TECHNISCHE UNIVERSITÄT MÜNCHEN
Ingenieurfakultät Bau Geo Umwelt
Fachgebiet für Risikoanalyse und Zuverlässigkeit

Bayesian inference of engineering models

Wolfgang Betz

Vollständiger Abdruck der von der Ingenieurfakultät Bau Geo Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs

genehmigten Dissertation.

Vorsitzender:	Prof. Dr.-Ing. Kai-Uwe Bletzinger
Prüfer der Dissertation:	
	1. Prof. Dr. sc. techn. Daniel Straub
	2. Prof. Siu-Kui Au, Ph.D. University of Liverpool
	3. Prof. Phaedon-Stelios Koutsourelakis, Ph.D.

Die Dissertation wurde am 04.09.2017 bei der Technischen Universität München eingereicht und durch die Ingenieurfakultät Bau Geo Umwelt am 27.11.2017 angenommen.

This work is dedicated to my children
Maike and Thomas

Acknowledgements

Over the last few years, many people have helped me to grow, both academically and professionally. Without their support, this work would not have been possible.

Prof. Dr. Daniel Straub offered me the opportunity to work on a topic that truly inspired me. He has been an excellent teacher as well as supervisor. At the right points in time during my research, he put my focus on Bayesian inference and showed me his ideas about BUS. However, he also gave me the freedom and provided the time necessary to develop my own ideas. This balance between guidance and freedom that I found in his supervision was essential to my research and to putting my ideas in shape. The time he invested in our fruitful discussions is highly appreciated. Daniel's leadership approach fostered the pleasant working atmosphere that prospered in his research group. I consider it a privilege to have been part of this team.

Dr. Iason Papaioannou has been a supportive and highly competent mentor for many years. He taught the first lecture I took when starting my studies in Munich, and he is the one that initially got me interested in reliability analysis. Iason supervised me as a student assistant, during my study project, my Master's thesis and the time it took me to work on and finish my PhD. He also established the initial contact to Daniel. Our weekly meetings constantly let to highly interesting and relevant discussions. The friendship that developed made the work with him even more enjoyable.

Prof. James L. Beck hosted me during my research stay at Caltech. I consider myself very lucky that I could take his lecture on Probability Theory – it was the last lecture he taught at Caltech. This very lecture as well as the enlightening discussions I had with him sparked my interest in the philosophical questions behind Bayesian inference.

I am honored that Prof. Siu-Kui Au and Prof. Faidon-Stelios Koutsourelakis reviewed my thesis and acted as referees. I am grateful to Prof. Armen der Kiureghian and Jasper Vrugt for accepting me as visiting researchers at UC Berkeley and UC Irvine, respectively. To Dr. Chin Man W. Mok I express my thanks for the helpful discussions regarding uncertainties in hydrological modeling, and for agreeing to associate my position with his Fellowship.

Last but not least I would like to express my deepest gratitude to my family; my parents Georg and Christa for their constant support during my education, my sister Steffi for proofreading the manuscript, and my wife Micha for putting up with my late hours working on the thesis.

Wolfgang Betz
Schwabach, February 2018

Abstract

A new method for numerical Bayesian inference, termed *aBUS*, is proposed, which is based on Subset Simulation applied within the BUS (Bayesian Updating with Structural reliability methods) approach. The performance of *aBUS* is independent of the number of uncertain parameters of the problem by virtue of Subset Simulation. *aBUS* produces samples from the posterior distribution and returns an estimate for the evidence of the Bayesian inference problem. Compared to the standard BUS approach, the proposed method does not require knowledge about the maximum of the likelihood function.

Besides *aBUS*, other methods for numerical Bayesian inference that generate samples from the posterior distribution and provide an estimate for the evidence are discussed. For the Transitional Markov Chain Monte Carlo (TMCMC) method and for Nested Sampling, potential modifications are proposed. For the BUS approach, a Metropolis-Hastings based post-processing step is proposed that enables application of standard BUS even if the maximum value of the likelihood function is unknown.

Furthermore, probabilistic modeling approaches for the prior distribution and probability models for error structures are investigated, and guidelines for probabilistic modeling are given. The contribution of prior distribution and likelihood function to the evidence of the Bayesian inference problem is discussed. It is shown that modeling as well as measurement errors can be represented in terms of the prior probabilistic model or in terms of the likelihood function without influence on the evidence of the inference problem, but with considerable consequences for the efficiency of the applied numerical methods.

Moreover, the notion of uncertainty is discussed and a probabilistic framework for uncertainty quantification is developed. The framework highlights the personal aspect of probability and argues against classifying probabilities as *subjective* or *objective*. The introduced framework is intended as a probabilistic foundation for stochastic forward analysis and Bayesian inference.

Zusammenfassung

Ein neues numerisches Verfahren, genannt *aBUS*, zur Durchführung einer Bayes'schen Inferenz wird eingeführt. Dieses basiert auf Subset Simulation, angewendet innerhalb des BUS (Bayesian Updating with Structural reliability methods) Ansatzes. Durch Subset Simulation als Basis ist die Performance des Verfahrens unabhängig von der Anzahl der unsicheren Modellparameter. *aBUS* erzeugt Stichproben der Posterior-Verteilung und schätzt die Plausibilität des verwendeten stochastischen Modells ab. Verglichen mit anderen BUS Verfahren benötigt der vorgeschlagene Ansatz das Maximum der Likelihood Funktion nicht als Eingangsgröße.

Neben *aBUS* werden weitere numerische Bayes'sche Inferenz-Methoden betrachtet, welche Stichproben der Posterior-Verteilung liefern und die Plausibilität des verwendeten stochastischen Modells abschätzen. Potentielle Änderungen werden für TMCMC (Transitional Markov Chain Monte Carlo) und für Nested Sampling vorgeschlagen. Für BUS wird ein auf dem Metropolis-Hastings Algorithmus basierender Nachbereitungsschritt vorgeschlagen, welcher es erlaubt, BUS selbst dann einzusetzen, wenn das Maximum der Likelihood Funktion nicht genau bekannt ist.

Darüber hinaus wird die probabilistische Modellierung der Prior-Verteilung und die probabilistische Beschreibung von Fehlermodellen diskutiert. Ratschläge für die probabilistische Modellierung werden gegeben. Der Einfluss der gewählten Prior-Verteilung und der Likelihood Funktion auf die berechnete Plausibilität des stochastischen Modells wird erläutert. Es wird gezeigt, dass Modell- und Messfehler sowohl durch die Prior-Verteilung als auch durch die Likelihood Funktion ausgedrückt werden können. Die berechnete Plausibilität hängt hiervon nicht ab, wohl aber der zur Lösung des Problems nötige Rechenaufwand.

Weiterhin wird die Interpretation von Unsicherheiten diskutiert und ein probabilistisches Grundgerüst für die Quantifizierung von Unsicherheiten erarbeitet. Der *persönliche* Aspekt der Unsicherheiten wird hervorgehoben und es wird sich gegen eine Klassifizierung von Unsicherheiten als entweder *subjektiv* oder *objektiv* ausgesprochen. Das eingeführte probabilistische Grundgerüst ist als Basis für sowohl einfache stochastische Analysen als auch für Bayes'sche Inferenzen gedacht.

Contents

Nomenclature

Terms, symbols and abbreviations are generally explained when first introduced.

List of Terms

Expression	**Description**
Bayesian ...	... framework/probability refers to the Cox-Jaynes interpretation of probability.
conditional on	"Statement A" being *conditional on* "Statement B" means that "Statement A" holds if "Statement B" is imposed as *true*.
model	is an approximate representation of a real *system*. In this work, *model* refers to the "tool" that is used to approximate the response of the *true* underlying *system*. In the simplest case a model is an explicit mathematical function that approximates the *system*. For more general cases, the model is a numerical approximate of the *system* response conditional on uncertain model parameters.
Probability Theory	refers to the Cox-Jaynes interpretation of probability.
proposition	is a statement that is either *true* or *false*
random variable	denotes a quantity that is uncertain.
stochastic variable	refers to a quantity that is fixed in reality, but whose value we are uncertain about. Note: The term *stochastic variable* is more specific than the expression *random variable*.
system	refers to an actual system in the real-world. In this context, the response of a *system* can never be represented exactly, but can only be approximated by a *model*. Examples for *systems* are: a tunnel to be constructed, an existing bridge, a hydrological catchment, ...

List of Symbols

Latin characters

Symbol	Description
$c_{v,\mathbf{X}}\left[g(\mathbf{X})\right]$	coefficient of variation (C.o.V.) of function $g : \mathbb{R}^M \to \mathbb{R}$ with respect to random variable $\mathbf{X} \in \mathbb{R}^M$.
$\mathrm{Cov}(A, B)$	covariance between random variables A and B
$\mathcal{D}$	observed data/information
$\mathrm{E}_{\mathbf{X}}\left[g(\mathbf{X})\right]$	expectation of function $g : \mathbb{R}^M \to \mathbb{R}$ with respect to random variable $\mathbf{X} \in \mathbb{R}^M$.
$\mathbf{f}$	model input; observed as $\mathbf{s}$
$I_g(\mathbf{X})$	indicator function: $$I_g(\mathbf{X}) = \begin{cases} 1 & \text{if } g(\mathbf{X}) \leq 0 \\ 0 & \text{otherwise} \end{cases}$$ with $g : \mathbb{R}^M \to \mathbb{R}$.
$L(\boldsymbol{\theta}\|\mathcal{D})$	Likelihood function in Bayesian inference
$L_{\max}$	maximum value that the likelihood function can take
M	number of random/stochastic variables in a problem
$\mathcal{M}$	information/knowledge available a-priori in a Bayesian framework. $\mathcal{M}$ is referred to as a *stochastic model class*. Note: all propositions are either implicitly or explicitly conditioned on π.
$\mathbf{M}$	set that contains m stochastic model classes; i.e., $\mathbf{M} = \{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_m\}$
m	number of stochastic model classes in a set $\mathbf{M}$ of stochastic model classes.
$\mathrm{P}[A]$	This notation is used to express the probability of event A if probability is not quantified according to Cox-Jaynes interpretation of probability.
P_f	probability of failure in reliability analysis
p_f	estimate for the probability of failure P_f
$p_X(\cdot)$	probability density function (PDF) of X
$P_X(\cdot)$	cumulative distribution function (CDF) of X
$\Pr[A\|\pi]$	probability that proposition A is *true* conditional on proposition π. *Probability* is quantified according to Cox-Jaynes interpretation of probability.
$\mathbf{q}$	*output* of a *model* that approximates the response of a *system*
$\mathbf{r}$	output generated by the "real" system; approximated by the model output $\mathbf{q}$ and observed as $\mathbf{z}$.
$\mathbf{s}$	observation of the model input $\mathbf{f}$.
$S(\mathcal{D}\|\mathcal{M})$	self-information in observation $\mathcal{D}$ conditional on the chosen probabilistic model $\mathcal{M}$

Symbol	Description
$\mathbf{U}$	vector of M independent standard Normal random variables
$\mathbf{u}$	realization of $\mathbf{U}$
$\mathrm{Var}_{\mathbf{X}}[g(\mathbf{X})]$	variance of function $g : \mathbb{R}^M \to \mathbb{R}$ with respect to random variable $\mathbf{X} \in \mathbb{R}^M$.
$\mathbf{v}$	*output prediction-error* that describes the relation between $\mathbf{q}$ and $\mathbf{r}$.
$\mathbf{w}$	*observation-error* that describes the relation between $\mathbf{z}$ and $\mathbf{r}$.
X	one-dimensional *random variable*
$\mathbf{X}$	M-dimensional *random variable*
$\mathbf{x}$	a realization of random vector $\mathbf{X}$
x_i	the ith component of vector $\mathbf{x}$
$\mathbf{Y}$	vector of M possibly correlated standard Normal random variables
$\mathbf{y}$	realization of $\mathbf{Y}$
$\mathbf{z}$	observed output of a system; observation of $\mathbf{r}$.

