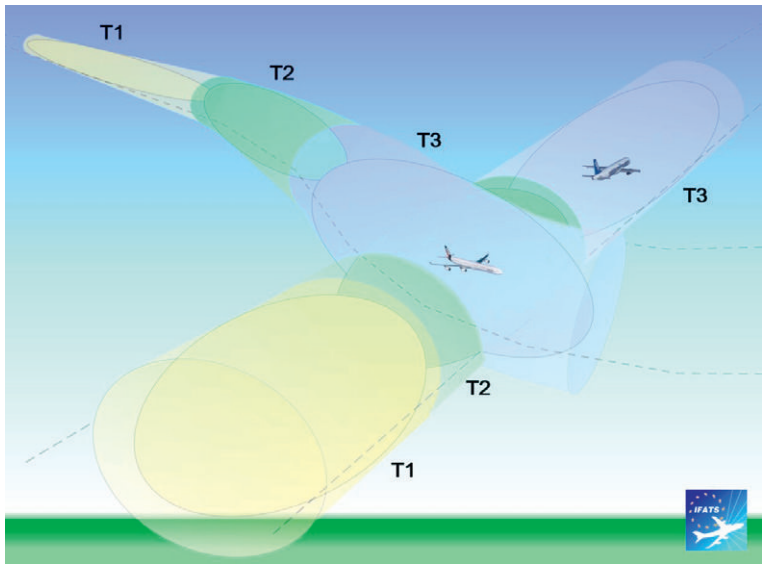

Figure 13.1 Conflict between aircraft. Uncertainty bubbles at time T1, T2, T3. (Courtesy of A. Joulia, Onera).

and $W_t^{(a)}$ and $W_t^{(c)}$ are two independent standard one-dimensional Brownian motions, with $W_0^{(a)} = W_0^{(c)} = 0$ and $X_t^{(a)}$, $X_t^{(c)}$, and $X_t^{(v)}$ denote the along track, cross-track, and vertical positions, respectively.

As commonly assumed (Blom, Krystul, Bekker, Klompstra, & Klein Obbink, 2007; Prandini, Blom, & Bakker, 2011), the relative mean square error of the distance between the expected and real positions of an aircraft is linear with the time t. The slope equals $r_a = 0.25$ nautical miles per minute (nmi min^{-1}) in the along track direction and $r_c = 0.2$ nmi min^{-1} in the cross-track direction.

We thus must determine some functions g^a and g^c so that

$$x_t^{(a)} = x_0^{(a)} + vt + \int_0^t g^a(s)\,\mathrm{d}W_s^{(a)} \sim \mathcal{N}\left(vt, (r_a t)^2\right)$$

and

$$x_t^{(c)} = x_0^{(c)} + \int_0^t g^c(s)\,\mathrm{d}W_s^{(c)} \sim \mathcal{N}\left(0, (r_c t)^2\right)$$

With Itō's isometry, checking that the following two functions satisfy the problem is straightforward

$$g^a(t) = r_a\sqrt{2t}, \quad g^c(t) = r_c\sqrt{2t}$$

Finally, if the speed vector $\mathbf{v}$ has angles θ_1 and θ_2 with the horizontal and the vertical axes, the aircraft position is obtained by applying the corresponding rotation matrix.

13.2 Simulation description

We consider a two-aircraft scenario with a given flight plan. The simulated time horizon is 20 min and the speed of the aircraft is set to 20 nmi min^{-1} (see Figure 13.2). The left panel presents a top view of the two aircraft trajectories, and the right panel describes the separation distance between aircraft as a function of time. For both panels, the dashed lines are the expected position and the expected separation distance. The continuous curves are a realization of the random trajectory of the aircraft given in Equation (13.1).

We are interested in estimating the probability that the distance between aircraft is lower than a given separation distance T during the flight duration S, that is,

$$\mathbb{P}(\mathcal{R}) = \mathbb{P}\left(\left\{\|\mathbf{X}_t^1 - \mathbf{X}_t^2\| \leq T, \text{for some } t \leq S\right\}\right) \tag{13.3}$$

where $\mathbf{X}_t^1$ and $\mathbf{X}_t^2$ denote the position of aircraft 1 and 2, respectively. It is straightforward that Equation (13.3) can be rewritten with the $\varphi(\cdot)$ function defined by

$$\varphi : \begin{cases} \mathbb{R}^3 \times \mathbb{R}^3 \longrightarrow \mathbb{R}^+ \\ (\mathbf{X}^1, \mathbf{X}^2) \longmapsto \|\mathbf{X}^1 - \mathbf{X}^2\| \end{cases} \tag{13.4}$$

We conclude for any $T \geq 0$ that

$$\mathcal{R} = \left\{\inf_{0 \leq t \leq S} \varphi(\mathbf{X}_t^1, \mathbf{X}_t^2) \leq T\right\} \tag{13.5}$$

13.3 Estimation results

The different algorithms proposed in Chapter 9 are applied to estimate $\mathbb{P}(\mathcal{R})$. For each algorithm, the stochastic process (13.1) is implemented using the Euler scheme with a discretization step $\Delta = 0.05$ min. Each estimation is performed with 50 retrials. A

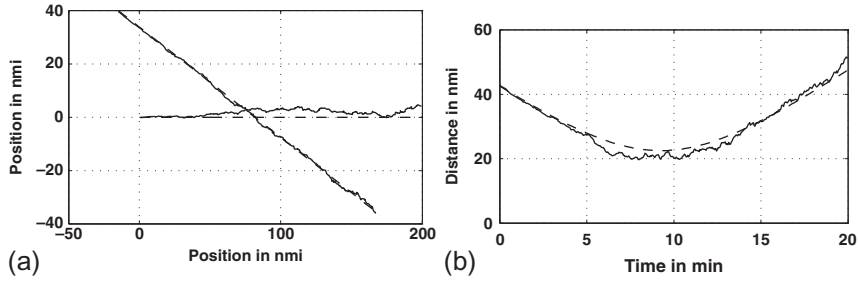

Figure 13.2 Aircraft positions (a) and separation distance as functions of time (b). The expected position and separation distance are in dashed lines, and the true position and separation distance are in continuous lines.

one-shot crude Monte Carlo (CMC) estimation of 10^8 samples is performed and shows that the target probability is approximately 2.26×10^{-6}. We will take this probability as a reference to estimate RE and RB for the different methods.

13.3.1 Importance splitting

General adaptive importance splitting algorithm (GAISA) is well adapted to Markovian simulation model. It is applied with the following parameters: the sample number $\tilde{N}$ is set to $\tilde{N} = 0.1 \times N$, and the quantile parameter ρ is equal to 0.5. Several GAISA sample paths are illustrated in Figure 13.3. Tuning of GAISA parameters is not complicated because their influence is not very important on the estimation accuracy. GAISA converges to the expected probability with a high efficiency compared to CMC (Table 13.1).

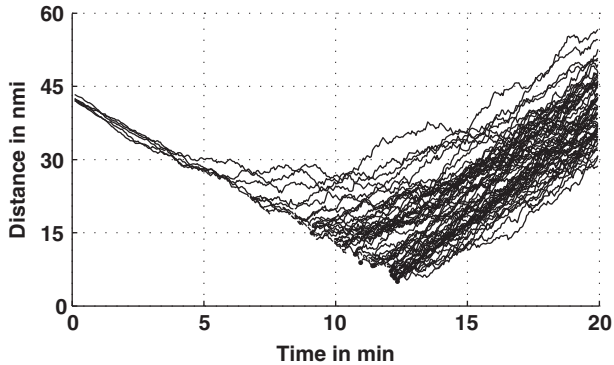

Figure 13.3 Typical distance processes obtained with GAISA.

Table 13.1 Results obtained with GAISA for aircraft collision case

$\hat{\mathbb{P}}^{\text{GAISA}}$	RB($\hat{\mathbb{P}}^{\text{GAISA}}$)	RE($\hat{\mathbb{P}}^{\text{GAISA}}$)	N	**Simulation budget**	ν^{GAISA}
1.22×10^{-6}	-45%	256%	100	685	98
2.22×10^{-6}	-1%	174%	1000	6295	23
2.28×10^{-6}	1%	42%	10,000	6.1×10^4	40

13.3.2 Weighted importance resampling

Weighted importance resampling (WIR) is also applied to the aircraft collision case. The parameter α must be tuned in this algorithm and has a high influence on the estimated probabilities. About $60,000$ samples are necessary in this use case to determine an efficient value with $\alpha = 1.14$. Table 13.2 shows the results for different budgets and α values. Moreover, typical stochastic processes obtained with WIR are