Greek characters

Symbol	Description
Γ	support of a stochastic quantity
Ω	failure domain of the reliability problem in the BUS approach; a sample from the prior distribution of the associated Bayesian inference problem that falls in this domain is a sample that follows the posterior distribution.
$\varphi(\cdot)$	PDF of the univariate standard Normal distribution
$\varphi_M(\mathbf{u})$	joint PDF of the M-dimensional vector $\mathbf{u}$ of independent standard Normal random variables
$\Phi(\cdot)$	CDF of the univariate standard Normal distribution
π	auxiliary random variable in the BUS approach
θ	one-dimensional *stochastic variable*
$\boldsymbol{\theta}$	M-dimensional *stochastic variable* The symbol $\boldsymbol{\theta}$ is used to denote both the realization and the random variable/vector. If a distinction between the actual realization and the random variable/vector is important, $\boldsymbol{\theta}$ is used for the realization, and Θ to denote the random variable/vector.
υ	denotes both the realization and the random variable with underlying uniform distribution on support $[0, 1]$.

other Mathematical operators and symbols

Symbol	Description
$b\|a$	proposition b conditional on proposition a. Note: As proposition a is only conditionally asserted, it does not mean that a must be *true* in reality.
$a \wedge b$	is *true* if and only if both proposition a and b are *true*.
$a \vee b$	is *true* if and only if either proposition a or b is *true*.
$\overline{b}$	is *true* if and only if proposition b is *false*.
$\mathbf{a} \leq \mathbf{b}$	the vector inequality is *true* if and only if $a_i \leq b_i \; \forall \; i \in M$; $\mathbf{a}, \mathbf{b} \in \mathbb{R}^M$. The operators $<$, $>$, $\geq$ are defined accordingly.

Abbreviations

Abbreviation	Description
BUS	Bayesian Updating using Structural reliability methods
C.o.V.	coefficient of variation
CDF	cumulative distribution function
CS	Conditional Sampling in standard Normal space MCMC
ESJD	expected squared jumping distance
MCMC	Markov chain Monte Carlo
MCS	Monte Carlo simulation
MH	Metropolis-Hastings MCMC
MEP ...	Maximum Entropy probability ...
PDF	probability density function
SuS	Subset Simulation
TMCMC	Transitional Markov chain Monte Carlo

Chapter 1

Introduction

1.1 Motivation

Numerical models are used in many industries to assess and predict the behavior of a real system. Engineers use numerical models to assess the safety and serviceability of structural designs, or to investigate how existing structures can be improved or repaired. In meteorology, weather forecasts are obtained by numerical models that are so complex that only supercomputers can handle them in time. In biology, numerical models provide the means to study complex phenomena and processes that are too expensive or too dangerous to study in an experimental setting. Numerical models are employed by the financial industry to predict future behavior of the stock market, by space agencies to perform a detailed simulation of the anticipated voyage of any spacecraft before it leaves earth, by authorities for emergency planning to simulate the evacuation of sport stadiums or subway stations, and so forth.

Increasing computing capacity of modern processors facilitates the use of computationally more and more complex models. However, no matter how complex our models become, any model is necessarily approximate in its representation of the real system of interest. This has mainly to be attributed to our limited knowledge and imprecise understanding of the real world and limited computational power. In this regard, it has to be acknowledged that any model-based assessment and prediction is – to a certain degree – uncertain. The input of the model, the parameters of the model, the error in the modeled system response and the error in the measured system response are quantities that cannot be specified with absolute certainty. In engineering, this lack of knowledge is usually compensated implicitly by means of safety factors; i.e., structural safety and serviceability are demonstrated on a consciously selected conservative deterministic model. Such an implicit representation of uncertainty is adopted for computational reasons, and because it is easier to understand by practical engineers. If uncertainty is explicitly taken into account, the computational demand required to process the model increases significantly. Besides reduced computational

requirements, the use of standardized safety factors in engineering is comprehensible and ensures that structures of similar type possess similar reliability – even if the anticipated reliability is not quantified directly. Contrary to that, the intellectual challenge of setting up an appropriate probabilistic model (a model that takes the uncertainties explicitly into account) is substantial. Inappropriate assumptions in the stochastic modeling of uncertainties can be hard to detect and may have severe consequences. If an engineering structure is, however, not covered by technical standards, if the uncertainties in the model response are the target quantity of interest, or if measured or observed data should be used to identify and learn the model, relevant uncertainties at hand should be explicitly taken into account.

One can principally distinguish between two stochastic analysis categories: (i) *stochastic forward analysis* and (ii) *Bayesian analysis*. In both categories, initially, the available information/knowledge is used to formulate the uncertainties and set up the model. In the first category, the considered uncertainties are propagated through the model. In the second category, measured or observed data is used to reduce the uncertainties considered in the stochastic model by means of an inverse analysis. Except for a few special cases, in both categories, the problem has typically to be solved numerically. Most numerical methods require multiple evaluations of the model with all uncertain parameters conditionally fixed. This is why a stochastic analysis entails a larger computational demand than a deterministic analysis. In this context, the numerical efficiency of stochastic analysis methods is usually assessed in terms of the required number of model evaluations.

In Bayesian analysis the aim is to identify the posterior distribution of the model parameters conditional on data and observations. Desired requirements for numerical Bayesian inference methods are: (i) An approximation of the posterior distribution, or samples of the posterior distribution should be returned. (ii) The evidence is ideally evaluated as a by-product of the method, where the evidence is a number that quantifies the plausibility of the employed stochastic model. (iii) The involved computational cost should be reasonable; i.e., the required number of model evaluations should not be too large. (iv) There should be a standard configuration of the method that works efficiently for a large variety of inference problems; i.e., the user should not have to worry about specifying configuration parameters of the method. (v) Moreover, the user should not have to worry about convergence of the generated samples to the posterior. (vi) The method should perform well for stochastic models with large as well as with small a numbers of uncertain parameters. (vii) The inference approach should be non-intrusive; i.e., for conditionally fixed model parameters, the underlying deterministic model is utilized as a "black box".

Markov chain Monte Carlo (MCMC) methods constitute a popular class of methods to sample from the posterior distribution [Gilks et al., 1996; Gelman et al., 2004a]. However, one problem of MCMC methods is that after an initial burn-in phase the samples may not yet have reached the stationary distribution of the Markov chain [Plummer et al., 2006]. That is,

finding an appropriate burn-in period in MCMC is often a non-trivial problem. Moreover, an estimate for the evidence of the stochastic model is not a direct by-product of MCMC methods. Another issue is that standard MCMC algorithms usually cannot be applied efficiently for problems with many uncertain parameters. Some specialized MCMC algorithms [Haario et al., 2005; Robert and Tweedie, 1996; Neal, 2011; Cheung and Beck, 2009] can cope with such high dimensional problems, they require however additional evaluations of the likelihood function or its gradient for each generated sample. Advanced methods based on MCMC sampling that try to overcome the afore mentioned issues include particle filter methods [Chopin, 2002] (e.g., the Transitional Markov Chain Monte Carlo (TMCMC) method [Ching and Chen, 2007]), the nested sampling approach (see e.g., [Skilling et al., 2006]) and Bayesian updating with structural reliability methods (BUS) [Straub and Papaioannou, 2015]. These methods essentially differ in how samples starting from the prior distribution propagate to the posterior distribution. An advantage of the BUS approach is that existing methods from structural reliability can be employed to solve the Bayesian inference problem. A particularly interesting combination is the use of Subset Simulation within BUS. By considering different resolutions of the underlying deterministic model, the overall computational cost can be reduced [Koutsourelakis, 2009a,b]. The numerical Bayesian analysis is started on the coarse mesh and the mesh resolution is gradually improved as the samples propagate to the posterior distribution. Approximate Bayesian computation (ABC) approaches [Beaumont et al., 2009; Csilléry et al., 2010; Turner and Van Zandt, 2012; Chiachio et al., 2014] are an alternative that bypass the direct evaluation of the likelihood function. Apart from sampling based methods, the posterior distribution can also be approximated directly. Methods in this category include the Laplace approximation [Laplace, 1986], variational Bayes theory [MacKay, 1995; Neal and Hinton, 1998], and sparse variational Bayes algorithms [Franck and Koutsourelakis, 2016].

Setting up a stochastic model for a Bayesian analysis is more involved than for a forward analysis. In a forward analysis, the impact of assumptions and probabilistic modeling choices on the stochastic response can be assessed by reasoning how they propagate through the model. In a Bayesian inference, uncertainties are reduced by means of a inverse analysis. In this case, the consequences of assumptions are more difficult to trace as the learning process is often not intuitively understand. Loosely speaking, learning in a Bayesian framework is mainly motivated by surprises. The more an observation surprises us in terms of model behavior, the more we can learn from it. However, all assumptions inevitably increase our uncertainty about the model and, thus, lead to surprises when comparing the model response to observed data. The crux of a Bayesian analysis is that we can make any assumptions, approximations or simplifications in our model as long as we are realistic about the modeling uncertainties that emerge from it. This is especially difficult for deliberately conservative assumptions, because the introduced bias needs to be corrected within the chosen probabilistic error model. Moreover, this is also one of the reasons why complex models in a Bayesian setting are not

always better than simpler models: For complex models, modeling uncertainties are typically much more difficult to quantify than for simple models.

If the modeling uncertainties are quantified inappropriately (e.g., with a too large or too small spread, with the wrong mean, or with an unfit dependency model), the learning effect is increased when comparing the model response to the observed system response due to larger surprises. On its own, this is generally not a problem as more has indeed to be learned under unrealistic assumptions. On a higher level, inappropriate models can be identified by comparing different probabilistic models with each other through their evidence (assuming that there are any "good" probabilistic models in the set). However, on the level of the affected model, unrealistic assumptions cannot be identified, as the Bayesian framework is formulated mathematically consistent, conditional on the assumptions made. Thus, within the Bayesian framework, the probabilistic model has to somehow cope with unrealistic modeling choices. This has the following consequences: (i) The posterior of uncertain parameters can be estimated with too much confidence due to large surprises in learning. (ii) Some uncertain parameters might change their interpretation to account for unexpected behavior of the observed response. If the model to be learned is used later merely for prediction of the system response, only the first point is of relevance. If, however, the uncertain model parameters are the target of interest, the second point is crucial.

For example, consider an engineering model that predicts displacements and has the Young's modulus as an uncertain parameter. The target of interest is the probabilistic description of the Young's modulus, which can be improved through observed displacements. Let us assume that the predicted displacements are on average too large because of a poor model and this is not accounted for in the probabilistic model error. Thus, the observed system response will possibly be smaller than what the model expects. On the one hand, if the observed system response is indeed smaller than expected, the stochastic model will artificially increase the value of the Young's modulus to account for the "surprisingly" small observed displacement. On the other hand, if the observed system response is close to what is expected, the mean of the Young's modulus will not change much in the inference; whereas knowing about the model deficiencies suggests that the value of the Young's modulus should actually be decreased. In both cases, the uncertain model parameter does not represent the actual Young's modulus, but the uncertainty about an artificial Young's modulus that additionally incorporates global modeling errors.

1.2 Scope of the thesis

This work focuses on the reduction of uncertainties through measured or observed data within a Bayesian framework. The scope of this thesis is twofold:

numerical methods Non-intrusive methods for numerical Bayesian inference are discussed. The focus is on methods that generate samples from the posterior and provide an estimate of the evidence. Only approaches that have the potential to perform well for a large number of uncertain model parameters are investigated. Methods based on asymptotic approximations or approximate Bayesian computation are not considered. The selected Bayesian inference methods are BUS, TMCMC and nested sampling. Emphasis is put on the combination of BUS with Subset Simulation (BUS-SuS), and a modified variant of this approach is proposed. In this context, also Subset Simulation needs close attention, as it is the cornerstone of BUS-SuS.

probabilistic framework Requirements for a Bayesian framework that performs reasonably are introduced and discussed. Within this context, the philosophical standpoint about uncertainty and probability theory needs to be defined. The *Cox-Jaynes interpretation of probability* is adopted, and it is argued that uncertainty should be regarded as neither *subjective* nor *objective*, but as *personal*. The influence of assumptions in prior and likelihood is studied.

1.3 Outline

Chapter 2 starts with a discussion of the meaning of *uncertainty*. In Section 2.1 different views on uncertainty are discussed: objective vs. subjective, aleatory vs. epistemic, frequency, design, level of confidence. The inevitable personal aspect in the interpretation of uncertainty is emphasized. Section 2.2 continues the discussion on a more mathematical level. The fundamentals of *Probability Theory* are presented adopting the *Cox-Jaynes interpretation of probability*. The axioms of probability logic derived by [Cox, 1946] are stated, and compared to the Kolmogorov axioms. Cox-Jaynes probability theory is linked to the Bayesian interpretation of probability. Furthermore, alternative interpretations of probability are briefly discussed. The section finishes with extending *Probability Theory* to continuous probability spaces. Section 2.3 is about modeling of uncertainty. The theory and notion of stochastic/random variables, vectors of stochastic variables and stochastic processes/fields is presented. Section 2.4 provides a short overview of information theory; terms like *entropy* and *Kullback–Leibler divergence* are introduced.

Chapter 3 is about generating samples of a distribution. The Rosenblatt transformation, the Nataf transformation, rejection sampling and Markov chain Monte Carlo (MCMC) are

presented in Sections 3.1 – 3.4. Section 3.5 looks in more detail at MCMC sampling focusing on target distributions arising in structural reliability. The following MCMC algorithms are presented in Sections 3.5.4 – 3.5.7: component-wise Metropolis-Hastings (cwMH), conditional Metropolis-Hastings (CMH), conditional sampling in standard Normal space (CS), and directional conditional sampling (DCS). Example problems studied throughout this thesis are introduced in Section 3.5.3. By means of these numerical examples, it is demonstrated in Section 3.5.8 that the *CS algorithm* proposed in [Papaioannou et al., 2015] is a convenient choice for this type of problems. Mathematical reasons why this algorithm is an excellent choice for problems with many uncertain parameters are given in Section 3.5.6.2. The chapter concludes with a discussion of adaptive learning of the spread of the MCMC proposal distribution in Section 3.5.9.

Chapter 4 deals with probabilistic modeling based solely on prior knowledge – without including measurement/observations in the analysis. Such models are referred to as *probabilistic forward models*. The concept of a *stochastic model class* is introduced in Section 4.1. Thereafter, a simple Monte Carlo based approach to simulate the probabilistic model response is presented (Section 4.2), and the difference between credible and confidence intervals is discussed (Section 4.3). Section 4.4 provides an overview of reliability analysis.

Chapter 5 presents selected numerical methods for reliability analysis. First, an overview of commonly used reliability methods is given and the merits of working in an underlying standard Normal space are explained in Section 5.1. Next, in Section 5.2, Monte Carlo simulation is presented and a Bayesian post-processing strategy for Monte Carlo simulation is introduced. Section 5.3 is dedicated to Subset Simulation (SuS). A general overview of SuS is given and the implementation of the method is discussed. Thereafter, the uncertainty about the estimated probability of failure is investigated in detail by means of numerical examples.

Chapter 6 introduces and discusses the theory behind *Bayesian analysis*. Section 6.1 presents the problem that is to be solved in a Bayesian analysis. Section 6.2 discusses the meaning and interpretation of the *evidence* in a Bayesian framework. Bayesian model class selection and model averaging is presented in Section 6.3. The Bayesian modeling framework is comprehensively discussed in Section 6.4. The section starts with discussing the interpretation of data. Next, the concept of a *stochastic model class* previously introduced in Chapter 4 is extended. Thereafter, the objectivity of the likelihood function is briefly discussed. This is followed by a discussion of probabilistic modeling approaches for the prior. Hierarchic stochastic models are presented subsequently. Probabilistic models for error structures are investigated. Section 6.5 concludes the chapter with an overview of different ways to formulate the likelihood function.

Chapter 7 investigates numerical methods for Bayesian inference. The focus is on methods that generate samples from the posterior distribution and provide simultaneously an estimate

for the evidence. First, the investigated numerical examples are presented in Section 7.2. Section 7.3 discusses the BUS approach. A numerically more beneficial variant of the BUS approach is suggested. Furthermore, a post-processing step is proposed to correct the results of BUS simulations when the scaling constant c^{-1} is selected too small. In Section 7.4 a modified variant of the BUS approach, referred to as aBUS, is proposed that does not require the scaling constant c^{-1} as input. The nested sampling method is presented in Section 7.5, similarities between nested sampling and Subset Simulation are highlighted, and a modification of nested sampling is proposed. Section 7.6 introduces the original variant of the TMCMC moethod and proposes a modified TMCMC variant.

Chapter 8 gives concluding remarks, lists the main contributions of this thesis, and provides an outlook.

In Appendix A numerical descriptors of random variables are defined. In Appendix B some common probability distributions are introduced. In Appendix C some continuous maximum entropy probability distributions are listed. Appendix D theory of stochastic fields and random field discretization is introduced. Appendix E some proofs are given that where omitted in the main part of the thesis. In Appendix F selected parts of statistical data analysis are presented.

Chapter 2

Representation of Uncertainties

2.1 The concept of uncertainty

This section informally discusses the notion of *uncertainty*. The viewpoint taken in this section serves as basis to formally introduce *Cox-Jaynes interpretation of probability* in Section 2.2. A short overview of alternative approaches to quantifying uncertainty is given in Section 2.2.7.

2.1.1 Introduction

The mass of an object or the distance between two points are properties attached to the world, the universe - as is the speed of light or gravity. We might disagree about their exact value, but this has to be attributed to our imprecise measurement devices: In the real world the mentioned quantities are fixed, they are said to be deterministic.[2] We are just uncertain about their true value. Uncertainty, however, is not a property attached to the universe[3]. Uncertainty relates our imprecise state of knowledge to the universe (see e.g., [Lindley, 1975]), it expresses our level of confidence.

For example, we can measure the mass of an object, but we know that if we repeat the measurement, or if we use a different measuring device, we might measure a different value. Moreover, our measuring devices have only finite precision. Thus, we are uncertain about the real mass of the investigated object. Somebody else could have measured the same object using the same or a different measuring device. Even if they measured exactly the same number, their belief in the accuracy of the measurement might be different from ours. Essentially, their uncertainty about the mass of the object will very likely be different from our

[2] This statement does not hold at the level of *quantum mechanics*. However, in engineering we are typically interested in much larger scales and can, thus, consider such quantities as fixed.

[3] Again, this statement does not hold at the level of *quantum mechanics*.

uncertainty about its mass. We are both uncertain, but to a varying degree. Alternatively, somebody else might have done the measurement for us. In this case we might judge the quality of the measurement differently than the one performing the measurement. We even could have asked more than one person to measure the mass of the object for us. Consequently, uncertainty relates to the observer and his state of knowledge.

The observer can be any person: you, your neighbor, your colleague, your boss, your medical advisor, or anybody else. Moreover, the observer can also be a company, an institution, or the government. They all see the world with different eyes. However, uncertainty does not only depend on the observer alone, it also depends on the state of knowledge of the observer; i.e., his level of confidence. Knowledge in this context stands for information and data that is available to the observer, as well as past experiences that he or she has made. Uncertainty can be reduced by an increase in knowledge: by learning.

All in all, it is important to highlight the personal aspect of uncertainty. Unlike mass or length that are properties of the real world, uncertainty describes our link to the real world. Uncertainty expresses our beliefs. Uncertainty is personal. For an extensive treatise on the meaning of uncertainty that uses only a bare minimum of mathematics, the easy-to-read book of [Lindley, 2006] is recommended.

2.1.2 Classification of uncertainties

2.1.2.1 Objective vs. subjective

In literature (this in particular includes statisticians arguing against the Bayesian point of view), much effort is dedicated to whether an approach to quantify uncertainty is *objective* or *subjective*. However, such a discussion is often misleading (see Section 2.2.5): (i) A unique objective approach does not exist – multiple approaches can be considered objective. (ii) Taking a subjective approach still requires one to follow rationality and consistency – both can be interpreted as applying objective rules conditional on the imposed subjective point of view. [Lindley, 2006] advocates the term *personal* instead of differentiating between *subjective* and *objective*.

2.1.2.2 Aleatory and epistemic

Uncertainties are often characterized as either *aleatory* or *epistemic* (see [Der Kiureghian and Ditlevsen, 2009] for a discussion of this topic). An *epistemic* uncertainty can be reduced by additional information, whereas an *aleatory* uncertainty cannot be reduced. Such a classification is only meaningful if done for each analysis individually: Uncertainties considered *aleatory* in one analysis might be classified as *epistemic* in another. The classification depends

on what data we believe can (actually) be gathered to reduce the identified uncertainties.

For example, such a classification depends on whether a bridge that is to be constructed is assessed, or a bridge that has already been constructed is assessed. For a completed bridge, one can simply go out and take measurements of its material properties. Thus, uncertainties about the material properties can be reduced in this case, and are classified as *epistemic*. If the bridge has not yet been constructed, obtaining data about its material properties is more involved. Consequently, in this case uncertainties about the material properties of the future bridge are typically classified as *aleatory*.

However, if we know in which factory the concrete will be made, then this information could, in principle, reduce our uncertainties about the material properties of the future bridge. Therefore, the classification depends on the data that one expects to become available – and not solely on what uncertainties could, in principle, be reduced. After all, all uncertainties are essentially due to a lack of knowledge (see Section 2.1.1).

The characterization into *aleatory* and *epistemic* can be meaningful. However, in the experience of the author, it is also the cause of unnecessary confusion. Moreover, within a Bayesian framework, only the *data* and the *model* decide what uncertainties can actually be reduced.

2.1.2.3 Frequency, design and level of confidence

In Section 2.1.1 it is argued that uncertainty is personal and relates to imprecise knowledge or lack of information. A seemingly valid counter-argument is that the uncertainty associated with the following actions is usually not considered to be personal: *throw the dice, toss a coin, participate in the lottery, play the slot machine*. Indeed, many people would agree that if a six sided dice with sides labeled *1,2,3,4,5,6* is thrown, we will observe a *5* in one out of six cases on average. The same can be said about tossing a coin: we usually agree that each side of the coin is equally likely to face upwards after it is thrown. Can we thus conclude that in these cases, uncertainty should be viewed as a universal property attached to the respective action?

In order to answer the question above, let us first look at a classification used in [Gigerenzer, 2013]: Uncertainty originates either from *frequency*, *design* or *level of confidence*.

Frequency means that we look at events occurring repeatedly, and count how often a specified event occurs in a certain number of cases. The aim is to determine the *long run frequency* of an event. Quantification of uncertainty in terms of *frequency* is discussed in Section 2.2.6.

For example, the number of days with an average temperature below zero degrees Celsius divided by the total number of days observed gives a frequency. The data that we have can stem from a single year, two years, ten years, 50 years or even from a

longer period. This estimated frequency can then be used to express our uncertainty about how many days in the coming year will have an average temperature below zero.

However, the use of *frequencies* is not free of personal beliefs:

- Is the frequency measured in the past applicable for the future?
  For example: How about global warming? Is the frequency of recorded past flood events a good estimate to quantify the frequency of future floods?
  It clearly is a (personal) assumption to regard the underlying frequency as stationary.

- Is the measured frequency the true underlying frequency? (Assuming that such a true underlying frequency even exists at all.)
  For example: Two different observation periods of the same length will very likely not result in the same outcome.
  As we do not know the underlying frequency, but only an estimate of it, employing either the measured frequency or a different frequency is an assumption.

Consequently, the way we handle information about the observed frequency has a personal component attached to it: We make assumptions that we deem justified by what we believe. The observed frequencies serve as a basis to quantify our uncertainty.

Design means that something is specifically made/designed to behave randomly. This includes *dice*, *roulette*, and *slot machines*.
Let us look at a single *dice* and assume that it is actually perfect – which is hard to achieve in practice. Knowing about who is going to throw the *dice* and what throwing technique will be used might change our belief about the outcome. Assigning equal probability to each side of the *dice* should usually be a good approximation (if no evidence is available that suggests otherwise), but it is only an assumption and not a property associated with the *dice*. Similar arguments can be applied for *roulette*, and also the outcome of a coin toss depends on how the coin is thrown [Jaynes, 2003, Chapter 10.3].
The *slot machine* is a slightly different case: A *pseudo random number generator* (PRNG) is employed by the machine. Thus, knowing the source code and the seed value of the PRNG as well as the starting time and the time-interval at which the numbers are generated, the behavior of the machine is actually deterministic.

Level of confidence This case is discussed in Section 2.1.1.