Table 13.2 Results obtained with WIR for aircraft collision case

$\hat{\mathbb{P}}^{\text{WIR}}$	α	RB ($\hat{\mathbb{P}}^{\text{WIR}}$)	RE ($\hat{\mathbb{P}}^{\text{WIR}}$)	N	ν^{WIR}
2.08×10^{-6}	1.14	−11%	55%	1000	1467
2.10×10^{-6}	1.14	−10%	16%	10,000	1728
2.10×10^{-6}	1.14	−10%	6%	100,000	1229
2.10×10^{-6}	0.5	−10%	47%	100,000	20
1.65×10^{-6}	2	−26%	60%	100,000	12.3

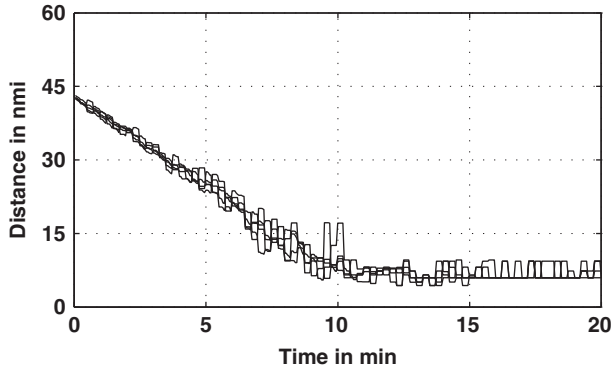

Figure 13.4 Typical distance processes obtained with WIR.

given in Figure 13.4. The tuning of α may be complicated and has a significant influence on the probability estimation, but if this parameter is accurately determined, the performances of WIR algorithm are very high comparatively to CMC and even GAISA.

13.4 Conclusion

For this use case, only two simulation techniques are adapted. Indeed, CMC is computationally intractable for this magnitude order of target probability. Moreover, it is impossible to find an efficient importance sampling (IS) distribution because of the distance process complexity and the problem dimensionality in the Markov space. Nor is extreme value theory adapted when resampling is possible and therefore has not been applied considered. Only WIR and GAISA can be applied with efficiency to estimate the probability of aircraft conflict proposed in this chapter and obtain accurate results. WIR is the most efficient method but needs the fine-tuning of a selection parameter α that significantly influences the WIR probability relative error. A simulation budget required to determine an efficient α must be considered. On the contrary, GAISA is more robust to its different parameter variation, but if it is well-tuned GAISA seems to give less accurate results than WIR.

References

Blom, H., Bakker, G., Everdij, M., and Van der Park, M. (2003). Collision risk modeling of air traffic. In *Proceedings of European control conference.* Cambridge, UK.

Blom, H., Krystul, J., Bekker, G., Klompstra, M., and Klein Obbink, B. (2007). Free flight collision risk estimation by sequential MC simulation. In C. Cassandra and J. Lygeros (Eds.), *Stochastic hydrid systems* (pp. 248–282). Boca Raton, FL: Taylor and Francis Group, CRC Press.

Hu, J., Lygeros, J., Prandini, M., and Sastry, S. (1999). Aircraft conflict prediction and resolution using Brownian motion. In *Proceedings of the 38th IEEE conference on decision and control* (Vol. 3, pp. 2438–2443). Phoenix, USA. December.

Paielli, R., and Erzberger, H. (1997). Conflict probability estimation for free flight. *Journal of Guidance Control and Dynamics*, *20*(3), 588–596.

Paielli, R., and Erzberger, H. (1999). Conflict probability estimation generalized to non-level flight. *Air Traffic Control Quarterly*, *7*(3), 195–222.

Prandini, M., Blom, H., and Bakker, G. (2011). Air traffic complexity and the interacting particle system method: An integrated approach for collision risk estimation. In *American Control Conference (ACC)* (pp. 2154–2159). San Francisco, USA.

Prandini, M., Hu, J., Lygeros, J., and Sastry, S. (2000). A probabilistic approach to aircraft conflict detection. *IEEE Transactions on Intelligent Transportation Systems*, *1*(4), 199–220 (December).

Part Four

Practical Guidelines of Rare Event Probability Estimation

This page intentionally left blank

Synthesis of rare event probability estimation methods for input–output systems

14

J. Morio, M. Balesdent

14.1 Synthesis

This book presents an overview of the different possible algorithms that are applicable to the estimation of rare event probabilities modeled by a threshold exceedance of an input–output function $\phi(\cdot)$. In general, we advise to consider the following algorithms in order to improve the probability estimation accuracy of crude Monte Carlo (CMC): nonparametric adaptive importance sampling (NAIS), cross entropy (CE), adaptive splitting technique (AST), first-order reliability method/second-order reliability method (FORM/SORM), line sampling (LS), directional sampling (DS), adaptive directional sampling (ADS), extreme value theory (EVT), and their potential association with a surrogate model. The algorithm domain of applicability can vary significantly from one algorithm to another (Table 14.1).

The proposed synthesis consists of a series of questions that can help the reader to choose the appropriate methods for a specific estimation problem.

1. Is the function $\phi(\cdot)$ available to sample the output? If resampling is not possible, that is, if we consider a fixed set of samples $\phi(\mathbf{X}_1), \ldots, \phi(\mathbf{X}_N)$, the only available method is EVT. A surrogate model could also be built if the samples $\mathbf{X}_1, \ldots, \mathbf{X}_N$ are also known, thus enabling to estimate the rare event probability with the surrogate model instead of the true function $\phi(\cdot)$. Notice that in that case, there might be a high risk of biased estimation if the surrogate is not accurate in the relevant zones for rare event estimation. If resampling is possible, the other proposed algorithms should be considered.
2. Is the input failure region multimodal? If so or if the answer to this question is not known, the use of first-order reliability method (FORM), second-order reliability method (SORM), or LS is not advised. CE can cope with a multimodal failure region if its parametric density family is well tuned. The other algorithms can deal with multimodal failure region.
3. What is the dimension d of the input? If $d < 10$ (value given as an order of magnitude), NAIS, ADS, CE, LS, FORM, and SORM are often very efficient on very general functions $\phi(\cdot)$. AST and DS are applicable but are not the best estimation techniques. If $10 < d < 30$ (values given as an order of magnitude), AST, CE, DS, LS, and FORM/SORM are the most efficient algorithms. If $d > 30$, AST, LS, and FORM/SORM are the most appropriate techniques. Consider the opportunity of performing sensitivity analysis to reduce the input dimension.

Estimation of Rare Event Probabilities in Complex Aerospace and Other Systems
http://dx.doi.org/10.1016/B978-0-08-100091-5.00014-9

4. What is the available simulation budget N? If $N < 1000$ (value given as an order of magnitude), FORM/SORM, LS, DS, or ADS must be used. Also CE, NAIS, and AST can be applied when $N < 1000$ and jointly used with a surrogate model. If $N > 1000$ (value given as an order of magnitude), CE, NAIS, and AST can then be applied, but LS, DS, or ADS are still efficient. When FORM/SORM requires more than 1000 samples, this is often because it has not converged.
5. Is the function $\phi(\cdot)$ highly nonlinear? Surrogate model and FORM/SORM can induce a bias in the estimation and must be applied carefully. The probability approximation proposed in LS, DS, and ADS also might be not accurate.
6. Is a control of the estimation error required? If so, FORM/SORM could not be applied. Depending on the algorithms, the estimation error can be estimated with retrials or an analytical formula.

14.2 Some remarks for a successful practical rare event probability estimation

The following are some tips for estimating rare event probability successfully:

- Pay great attention to the computer code modeling and the joint pdf input definition (particularly the statistical dependence between the inputs).
- Learn some computer code characteristics (the most influential parameters, etc.).
- If the simulation budget is restrained, consider the use of a surrogate model instead of the complex function $\phi(\cdot)$ for the probability estimation. For high-dimensional systems, evaluate a possible reduction of the input dimension. In both cases, be aware that it could result in a biased probability estimate.
- Avoid the use of algorithms unadapted to the structural characteristics of the input space or of the input–output function (e.g., application of NAIS on a high-dimensional system).
- Using your experience, tune the parameters of the algorithms if the simulation budget is important (the parameter tuning proposed in this book is of course not optimal for all situations).
- Characterize the estimation of the rare event probability by an error indicator (a relative error, a bound, etc.).
- Be sure that the input–output function's domain of definition is well characterized. Indeed, because most of the rare event estimation techniques generate samples in zones that are not usually activated for classical use of the black-box function, we must verify that this function will provide consistent results in these zones.