Table 2.1: Fatality rates with respect to distance. (The numbers are subjected to uncertainty.)

event description	rate	base	reference
road deaths (total)	$5.8 \cdot 10^{-9}$	km (in the EU, 2014)	[Adminaite et al., 2015]
by car on road	$7 \cdot 10^{-9}$	km (in the EU, 2001/2002)	[ETSC, 2003]
by bus on road	$7 \cdot 10^{-10}$	km (in the EU, 2001/2002)	[ETSC, 2003]
by motorcycle/moped on road	$1.4 \cdot 10^{-7}$	km (in the EU, 2001/2002)	[ETSC, 2003]
by cycle on road	$5.4 \cdot 10^{-8}$	km (in the EU, 2001/2002)	[ETSC, 2003]
by foot on road	$6.4 \cdot 10^{-8}$	km (in the EU, 2001/2002)	[ETSC, 2003]
railway	$3.5 \cdot 10^{-10}$	km (in the EU, 2001/2002)	[ETSC, 2003]
airplane (within EU)	$3.5 \cdot 10^{-10}$	km (in the EU, 2001/2002)	[ETSC, 2003]

2.1.3 Uncertainties related to individuals vs. populations

2.1.3.1 Statistical analysis of the behavior of a population

The behavior of a specified population within a given time-interval (for which data is available) can be assessed by means of statistical data analysis. Exemplary questions to be answered are: What is the average probability of a European citizen to die in a car accident? What is the average probability of a citizen to die at the age of 30? What is the average probability that an industrially manufactured device is defective?

Assessing the average behavior associated with an entire population can often be tackled well by means of standard statistical techniques in an objective manner, provided that sufficient data to capture the quantities of interest is available. For example, if the goal is to assess the fatality risk of traveling, one can count the total number of fatalities within a certain time-interval and divide this number by the total distance traveled by the entire reference population. Selected fatality risks with respect to traveled distance are listed in Table 2.1. We can also assess the average fatality risk of performing a certain action for a specified duration. Selected fatality rates with respect to time are listed in Table 2.2. Note that the fatality rates listed in Tables 2.1 and 2.2 are based on a certain population and time-interval. This means: (i) The fatality rates can (and do) change over time. (ii) The fatality rates depend on the underlying population. For example, in Fig. 2.1 the death rates (per year) of German males and females are listed as a function of the age. To compare the different fatality rates listed in Table 2.2, the hourly rates should be converted to annual fatality rates by multiplying the hourly fatality rate with the yearly exposure; i.e., the average time in *hours per year* that an individual is performing (or, exposed to) the indicated action. Annual fatality rates that are typically accepted are around 10^{-5} [Melchers, 1999, Table 2.6, Broad indicators of tolerable risk]. Events with annual fatality rates below 10^{-6} are not of great concern to an average individual [Melchers, 1999, Table 2.6, Broad indicators of tolerable risk].

It is important to highlight that if we perform an analysis based on data from an entire

Table 2.2: Fatality rates with respect to time. (The numbers are subject to uncertainty.)

event description	rate	base	reference
lowest death rate females (age 10)	$7.3 \cdot 10^{-5}$	a (in DE, 2010/12)	[DESTATIS, 2015]
lowest death rate population (age 10)	$8.4 \cdot 10^{-5}$	a (in USA, 2011)	[Arias, 2011]
lowest death rate males (age 9)	$9.1 \cdot 10^{-5}$	a (in DE, 2010/12)	[DESTATIS, 2015]
lowest death rate in population (girls age 10 to 14)	$1 \cdot 10^{-4}$	a (in NL, around 1990)	[Paté-Cornell, 1994; Ale, 1991]
death rate infant (1st year)	$3.4 \cdot 10^{-3}$	a (in DE, 2010/12)	[DESTATIS, 2015]
due to structural failure	$2 \cdot 10^{-11}$	a	[Melchers, 1999, Table 2.5]
at work	$7 \cdot 10^{-9}$	h (in EU, 2001/02)	[ETSC, 2003]
railway	$2 \cdot 10^{-8}$	h (in EU, 2001/02)	[ETSC, 2003]
by bus on road	$2 \cdot 10^{-8}$	h (in EU, 2001/02)	[ETSC, 2003]
at home	$2.3 \cdot 10^{-8}$	h (in EU, 2001/02)	[ETSC, 2003]
hiking: via ferrata	$1 \cdot 10^{-7}$	h (DAV, 2013)	[DAV, 2014]
airplane (within EU)	$1.6 \cdot 10^{-7}$	h (in EU, 2001/02)	[ETSC, 2003]
hiking: mountaineering	$2 \cdot 10^{-7}$	h (DAV, 2013)	[DAV, 2014]
by car on road	$2.5 \cdot 10^{-7}$	h (in EU, 2001/02)	[ETSC, 2003]
by foot on road	$2.5 \cdot 10^{-7}$	h (in EU, 2001/02)	[ETSC, 2003]
road deaths (total)	$2.8 \cdot 10^{-7}$	h (in EU, 2001/02)	[ETSC, 2003]
hiking: high-altitude mountain tours	$4 \cdot 10^{-7}$	h (DAV, 2013)	[DAV, 2014]
ski hiking (backcountry skiing)	$4 \cdot 10^{-7}$	h (DAV, 2013)	[DAV, 2014]
by cycle on road	$7.5 \cdot 10^{-7}$	h (in EU, 2001/02)	[ETSC, 2003]
alpine climbing	$1.4 \cdot 10^{-6}$	h (DAV, 2013)	[DAV, 2014]
by motorcycle/moped on road	$4.4 \cdot 10^{-6}$	h (in EU, 2001/02)	[ETSC, 2003]

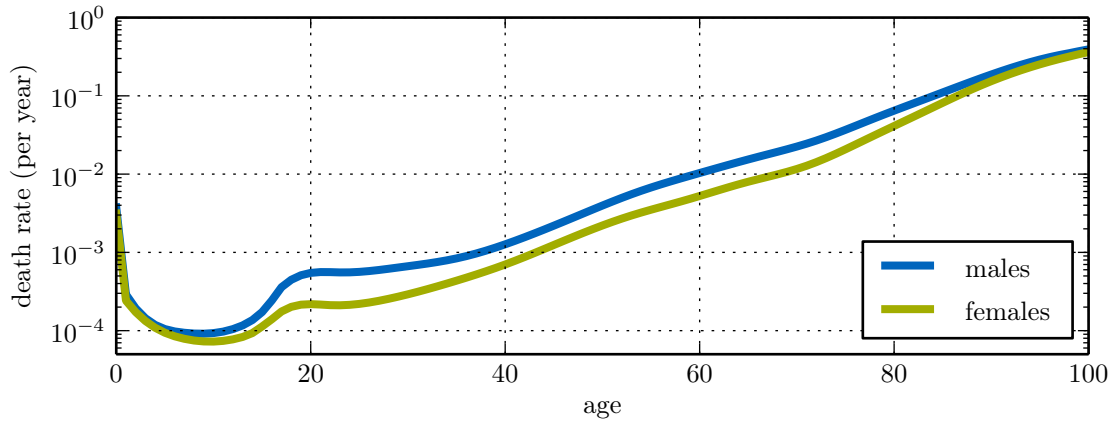


Figure 2.1: Death rates per year in 2010/12 of males and females for different age in Germany. Data taken from [DESTATIS, 2015].

population, the results are conditional to the entire population – which does not necessarily hold for a specific individual member of that group.

2.1.3.2 Stochastic assessment of individuals

Assessing the uncertainties associated with an individual[1] is more involved than looking at the statistics of an entire population. The main reason is that there is usually information available which distinguishes the individual from the "average" population. Exemplary questions to be answered are:

- *What is the probability that Jane Doe dies within the next year?*
  The probability depends on the age, gender, country of residence, social state, job, lifestyle, hobbies, ... of Jane Doe. Reliable statistical data that exactly fits to Jane Doe is not available. The more we know about Jane Doe, the more involved the task becomes.
  The associated event will happen at most once: Either she dies within the next year, or she does not die. As a sufficient data-basis is not available, the assessment will be highly personal: If you ask this question to different people, you will very likely get quite diverse answers.

- *What is the probability of John Doe to be involved in a car accident on his way to work tomorrow?*
  Similar to the previous question, we cannot answer this question based on an available data-basis: Where does he live and work? What car does he drive? What is his driving experience? Which mood is he in? Is he sober? How long did he sleep the previous night(s)?

- *What is the probability that the new building on Main Street will become unusable within the next 50 years?*
  What is the building used for? What is the structure of the building? Where is the building located? What live-load is acting on the structure? What material was used for construction? What are the environmental conditions that have an influence?

For this type of problems, the outcome of the analysis will be based on the state of knowledge and experience of the person conducting the analysis.

[1] "*Individual*" in this context is not restricted to people, but can be any particular member out of a larger group/population; e.g., an engineering structure, a machine, a hydrological catchment.

2.2 Probability Theory

2.2.1 Introduction

The *Cox-Jaynes interpretation of probability* [Cox, 1946; Jaynes, 2003] (in the following referred to as *Probability Theory*) provides a rigorous foundation for stochastic modeling and Bayesian updating [Beck, 2010]. Probability Theory extents *Boolean algebra* for quantification of *plausible reasoning* under *incomplete information*. Probability Theory is also referred to as *probability logic*. The underlying axioms of the theory were stated by [Cox, 1946]. The work of [Jaynes, 2003] is a treatise on the theory that is illustrated by applications. Moreover, [Jaynes, 2003] comprehensively compares *Probability Theory* to the *frequentist* point of view and resolves published paradoxes related to the axioms derived by [Cox, 1946]. Another noteworthy contribution that influenced me when writing this chapter is the work of [Beck, 2010]: It briefly summarizes the work of [Cox, 1946] and [Jaynes, 2003], and introduces terminology useful for Bayesian system identification.

The key idea of Probability Theory is to interpret the probability $\Pr[b|a]$ as the *plausibility* of proposition[1] b based on the information in proposition a. Note that $\Pr[b|a]$ does not mean that proposition a has to be *true*, since it is only conditionally asserted.

Besides $\Pr[b|a]$ the following notation is used:

- $\Pr[\overline{b}|a]$ denotes the probability that proposition b is *false* given the information in proposition a.
- $\Pr[b \wedge c|a]$ denotes the probability that both proposition b and proposition c are *true* given the information in proposition a.
- $\Pr[b \vee c|a]$ denotes the probability that at least proposition b or proposition c is *true* given the information in proposition a.

2.2.2 Axioms of probability logic

Probability $\Pr[b|a]$ is a measure to quantify the plausibility of proposition b being *true* conditional on proposition a being *true*. As there is no natural scale for probability, arbitrary measures can be used to quantify plausibility [Beck, 2010; Cox, 1946]. Typically, a linear relation is used to relate plausibility and probability: e.g., if $b|a$ is as likely as $\overline{b}|a$, then $\Pr[b|a] = 0.5$; if $b|a$ is twice as likely as $\overline{b}|a$, then $\Pr[b|a] = 2/3$; if $b|a$ is three times as likely as $\overline{b}|a$, then $\Pr[b|a] = 0.75$. For alternative measures of plausibility, see Section 2.2.3.

[1] *proposition* in this context refers to a statement that is either *true* or *false*.

For the linear relation between plausibility and probability, the axioms of probability logic derived by [Cox, 1946] are [Beck, 2010]:

(P1) $\Pr[b|a] \geq 0$ (by convention)

(P2) $\Pr[\bar{b}|a] = 1 - \Pr[b|a]$ (negation function)

(P3) $\Pr[b \wedge c|a] = \Pr[c|b \wedge a] \Pr[b|a]$ (conjunction function)

Based on (P1)-(P3), the following relations can be derived [Beck, 2010, 2008; Cox, 1946]:

(P4a) $\Pr[b|b \wedge a] = 1$

(P4b) $\Pr[\bar{b}|b \wedge a] = 0$

(P4c) $\Pr[b|a] \in [0, 1]$

(P5a) $\Pr[c|(c \Rightarrow b) \wedge a] \leq \Pr[b|(c \Rightarrow b) \wedge a]$
where $(c \Rightarrow b)$ is used to denote that c is contained in b.

(P5b) $\Pr[c|(c \Leftrightarrow b) \wedge a] = \Pr[b|(c \Leftrightarrow b) \wedge a]$
where $c \Leftrightarrow b$ means that c is equivalent to b.

(P6) $\Pr[b \vee c|a] = \Pr[b|a] + \Pr[c|a] - \Pr[b \wedge c|a]$

Let proposition a state that propositions $b_1, \ldots, b_N$ are mutually exclusive and collectively exhaustive (i.e., one and only one of propositions $b_1, \ldots, b_N$ can be *true*), then:

(P7a) $\Pr[c|a] = \sum_{n=1}^{N} P[c \wedge b_n|a]$ (marginalization theorem)

(P7b) $\Pr[c|a] = \sum_{k=1}^{N} \Pr[c|b_k \wedge a] \Pr[b_k|a]$ (total probability theorem)

(P7c) $\Pr[b_k|c \wedge a] = \frac{\Pr[c|b_k \wedge a] \Pr[b_k|a]}{\sum_{n=1}^{N} \Pr[c|b_n \wedge a] \Pr[b_n|a]}$ (Bayes' theorem[1])
for $k = 1, \ldots, N$.