Table 14.1 Synthesis table

	Impossibility of resampling	**Failure region disjoint; info not available**	**Dimension d (order of magnitude)**	**Simulation budget N**	**Control of estimation error**	**ϕ Nonlinear**	**Potential efficiency**
NAIS	×	√	<10	>1000 (<1000 with surrogate)	Analytic formula or retrials	√	+++
CE	×	×	<30	>1000 (<1000 with surrogate)	Analytic formula or retrials	√	+++
AST	×	√	≥1	>1000 (<1000 with surrogate)	With retrials	√	++
FORM/SORM	×	×	≥1	>10	Impossible	×	not applicable
LS	×	×	≥1	>100	With retrials	×	++
DS	×	√	<30	>1000	With retrials	×	++
ADS	×	√	<10	>100	With retrials	×	+++
EVT	√	√	≥1	>100	With bootstrap	√	+

"√"—The method presents some advantages for the characteristic being considered.
"×"—The method presents some drawbacks for the characteristic being considered.
+++: very good; ++: good; +: medium.

This page intentionally left blank

Synthesis for time-variant systems

15

D. Jacquemart, J. Morio, M. Balesdent, F. Le Gland

15.1 Synthesis

Over the past 50 years, different algorithms have been worked out for the probability estimation of identifying a rare set in a dynamic Markov framework. Selecting the most adapted method for the problem modeling requires a study of the state of the art. In our opinion, four main algorithms can be considered for rare event probability estimation: extreme value theory (EVT), importance sampling (IS), weighted importance resampling (WIR), and general adaptive importance splitting algorithm (GAISA). IS can provide accurate results in some cases, especially when the state spaced is finite or countable. Otherwise, GAISA and WIR algorithms are often preferable. To choose the best adapted algorithm for a given case, we propose a decision tree (Figure 15.1). Some information on the different leaves of this decision tree are given in the following.

- *Leaf* 1. EVT is the only alternative when no resampling is possible. It can also be viewed as a quick estimation to get a rough idea of the order of magnitude of the probability of interest.
- *Leaf* 2. Because of the broad diversity of complex problems, there is, to our knowledge, no specific procedure for such a case in the literature. If the simulation time is continuous and if the rare event is modeled by a threshold exceedance of a real score function φ, GAISA techniques can be used. A continuous time interpolation makes a splitting plan possible, setting $\mathbf{X}_t = \mathbf{X}_{i/n}$ for $i/n \leq t < (i+1)/n$ with a score function φ. IS can be efficient on a case-by-case basis for finite or countable state spaces if we have a good idea of auxiliary sampling distributions.
- *Leaf* 3. For time continuous problems and for rare events that are modeled by a threshold exceedance of a real score function φ, GAISA can be applied. For discrete time processes, WIR algorithm is efficient, especially more than GAISA, but we must choose its parameter tuning properly. The same continuous time interpolation as for Leaf 2 is required when using a splitting plan.
- *Leaf* 4. Efficient IS algorithms have been worked out and are often considered in practice. GAISA can also be applied for real function threshold exceedance characterization of the rare event. Even if the WIR algorithm is tractable, its use is not always recommended in this case because of its sensitive tunings in comparison with IS algorithms.
- *Leaf* 5. The implementation of splitting requires an easy tuning and is the best candidate. Stopping time S can be either random or deterministic.