In a "Bayesian" context, $\Pr[b_k|a]$ is referred to as the *prior*, $\Pr[c|b_k \wedge a]$ is the *likelihood*, and $\Pr[b_k|c \wedge a]$ is the *posterior*.

Note that *Probability Theory* can be viewed as an extension of *Boolean algebra* [Beck, 2010]: Whereas *Boolean algebra* requires complete information (we have to assign a value of either *true* or *false* to the involved propositions), *Probability Theory* can handle incomplete information (we are "merely" required to specify a plausibility that a specific proposition is *true*).

Derivations of axioms (P4a) to (P7c) are given in Appendix E.1.

[1] Bayes' theorem is named after *Thomas Bayes*. More information about Thomas Bayes is presented in Section 2.2.5.

2.2.3 Quantifying plausibility

In [Cox, 1946], Cox derives the axioms of probability logic without making strong assumptions about the scale of probability. Let $\mathrm{P}[b|a] = B$ if proposition b is certain[1] given a; and let $\mathrm{P}[b|a] = 0$ if proposition b is impossible given a, where $B \in \{x \in \mathbb{R} | x > 0\}$. [Cox, 1946] assumes that functions $C : [0, B] \times [0, B] \to [0, B]$ and $N : [0, B] \to [0, B]$ exist such that

$$\mathrm{P}[b \wedge c|a] = C\left(\mathrm{P}[c|b \wedge a], \mathrm{P}[b|a]\right) \tag{2.1}$$

$$\mathrm{P}[\bar{b}|a] = N\left(\mathrm{P}[b|a]\right) \tag{2.2}$$

[Cox, 1946] shows that C and N must have the form:

$$C(u, v) = \phi^{-1}\left(\frac{\phi(u)\phi(v)}{D}\right) \tag{2.3}$$

$$N(u) = \phi^{-1}\left(D - \phi(u)\right) \tag{2.4}$$

where $\phi(\cdot)$ is a continuous and strictly increasing function with $\phi(0) = 0$ and $\phi(B) = D$ (compare [Beck, 2010]), $\phi^{-1}(\cdot)$ is the inverse function of $\phi(\cdot)$ and $u, v \in [0, B]$. If $\phi(u) = u$, $B = 1$ and $D = \phi(B) = 1$, then we get the axioms (P1)-(P3) presented in Section 2.2.2.

Example 2.1. *Exponential relation for plausibility calculation*:
Let $\phi(u) = \exp(u) - \exp(0)$, $B > 0$ and $D = \phi(B) = \exp(B) - \exp(0)$.
The inverse function of $\phi(\cdot)$ is:

$$\phi^{-1}(\phi) = \ln\left[\phi + \exp(0)\right]$$

The negation function takes the form:

$$\begin{aligned} N(u) &= \ln\left[D - \phi(u) + \exp(0)\right] \\ &= \ln\left[\exp(B) - \exp(u) + \exp(0)\right] \end{aligned}$$

The validity of the negation function can be verified by:

$$\begin{aligned} N(N(u)) &= \ln\left[\exp(B) - \left(\exp(B) - \exp(u) + \exp(0)\right) + \exp(0)\right] \\ &= \ln\left[\exp(u)\right] = u \end{aligned}$$

The conjunction function takes the form:

$$C(u, v) = \ln\left[\frac{\left(\exp(u) - \exp(0)\right)\left(\exp(v) - \exp(0)\right)}{D} + \exp(0)\right]$$

It is trivial to show that if $u = 0$ or $v = 0$, then $C(u, v) = \ln[\exp(0)] = 0$. Moreover, for $v = B$: $C(u, B) = u$.

[1] To highlight this ambiguity in quantifying plausibility, the notation $\mathrm{P}[b|a]$ is used in this section to denote a measure for the plausibility of proposition b being *true* conditional on proposition a being *true*.

For such a measure of plausibility, the axioms of probability theory are:

(P1*) $\mathrm{P}[b|a] \geq 0$ (by convention)

(P2*) $\mathrm{P}[\overline{b}|a] = \ln\left[\exp(B) - \exp(\mathrm{P}[b|a]) + \exp(0)\right]$ (negation function)

(P3*) $\mathrm{P}[b \wedge c|a] = \ln\left[\frac{(\exp(\mathrm{P}[c|b\wedge a]) - \exp(0))(\exp(\mathrm{P}[b|a]) - \exp(0))}{D} + \exp(0)\right]$ (conjunction function)

2.2.4 Kolmogorov axioms

Instead of using the *Cox-Jaynes interpretation of probability* (Section 2.2.2), the foundation of probability theory is commonly defined in terms of *Kolmogorov axioms* [Kolmogorov, 1933]. Unlike *Cox-Jaynes interpretation of probability*, the Kolmogorov axioms are neutral on how to interpret probability [Beck, 2010].

2.2.4.1 Definition

Let (Ω, S, P) be a so-called probability space, where Ω denotes the sample space which contains all possible outcomes of a *model*[1], S is a set of events that contains zero or more outcomes (S is also referred to as σ-algebra), and P is a probability measure[2] that assigns probabilities to the events $A \in S$. The probability function P obeys the following rules (known as *Kolmogorov axioms*):

(K1) $\mathrm{P}[A] \geq 0$ (non-negativity)

(K2) $\mathrm{P}[\Omega] = 1$ (normalization)

(K3) $\mathrm{P}[A_1 \vee A_2 \vee \ldots] = \sum \mathrm{P}(A_i)$ (countable additivity)
if $A_1, A_2, \ldots$ are mutually exclusive.

Note the difference between an *event* and a *proposition*: A *proposition* is a statement that is either *true* or *false*; probability in this context is the plausibility that we assign to the statement being *true*. An *event* is defined as a set of outcomes of a model; i.e., a subset of the sample space. For example, the case "the proposition is *true*", can be interpreted as an *event*. However, the notion of an *event* is more general than the term *proposition*. Contrary to a *proposition*, an *event* is – in principle – repeatable.

[1] The term *model* can refer to any type of model. In this contribution, the term *model* is usually employed to refer to a model of a real-world system. In mathematical and statistical literature, often the term *experiment* is used instead, when introducing the notion of a probability space.

[2] The notion $\mathrm{P}(A)$ is used instead of $\Pr(A)$ to highlight the difference to *Cox-Jaynes interpretation of probability*. In *Cox-Jaynes interpretation of probability* (Section 2.2.2), probability is always defined conditional on a proposition.

2.2.4.2 Kolmogorov axioms in terms of Probability Theory

In this contribution, *Cox-Jaynes interpretation of probability* is preferred over *Kolmogorov's probability theory*. [Cox, 1946] defines all probabilities conditional on a proposition a that represents the information, knowledge and assumptions behind the probabilistic model. This interpretation is helpful for uncertainty quantification in combination with engineering models and for Bayesian inference. Contrary to that, *Kolmogorov's probability theory* requires the mathematical construct of a probability space (Ω, S, P) that does not explicitly[1] condition all probabilities on proposition a.

[Beck, 2010] shows that the *Kolmogorov axioms* can be deduced from the axioms of probability theory introduced by [Cox, 1946]: Let x denote a stochastic variable[2] that can take values in X. Furthermore, let A be a subset of X. The probability $\mathrm{P}(A)$ in *Kolmogorov's probability theory* can be expressed as $\Pr[x \in A|\mathcal{M}]$ using *Cox-Jaynes interpretation of probability*. Here proposition $\mathcal{M}$ states that $x \in X$, and contains the probability model used to represent x. The first axiom (K1) directly follows from (P1). The second axiom (K2) is obtained through (P2) with $P(X) = \Pr[x \in X|\mathcal{M}] = 1$. The third axiom (K3) follows from (P6), requiring the events to be mutually exclusive[3]. A discussion of the *Kolmogorov axioms* in light of *Cox-Jaynes interpretation of probability* can be found in [Jaynes, 2003, Appendix A].

2.2.4.3 Conditional probability

In Kolmogorov's approach, *conditional probability* $\mathrm{P}[A|B]$ is defined as:

$$\mathrm{P}[A|B] = \frac{\mathrm{P}[A \wedge B]}{\mathrm{P}[B]} \tag{2.5}$$

for $\mathrm{P}[B] > 0$. Contrary to that, in *Cox-Jaynes interpretation of probability* all probabilities are interpreted as conditional probabilities. Consequently, the definition in Eq. (2.5) appears directly as axiom (P3) in *Cox-Jaynes interpretation of probability* [Beck, 2010].

[1] One can argue that the probability space (Ω, S, P) incorporates proposition a; i.e., that the elements of (Ω, S, P) take the available information, knowledge and all assumptions into account. However, *Cox-Jaynes interpretation of probability* and in particular the notation introduced in Section 2.2.2 makes this dependency explicit. This is especially helpful if different competing probabilistic models are compared.

[2] The term *stochastic variable* is introduced in Section 2.3.1.

[3] Actually, from (P6) we get only *finite additivity*, and not *countable additivity* as required by (K3). Note that *finite additivity* is more general (i.e., less restrictive) than *countable additivity* and, thus, more complex to handle. The issue whether *finite additivity* or *countable additivity* should be required is controversial (see e.g., [Bingham, 2010; Easwaran, 2013; De Finetti, 2008]). From a theoretical (mathematical) point of view, it is indeed important to distinguish between finite and countable additivity. Problems with the weaker assumption of finite additivity can arise in combination with infinite sets. However, as [Jaynes, 2003] points out, such problems arise because *Cox-Jaynes interpretation of probability* is violated; e.g., by the use of improper prior distributions in Bayesian inference. As long as we adhere to *Cox-Jaynes interpretation of probability*, the discussion about *finite additivity* or *countable additivity* is dispensable. This is especially true for typical problems of practical relevance.

2.2.5 Bayesian interpretation of probability

In *Bayesian probability theory*, probability represents a degree of belief in a proposition[1] [Beck, 2010]. Consequently, probabilities appear as conditional probabilities; i.e., probabilities are viewed as conditional on a proposition. In this regard, *Cox-Jaynes interpretation of probability* is consistent with the *Bayesian point of view*.

Bayesian probability theory is named after the English mathematician and Presbyterian minister *Thomas Bayes* (1701-1761)[2] [Bellhouse, 2005]. Thomas Bayes wrote a manuscript that discusses a special case[3] of what later became known as *Bayes' theorem*. This manuscript [Bayes and Price, 1763] was edited and published posthumously by Richard Price[4] [Bellhouse, 2005]. Pierre-Simon Laplace (1749–1827) independently generalized the theorem proposed in [Bayes and Price, 1763], and established the foundation for what is referred to as *Bayesian probability theory* today [Stigler, 1986; Beck, 2014].

The name *Bayesian probability theory* is due to the many ways *Bayes' theorem* can be applied if probability is interpreted as *degree of belief*. Compared to classical statistics, *Bayes' theorem* plays a major role in Bayesian probability theory (see Chapter 6). However, Bayesian probability theory is a particular interpretation of the concept of probability and does not necessarily require application of *Bayes' theorem*; this is why e.g., [Edwards et al., 1963] considers the nomenclature somewhat inadequate.

There is no clear consensus on how exactly the *Bayesian interpretation of probability* is defined. Typically, the philosophy of Bayesian probability theory is classified in two main branches: *objective* and *subjective* Bayesian probability. (i) In the so-called *objective* view, probability is interpreted as a reasonable belief: Everyone with the same knowledge would come to the same conclusion. However, within this category there is no clear consensus on what it actually means to be objective. Most discussions are about how to formulate prior information in an objective manner. Some (e.g., [Jeffreys, 1946, 1998; Hartigan, 2012]) suggest the use of weakly-informative or uninformative priors, which can lead to problems (theoretical as well as numerical) – especially in combination with improper[5] prior distributions. Others advocate prior distributions selected based on the Principle of Maximum Information Entropy

[1] The term *proposition* is introduced in Section 2.2.1.

[2] It is most likely that Thomas Bayes was born in 1701, however, in general all that is known is that he was born between July 1701 and April 1702 [Bellhouse, 2005]. He died on April 7, 1961, at the age of 59 [Bellhouse, 1988].

[3] Thomas Bayes considered the case where *prior* and *posterior* follow a *Beta distribution* and the data comes from *Bernoulli trials*.

[4] *Richard Price* (1723-1791) was a British pastor, political philosopher and mathematician, and a friend of *Thomas Bayes* [Bellhouse, 2005]. He is also considered an intimate friend of *Benjamin Franklin*, and was visited by other *Founding Fathers of the United States* such as John Adams, Thomas Jefferson, and Thomas Paine. His most famous achievement was a pamphlet that he published in 1776, supporting British North America in the American War of Independence. This pamphlet is said to have contributed to the Americans declaring their independence in 1776. [Richard Price. (n.d.). In *Wikipedia*. Retrieved June 7, 2016, from `https://en.wikipedia.org/wiki/Richard_Price`]

[5] An improper prior distribution integrates to infinity.

[Jaynes, 1957, 2003], which depends on the chosen parametrization of the problem. Seemingly objective choices (for both the prior and the likelihood) will be discussed in Chapter 6. (ii) In the *subjective* view, probability is regarded as a personal belief: Probability is constrained by rationality and coherence, but can vary within those constraints. Here, probabilities are typically expressed and interpreted in terms of betting odds. The subjective point of view is usually based on the work of [De Finetti, 1964; Ramsey, 1931; De Finetti, 2008]. For an introduction to subjective Bayesian probability theory, see [Kadane, 2011].

[Lindley, 2006] notes that the discussion about *subjective* or *objective* Bayesian probabilities is often not helpful and can be potentially misleading; he suggests to use the expression *personal* instead. Such an interpretation of uncertainty[1] goes hand-in-hand with *Cox-Jaynes interpretation of probability*[2]. The angle on *Bayesian probability theory* taken in this contribution is: We have to admit that there is not a single objective approach to interpreting probability, as objectivity depends on information and experience (i.e., knowledge). People with different states of knowledge will inevitably consider different approaches to be objective[3]. However, conditional on our knowledge, we should aim to remain objective in our choices. As thus, the *personal* aspect of probability is central in the Bayesian point of view. Such an interpretation of *Bayesian probability theory* is in accordance with e.g., [Beck, 2008, 2010; Vanik et al., 2000; Beck and Yuen, 2004; Muto and Beck, 2008; Cheung and Beck, 2009, 2010; Zuev et al., 2012; Beck, 2014].

2.2.6 The frequentist interpretation of probability

The *frequentist* interpretation of probability associates probability $\mathrm{P}[A]$ with the long run average (i.e., the frequency) that event A occurs:

$$\mathrm{P}[A] = \lim_{N\to\infty} \frac{H}{N} \tag{2.6}$$

where N denotes the number of trials, and H is the number of times event A occurred in N trials. One problem of this interpretation is that the exact frequency of event A cannot be observed as N will always be finite in practice. Note that a strict association of frequentism with limiting frequencies is considered outdated by [Bingham, 2010].

[1] see Section 2.1

[2] *Cox-Jaynes interpretation of probability* is sometimes considered to be in the category of *objective* Bayesian probability. However, it should more appropriately be regarded as in between the *objective* and *subjective* point of view; as it accentuates the personal aspect of probability besides advocating objectivity. Moreover, note that improper prior distributions (that are considered as an objective Bayesian approach) are not in accordance with *Cox-Jaynes interpretation of probability*.

[3] Consider, for example, a scientist. It is commonly believed that scientists needs to be objective when doing research. However, his work is conditional on the underlying theory that he employs. If a flaw in the employed theory is detected, the strategy that is to be considered objective might change conditional on the new information. Nevertheless, his approach has to be considered objective, conditional on the knowledge he had at the time conducting the research.

The main differences between the frequentist interpretation and the view on probability introduced in Section 2.2.2 are:

- The frequentist interpretation postulates that the event A can be repeated arbitrarily often. If event A is unrepeatable or if no data is available to assign a frequency to event A, the probability of A cannot be specified.
- In frequentism, probability is regarded as a property of the *universe*[1] that can be measured only imprecisely from a finite number of observations.

See also [Beck, 2014] for critical remarks regarding the frequentist interpretation from a Bayesian perspective. From a Bayesian perspective, observed frequencies should be used in a Bayesian framework to express our knowledge conditional on this information – and not to directly associate probabilities with frequencies.

2.2.7 Overview: different interpretations of probability/uncertainty

The two most widely spread and discussed views on the interpretation of probability are the *frequentist* and the *Bayesian* point of view. An overview of different interpretations of probability is presented in [Hájek, 2012]. In the following, some interpretations of probability/uncertainty are briefly presented:

Bayesian point of view In this work, uncertainty is quantified according to *Cox-Jaynes interpretation of probability* (see Sections 2.2.1 and 2.2.2). This interpretation takes a Bayesian viewpoint on uncertainties (see Section 2.2.5): Probabilities represent a degree of belief in a proposition. Within the category of Bayesian probability interpretations, typically two classes are distinguished: the *objective* and the *subjective* Bayesian approach. In Section 2.2.5, it is suggested that regarding probabilities as *personal* (instead of as *objective* or *subjective*) can be more appropriate with respect to *Cox-Jaynes interpretation of probability*.

Frequentist point of view In Section 2.2.6, the *frequentist* interpretation of probability is presented, where probability is interpreted as the frequency of an event. This approach has problems with events that are unrepeatable.

Propensity probability Probability is interpreted as a physical property (propensity) [Popper, 1957]. Long run frequencies are viewed as the result of a concatenation of single-case probabilities (known as propensities). This interpretation of probability is motivated by problems arising in quantum mechanics.

[1] see discussion in Section 2.1.1.

Fuzzy logic Fuzzy logic [Zadeh, 1965] does not use *probability* to represent uncertainty. Fuzzy logic is an extension of many valued logic[1] where an object can belong to multiple classes. From a Bayesian viewpoint, a proposition is either *true* or *false*, and probability represents our belief about the actual state of the proposition. Contrary to that, fuzzy logic expresses how much a quantity is in a certain set.

2.2.8 Continuous probability spaces

2.2.8.1 Introduction

Let $X \in \mathbb{R}$ be a continuous quantity of interest; X is interpreted as a fixed[2] quantity that we do not know. Let the proposition F be defined as $x \leq X$, were $x \in \mathbb{R}$. The probability that proposition F is *true* is denoted as

$$P_{X|\mathcal{M}}(x|\mathcal{M}) = \Pr(x \leq X|\mathcal{M}) = \Pr(F|\mathcal{M}) \tag{2.7}$$

where $\mathcal{M}$ represents the available information/knowledge. $P_{X|\mathcal{M}}(\cdot|\mathcal{M})$ is referred to as *cumulative distribution function* (CDF) for[3] X. Furthermore, let proposition W be defined as $a \leq X < b$, with $a \leq b$. It is straight-forward to show that $\Pr(W|\mathcal{M}) = P_{X|\mathcal{M}}(b|\mathcal{M}) - P_{X|\mathcal{M}}(a|\mathcal{M})$. If $P_{X|\mathcal{M}}(\cdot|\mathcal{M})$ is continuous and differentiable, we can write

$$\Pr(W|\mathcal{M}) = \int_a^b p_{X|\mathcal{M}}(x|\mathcal{M})\,\mathrm{d}x = P_{X|\mathcal{M}}(b|\mathcal{M}) - P_{X|\mathcal{M}}(a|\mathcal{M}) \tag{2.8}$$

with

$$p_{X|\mathcal{M}}(x|\mathcal{M}) = \frac{\mathrm{d}P_{X|\mathcal{M}}(x|\mathcal{M})}{\mathrm{d}x} \tag{2.9}$$

$p_{X|\mathcal{M}}(x|\mathcal{M})$ is referred to as the *probability density function* (PDF) for[4] X.

2.2.8.2 Continuous one-dimensional prior

Let proposition A state that $x \leq X < x + \mathrm{d}x$, where $\mathrm{d}x$ represents an infinitely small change in x. Thus, $\Pr(A|\mathcal{M}) = p_{X|\mathcal{M}}(x|\mathcal{M}) \cdot \mathrm{d}x$. Based on this, the *total probability theorem* (rule

[1] Contrary to *boolean logic* where only two states exist (e.g., *true* and *false*), there can be more than two possible states in *many valued logic*.

[2] Note that X is not random/varying in reality, only our knowledge about X is uncertain. This distinction is important (compare [Jaynes, 2003, Chapter 4]).

[3] Often, the expressions "CDF for X" and "CDF of X" are used interchangeably. [Jaynes, 2003, Chapter 4] suggests to use "CDF for X", as quantity X is not uncertain – only our knowledge about X is uncertain. [Jaynes, 2003, Chapter 4] states that not distinguishing between "of" and "for" has let to paradoxes.

[4] As in case of the CDF, "PDF for X" and "PDF of X" are often used interchangeably. [Jaynes, 2003, Chapter 4] suggests to avoid the use of "PDF of X" – as as quantity X itself is not uncertain.

(P7b) in Section 2.2.2) can be transformed as:

$$\Pr(D|\mathcal{M}) = \int_{\mathbb{R}} \Pr(D|A \wedge \mathcal{M}) \cdot \Pr(A|\mathcal{M})$$
$$\Pr(D|\mathcal{M}) = \int_{\mathbb{R}} \Pr(D|(X = x) \wedge \mathcal{M}) \cdot p_{X|\mathcal{M}}(x|\mathcal{M})\,\mathrm{d}x \tag{2.10}$$

Equivalently, *Bayes theorem* (rule (P7c) in Section 2.2.2) becomes:

$$\Pr(A|D \wedge \mathcal{M}) = \frac{\Pr(D|A \wedge \mathcal{M}) \cdot \Pr(A|\mathcal{M})}{\Pr(D|\mathcal{M})}$$
$$\Pr(A|D \wedge \mathcal{M}) = \frac{\Pr(D|A \wedge \mathcal{M}) \cdot p_{X|\mathcal{M}}(x|\mathcal{M})}{\Pr(D|\mathcal{M})}\,\mathrm{d}x \tag{2.11}$$

From Eq. (2.11) it follows that $\int_{\mathbb{R}} \Pr(A|D \wedge \mathcal{M}) = 1$. Thus, we can write $\Pr(A|D \wedge \mathcal{M}) = p_{X|D,\mathcal{M}}(x|D, \mathcal{M})\,\mathrm{d}x$. Consequently, Eq. (2.11) becomes:

$$p_{X|D,\mathcal{M}}(x|D, \mathcal{M}) = \frac{\Pr(D|A \wedge \mathcal{M}) \cdot p_{X|\mathcal{M}}(x|\mathcal{M})}{\Pr(D|\mathcal{M})} \tag{2.12}$$

2.2.8.3 Continuous one-dimensional prior and likelihood

Let proposition D state that $h \leq f(X) < h + \mathrm{d}h$, where $f : \mathbb{R} \to \mathbb{R}$. Thus, $\Pr(D|A \wedge \mathcal{M}) = p_{H|X,\mathcal{M}}(h|x, \mathcal{M}) \cdot \mathrm{d}h$, with $H = f(X)$. Starting from Eq. (2.10), the *total probability theorem* can then be transformed as:

$$\Pr(D|\mathcal{M}) = \left(\int_{\mathbb{R}} p_{H|X,\mathcal{M}}(h|x, \mathcal{M}) \cdot p_{X|\mathcal{M}}(x|\mathcal{M})\,\mathrm{d}x \right) \mathrm{d}h \tag{2.13}$$

The integration of Eq. (2.13) over $\mathbb{R}$ with respect to h gives by definition 1. Consequently, we can write $\Pr(D|\mathcal{M}) = p_{H|\mathcal{M}}(h|\mathcal{M}) \cdot \mathrm{d}h$; and, Eq. (2.13) becomes:

$$p_{H|\mathcal{M}}(h|\mathcal{M}) = \int_{\mathbb{R}} p_{H|X,\mathcal{M}}(h|x, \mathcal{M}) \cdot p_{X|\mathcal{M}}(x|\mathcal{M})\,\mathrm{d}x \tag{2.14}$$

Bayes theorem (Eq. (2.12)) can be written as:

$$p_{X|H,\mathcal{M}}(x|h, \mathcal{M}) = \frac{p_{H|X,\mathcal{M}}(h|x, \mathcal{M}) \cdot p_{X|\mathcal{M}}(x|\mathcal{M})}{p_{H|\mathcal{M}}(h|\mathcal{M})} \tag{2.15}$$