Estimation of Rare Event Probabilities in Complex Aerospace and Other Systems
http://dx.doi.org/10.1016/B978-0-08-100091-5.00015-0

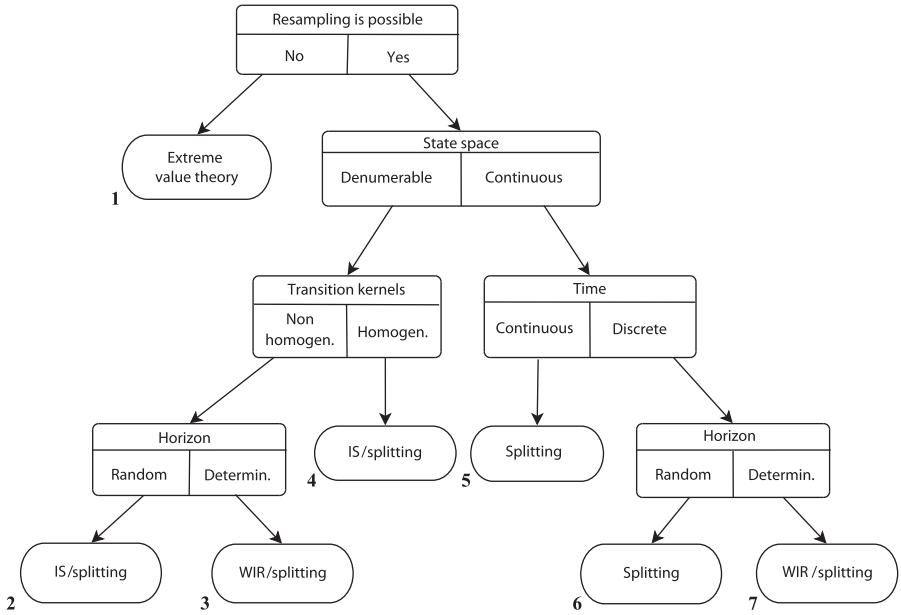

Figure 15.1 Algorithm decision tree for rare event probability estimation of time-variant systems (Homogen. = Homogeneous and Determin. = Deterministic).

- *Leaf* 6. GAISA algorithm with a time continuous interpolation of the process will provide good results. An IS plan is still possible, but determining an efficient auxiliary distribution is a complicated task in general.
- *Leaf* 7. GAISA and WIR algorithms are both tractable. WIR is possibly difficult to tune correctly and efficiently. Nevertheless, if WIR is well tuned, it often performs better than GAISA.

15.2 Some remarks for a successful practical rare event probability estimation

The following are some things to know in order to estimate rare event probability successfully:

- Pay attention to the transition kernel modeling, particularly the tail.
- For time-continuous processes, remember that the discretization step tuning is a trade-off between computational time and probability bias.
- Tune the parameters of the probability estimation algorithms if the simulation budget is important. If the simulation budget is limited, it is more often adapted to the use of GAISA than WIR.
- Be careful when using IS. Most of the time, defining an efficient auxiliary sampling distribution is complicated.
- Characterize the estimation of the rare event probability with an error indicator (a relative error, a bound, etc.).

Index

WOODHEAD PUBLISHING
IN MECHANICAL ENGINEERING

For reliability and safety reasons, accurate estimation of critical rare event probabilities (10^{-4} and less) is required in a large range of applications and notably in aerospace systems. *An Estimation of Rare Event Probabilities In Complex Aerospace And Other Systems* aims to provide practioners with the most adapted methods depending on their applications for the estimation of very low probabilities. This book also gives a broad view of the current research on rare event and thus, can be potentially interesting for graduate students and young researchers.

Dr. Jérôme Morio is a Research fellow at Onera, the French Aerospace Lab. He graduated from the Ecole Centrale of Marseille, France, and obtained a Ph.D. degree in physics from University Aix-Marseille Paul Cezanne, France. His main research interests include rare event simulation, sensitivity analysis, uncertainty propagation and optimization for large scale and time-dependent systems. He has published more than 30 articles in scientific publications and supervised 4 Ph.D. students on these subjects.

Dr. Mathieu Balesdent is a Research Engineer at Onera, the French Aerospace Lab. He graduated from the Institut Superieur d'Electronique et du Numerique Lille, France and from the Ecole Centrale and the University of Science, Lille, France. He obtained a Ph.D. in Mechanical Engineering from the Ecole Centrale, Nantes, France in 2011, and won two prizes for his thesis in 2012; Centrale Innovation's Best Ph.D. thesis award and PERSEUS award. His main research interests include multidisciplinary optimization, rare event estimation and optimization under uncertainty. He has published more than 15 articles in scientific publications of conference proceedings relative to these topics.

ELSEVIER
WP
WOODHEAD
PUBLISHING
An imprint of Elsevier • store.elsevier.com

ISBN 978-0-08-100091-5
9 780081 000915