2.2.8.4 Multi-dimensional problems

Consider the case where $\mathbf{X} \in \mathbb{R}^M$ is a M-dimensional vector. The multivariate CDF $P_{\mathbf{X}|\mathcal{M}}(\mathbf{x}|\mathcal{M})$ for $\mathbf{X}$ is then defined as[1]:

$$P_{\mathbf{X}|\mathcal{M}}(\mathbf{x}|\mathcal{M}) = \Pr(\mathbf{x} \leq \mathbf{X}|\mathcal{M}) \tag{2.16}$$

The multivariate PDF $p_{\mathbf{X}|\mathcal{M}}(\mathbf{x}|\mathcal{M})$ for $\mathbf{X}$ is defined through proposition $W : \mathbf{a} \leq \mathbf{X} < \mathbf{b}$, with $\mathbf{a}, \mathbf{b} \in \mathbb{R}^M$ and $\mathbf{a} \leq \mathbf{b}$:

$$\Pr(W|\mathcal{M}) = \int_{a_1}^{b_1} \dots \int_{a_M}^{b_M} p_{\mathbf{X}|\mathcal{M}}(\mathbf{x}|\mathcal{M}) \, \mathrm{d}\mathbf{x}_1 \dots \mathrm{d}\mathbf{x}_M = P_{\mathbf{X}|\mathcal{M}}(\mathbf{b}|\mathcal{M}) - P_{\mathbf{X}|\mathcal{M}}(\mathbf{a}|\mathcal{M}) \tag{2.17}$$

Eqs. (2.14) and (2.15) can be extended to multi-dimensional problems:

$$p_{\mathbf{H}|\mathcal{M}}(\mathbf{h}|\mathcal{M}) = \int_{\mathbb{R}^M} p_{\mathbf{H}|\mathbf{X},\mathcal{M}}(\mathbf{h}|\mathbf{x},\mathcal{M}) \cdot p_{\mathbf{X}|\mathcal{M}}(\mathbf{x}|\mathcal{M}) \, \mathrm{d}\mathbf{x} \tag{2.18}$$

$$p_{\mathbf{X}|\mathbf{H},\mathcal{M}}(\mathbf{x}|\mathbf{h},\mathcal{M}) = \frac{p_{\mathbf{H}|\mathbf{X},\mathcal{M}}(\mathbf{h}|\mathbf{x},\mathcal{M}) \cdot p_{\mathbf{X}|\mathcal{M}}(\mathbf{x}|\mathcal{M})}{p_{\mathbf{H}|\mathcal{M}}(\mathbf{h}|\mathcal{M})} \tag{2.19}$$

2.2.8.5 Paradoxes and infinite sets

The axioms of Probability Theory (Section 2.2.2) are stated in terms of finite sets. In this section (Section 2.2.8), Probability Theory is extended to continuous quantities $\mathbf{X}$ on $\mathbb{R}^M$; i.e., the support of $\mathbf{X}$ is a subset of $\mathbb{R}^M$ and, thus, an infinite set. For problems that typically arise in engineering, the extension of Probability Theory to infinite dimensional spaces is straight-forward (see Sections 2.2.8.1 – 2.2.8.4).

However, in general, care must be taken when working with conditional PDFs, as conditioning on a continuous quantity requires an infinite amount of information [Beck, 2014]. In principal, such a case should be approached as limit of a case with finite information [Beck, 2014]. Working directly with infinite sets in more general settings is the cause for many paradoxes [Jaynes, 2003, Chapter 4]. The problem is often a neglect in specifying how the limit $\mathrm{d}x \to 0$ is approached [Jaynes, 2003, Chapter 4].

[1] The vector inequality $\mathbf{x} \leq \mathbf{X}$ is defined as $x_i \leq X_i \; \forall \; i \in M$.

2.3 Modeling of uncertainty

2.3.1 Stochastic variables / random variables

2.3.1.1 Notation

The term *stochastic variable* was introduced by [Beck, 2010] to refer to a variable with uncertain value. The uncertainty associated with a stochastic variable is quantified by means of a probabilistic model. [Beck, 2010] uses the notion of a *stochastic variable* instead of the more commonly used notion of a *random variable* to highlight that the actual value of the underlying quantity is not necessarily random – only our knowledge about the value of the underling quantity is uncertain.

In this work, the term *stochastic variable* is used to explicitly denote a quantity whose value we are uncertain about, but that is fixed in the *real world*. Most of the uncertainties that arise in engineering can be represented by means of *stochastic variables*. The term *random variable* is used more general as a mathematical construct to denote both quantities that are unknown but fixed, and quantities that are fluctuating.

The focus in this thesis is primarily on real-valued uncertain quantities that are continuous (and not discrete).

2.3.1.2 Definition

A real-valued stochastic (or random) variable θ is pragmatically defined as $\theta \in \Gamma$, where $\Gamma \subseteq \mathbb{R}$ is referred to as the support of θ. The probabilistic model behind θ is contained in the stochastic model class $\mathcal{M}$, which will be formally introduced in Section 4.1. The entire analysis is always (either explicitly or implicitly) conditional on $\mathcal{M}$. Let $W = [a, b]$ with $a, b \in \mathbb{R}$ and $a \leq b$ represent an arbitrary subset of $\mathbb{R}$. The probabilistic model of θ contained in $\mathcal{M}$ assigns each W a probability that $\theta \in W$; i.e., $\Pr[\theta \in W|\mathcal{M}]$. For $W = \mathbb{R}$, by definition $\Pr[\theta \in \mathbb{R}|\mathcal{M}] = 1$. The support Γ is formally defined as

$$\Gamma = \left\{ a \in \mathbb{R} \middle| \lim_{\mathrm{d}a \to 0} \frac{\Pr[a \leq \theta < a + \mathrm{d}a|\mathcal{M}]}{\mathrm{d}a} > 0 \right\} \tag{2.20}$$

Note: The above definition is different from the classical definition of a random variable in *Kolmogorov's probability theory*. Classically, a real-valued random variable is defined as: Let (Ω, S, P) be a so-called probability space (see Section 2.2.4). A random variable X is defined as the function $X : \Omega \to \mathbb{R}$ that maps elements of the sample space Ω to a real-valued quantity.

Interpreting a stochastic variable θ as a real-valued quantity and not as a mapping is in

accordance with regarding the value of a stochastic variable as uncertain, but fixed in the *real world*.

2.3.1.3 Cumulative distribution function (CDF)

The *cumulative distribution function* (CDF) for random variable X is defined as (see also Section 2.2.8):

$$P_{X|\mathcal{M}}(x|\mathcal{M}) = \Pr(x \leq X|\mathcal{M}) \tag{2.21}$$

The CDF of X has the following properties:

- Any CDF is a *non-decreasing* and *right-continuous*[1] function.
- The CDF is *zero* at $\min_{x\in\Gamma}(x)$; i.e.,:

$$\lim_{x\to-\infty} P_{X|\mathcal{M}}(x|\mathcal{M}) = 0 \tag{2.22}$$

- The CDF is *one* at $\max_{x\in\Gamma}(x)$; i.e.,:

$$\lim_{x\to\infty} P_{X|\mathcal{M}}(x|\mathcal{M}) = 1 \tag{2.23}$$

2.3.1.4 Probability density function (PDF)

The *probability density function* (PDF) for random variable X is defined as:

$$P_{X|\mathcal{M}}(b|\mathcal{M}) - P_{X|\mathcal{M}}(a|\mathcal{M}) = \int_a^b p_{X|\mathcal{M}}(x|\mathcal{M})\,\mathrm{d}x \tag{2.24}$$

$p_{X|\mathcal{M}}(x|\mathcal{M})$ is equal to the derivative of $P_{X|\mathcal{M}}(x|\mathcal{M})$ *almost everywhere*.

2.3.1.5 Quantile function

The *quantile function* for X is the inverse function of the CDF for X.

$$x = P^{-1}_{X|\mathcal{M}}(p) \tag{2.25}$$

where $p \in [0, 1]$.

The x that is associated with a certain p is referred to as the p-quantile for X.

[1] *right-continuous*: Loosely speaking, in a right-continuous functions no jump occurs if the limit is approached from the right.

As the quantile function $P_X^{-1}(\cdot)$ is the inverse function of $P_X(\cdot)$, the derivative of the quantile function is:

$$\frac{\mathrm{d}P_X^{-1}(p)}{\mathrm{d}p} = \frac{1}{\frac{\mathrm{d}P_X\left(P_X^{-1}(p)\right)}{\mathrm{d}x}} = \frac{1}{p_X\left(P_X^{-1}(p)\right)} \tag{2.26}$$

To keep the notation simple, the conditional dependency on $\mathcal{M}$ is omitted in above equations.

2.3.1.6 Transformation of random variables

A transformation of random variable X to random variable Z is a function $Z = \mathrm{T}(X)$. The corresponding inverse transformation is usually denoted as: $X = \mathrm{T}^{-1}(Z)$. If the distributions of both X and Z are known, one can express the transformation $\mathrm{T} : X \to Z$ as:

$$\mathrm{T}(x) = P_Z^{-1}\left(P_X(x)\right) \tag{2.27}$$

which holds if the CDF of X and Z is strictly increasing. Equivalently, the inverse transformation $\mathrm{T}^{-1} : Z \to X$ can be written as:

$$\mathrm{T}^{-1}(z) = P_X^{-1}\left(P_Z(z)\right) \tag{2.28}$$

The derivative of the transformation is:

$$\frac{\mathrm{d}z}{\mathrm{d}x} = \frac{p_X(x)}{p_Z\left(\mathrm{T}(x)\right)} = \frac{p_X(x)}{p_Z\left(z\right)} \tag{2.29}$$

Proof 2.1. The proof is based on the *chain rule*:

$$\frac{\mathrm{d}z}{\mathrm{d}x} = \frac{\mathrm{d}\mathrm{T}(x)}{\mathrm{d}x} = \frac{\mathrm{d}P_Z^{-1}\left(P_X(x)\right)}{\mathrm{d}x} = \frac{1}{p_Z\left(P_Z^{-1}\left(P_X(x)\right)\right)} \cdot p_X(x) = \frac{p_X(x)}{p_Z\left(\mathrm{T}(x)\right)}$$

□

2.3.2 Numerical descriptors of random variables

Properties of random variables are typically quantified by means of numerical descriptors (e.g., expectation, variance). Some numerical descriptors for random variables are listed in Appendix A.

2.3.3 Probability distributions

Usually uncertainties are expressed in terms of probability distributions that are of a standardized type. Some common types of discrete and continuous probability distributions are listed in Appendix B.

2.3.4 Vectors of stochastic variables

2.3.4.1 Definition

The clustering of stochastic variables θ_i, $i = 1, \ldots, M$ in a vector is referred to as stochastic vector $\boldsymbol{\theta} \in \mathbb{R}^M$. The joint CDF of $\boldsymbol{\theta}$ is defined as:

$$P_{\boldsymbol{\theta}|\mathcal{M}}(\mathbf{t}|\mathcal{M}) = \Pr(\theta_1 \leq t_1 \wedge \ldots \wedge \theta_M \leq t_M|\mathcal{M}), \quad \mathbf{t} \in \mathbb{R}^M \tag{2.30}$$

The joint PDF is the derivative of the joint CDF:

$$p_{\boldsymbol{\theta}|\mathcal{M}}(\boldsymbol{\theta}|\mathcal{M}) = \frac{\partial^M P_{\boldsymbol{\theta}|\mathcal{M}}(\boldsymbol{\theta}|\mathcal{M})}{\partial\theta_1 \ldots \partial\theta_M} \tag{2.31}$$

The marginal density of component θ_i is:

$$p_{\theta_i|\mathcal{M}}(\theta_i|\mathcal{M}) = \int_{\mathbb{R}^{M-1}} p_{\boldsymbol{\theta}|\mathcal{M}}(\boldsymbol{\theta}|\mathcal{M})\, \mathrm{d}\boldsymbol{\theta}_{-i} \tag{2.32}$$

with $\mathrm{d}\boldsymbol{\theta}_{-i} = \mathrm{d}\theta_1 \ldots \mathrm{d}\theta_{i-1}\mathrm{d}\theta_{i+1} \ldots \mathrm{d}\theta_M$.

2.3.4.2 Vector of two correlated Normal random variables

Let $\boldsymbol{\theta} = [\theta_1, \theta_2]^{\mathrm{T}}$, where θ_1 and θ_2 are Normal random variables with mean μ_1 and μ_2, and standard deviation σ_1 and σ_2, respectively. The covariance between θ_1 and θ_2 is $\mathrm{Cov}[\theta_1, \theta_2] = \sigma_1\sigma_2\rho$, where $\rho \in [-1, 1]$ is the coefficient of linear correlation between θ_1 and θ_2. Furthermore, let $Z = \theta_1 + \theta_2$.

2.3.4.2.1 Joint PDF of $\boldsymbol{\theta}$

The joint PDF of $\boldsymbol{\theta}$ is:

$$\begin{aligned} p_{\theta_1,\theta_2}(\theta_1, \theta_2) =& \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \\ & \cdot \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(\theta_1-\mu_1)^2}{\sigma_1^2} + \frac{(\theta_2-\mu_2)^2}{\sigma_2^2} - \frac{2\rho(\theta_1-\mu_1)(\theta_2-\mu_2)}{\sigma_1\sigma_2}\right)\right] \end{aligned} \tag{2.33}$$

2.3.4.2.2 Distribution of $\theta_1|\theta_2$

The distribution of $\theta_1|\theta_2$ is Normal with mean $\mu_{1|2}$ and standard deviation $\sigma_{1|2}$.

$$\mu_{1|2} = \mu_1 + \frac{\sigma_1}{\sigma_2}\rho\,(\theta_2 - \mu_2) \tag{2.34}$$

$$\sigma_{1|2} = \sigma_1\sqrt{1-\rho^2} \tag{2.35}$$

2.3.4.2.3 Distribution of the sum Z

The sum of two correlated Normal random variables follows a Normal distribution with mean $\mu_Z = \mu_1 + \mu_2$ and standard deviation $\sigma_Z = \sqrt{\sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2}$. Thus, the distribution of Z can be stated as:

$$p_Z(Z) = \frac{1}{\sqrt{\sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2}\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(Z-\mu_1-\mu_2)^2}{\sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2}\right) \tag{2.36}$$

2.3.4.2.4 Distribution of $\theta_1|Z$

The distribution $p_{\theta_1|Z}(\theta_1|Z)$ is a Normal distribution with mean μ_* and standard deviation σ_*.

$$\mu_* = \frac{\mu_1\sigma_2^2 + (Z-\mu_2)\sigma_1^2 + \rho\sigma_1\sigma_2(Z-\mu_2+\mu_1)}{\sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2} \tag{2.37}$$

$$\sigma_* = \frac{\sigma_1\sigma_2\sqrt{1-\rho^2}}{\sqrt{\sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2}} \tag{2.38}$$

A proof for Eqs. (2.37) and (2.38) is given in Appendix E.2 (Proof E.10).

2.3.4.3 Vectors of independent standard Normal random variables

Let $\mathbf{u} = [u_1, \ldots, u_M]^{\mathrm{T}}$ be a M-dimensional vector, where each component u_i, $i \in \{1, \ldots, M\}$, of $\mathbf{u}$ is a standard Normal random variable that is independent of all other components of $\mathbf{u}$. Let R denote the length of $\mathbf{u}$, i.e., $R = \|\mathbf{u}\|$ is the Euclidean norm of $\mathbf{u}$. Thus, the squared length of $\mathbf{u}$ is R^2; i.e., $R^2 = \sum_{i=1}^{M} u_i^2$.

2.3.4.3.1 Joint PDF

The joint PDF of random vector $\mathbf{u}$ is

$$p(\mathbf{u}) = \prod_{i=1}^{M} \varphi(u_i) \tag{2.39}$$

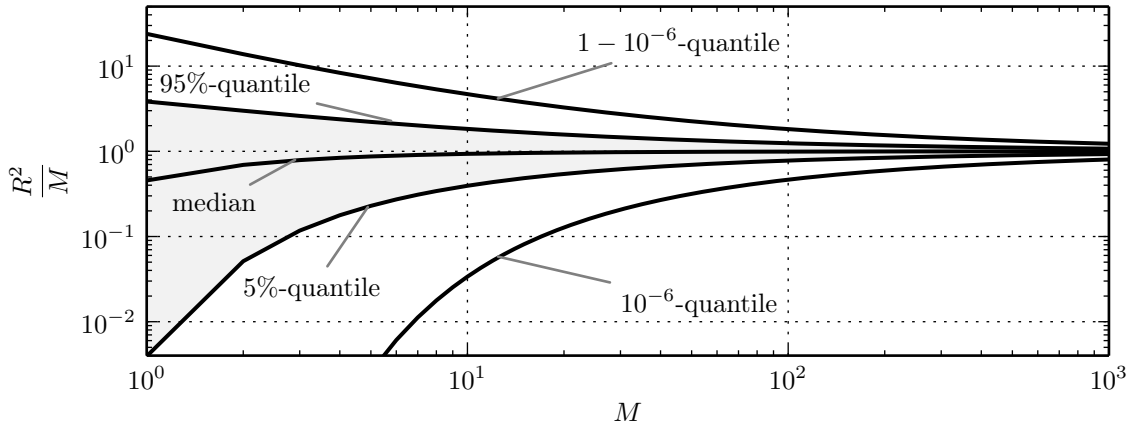


Figure 2.2: Selected quantiles of R^2/M for increasing values of M.

where $\varphi(\cdot)$ denotes the PDF of the standard Normal distribution. In terms of numerical performance, Eq. (2.39) can be expressed more efficiently as

$$p(\mathbf{u}) = \left(\frac{1}{\sqrt{2\pi}}\right)^N \exp\left(-\frac{1}{2}R^2\right) \tag{2.40}$$

The log-transform of the joint PDF of $\mathbf{u}$ is:

$$\ln\left[p(\mathbf{u})\right] = -\frac{1}{2}\left[N \cdot \ln\left(2\pi\right) + R^2\right] \tag{2.41}$$

2.3.4.3.2 Distribution of the squared length R^2

The squared length R^2 of a M-dimensional vector of independent standard Normal variables follows a chi-squared distribution with M degrees of freedom. Thus, the mean and standard deviation of R^2 is M and $\sqrt{2M}$, respectively. For increasing M, the coefficient of variation c_{v,R^2} of R^2 decreases:

$$c_{v,R^2} = \sqrt{\frac{2}{M}} \tag{2.42}$$

Thus, $c_{v,R^2} \to 0$ as $M \to \infty$. Consequently, for large M, the points $\mathbf{u}$ are located on the surface of a M-dimensional hypersphere that has radius $\sqrt{M}$. Selected quantiles of R^2/M for increasing values of M are depicted in Fig. 2.2.

2.3.4.3.3 Distribution of the angle between two independent vectors

Let $\mathbf{v}$ be an arbitrary M-dimensional vector. The angle between $\mathbf{u}$ and $\mathbf{v}$ is denoted as ω. We can, without loss of generality, assume that $\mathbf{v} = [1, 0, 0, \ldots]$. In this case, the cosine of ω can be expressed as:

$$\cos\omega = \frac{u_1}{\|\mathbf{u}\|} \tag{2.43}$$

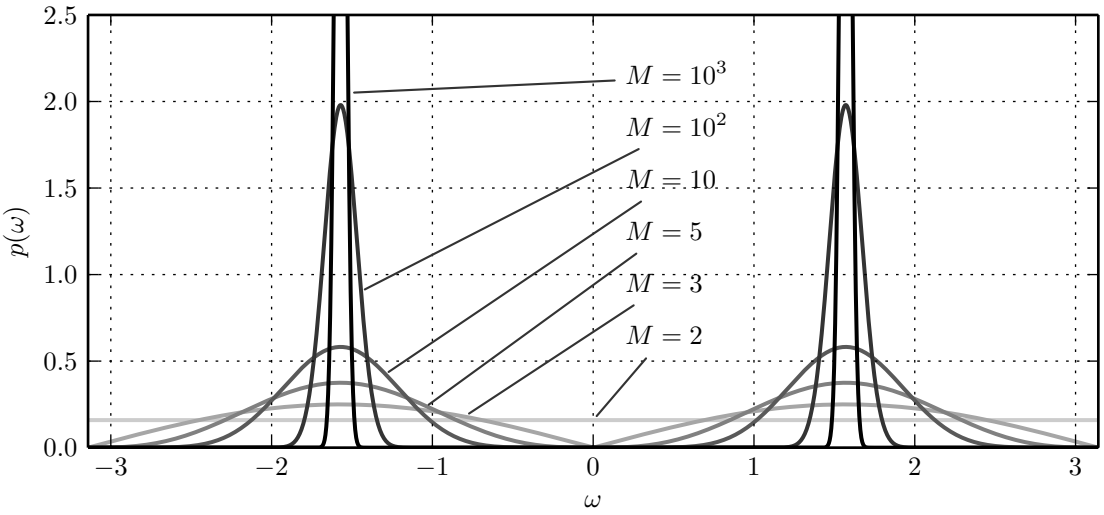


Figure 2.3: PDF of ω for different dimensions M.

Squaring both sides in Eq. (2.43), we get:

$$\cos^2 \omega = \frac{u_1^2}{R^2} = \frac{u_1^2}{u_1^2 + \sum_{i=2}^{M} u_i^2} \tag{2.44}$$

The quantity u_1^2 follows a chi-squared distribution with a single degree of freedom, and the quantity $\sum_{i=2}^{M} u_i^2$ follows a chi-squared distribution with $M-1$ degrees of freedom. Moreover, u_1^2 and $\sum_{i=2}^{M} u_i^2$ are independent. Thus, $\cos^2 \omega$ defined in Eq. (2.44) is beta-distributed with shape parameters $\alpha = 0.5$ and $\beta = (M-1)/2$.

Based on the beta-distributed $\cos^2 \omega$, the density of ω can be derived as

$$p(\omega) = \frac{|\sin \omega|^{M-2}}{2 \cdot B\left(\frac{1}{2}, \frac{M-1}{2}\right)} \tag{2.45}$$

where the support of ω is $[-\pi, \pi]$ and $B(\cdot, \cdot)$ denotes the beta function. The density $p(\omega)$ is depicted in Fig. 2.3 for different M. Selected quantiles of $|\omega|$ for increasing values of M are depicted in Fig. 2.4. Note that as $M \to \infty$, we have $|\omega| \to \frac{\pi}{2}$.

A proof for Eq. (2.45) is given in Appendix E.2 (Proof E.11).

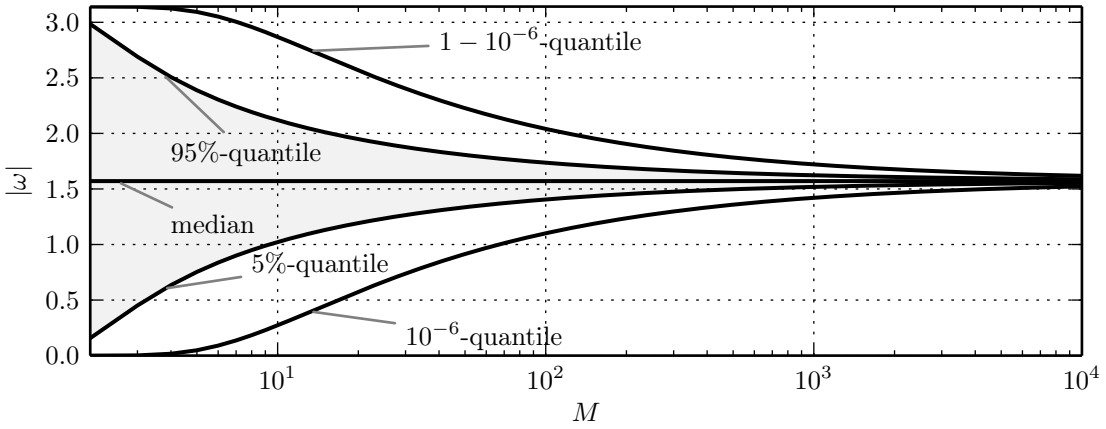


Figure 2.4: Selected quantiles of $|\omega|$ for increasing values of M.

2.3.5 Stochastic processes

2.3.5.1 Overview

Let $T \subseteq \mathbb{R}$ be a discrete or continuous set of real numbers[1]. A *stochastic process* X is a collection $\{X_t : t \in T\}$, where each X_t is a random variable. The quantity t is usually referred to as *time*. If T is discrete, the process is called *discrete time* stochastic process; if T is continuous, the process is referred to as *continuous time* stochastic process. The space of possible values that X_t, with $t \in T$, can take is called the *state space* of the process. It is typically distinguished between discrete and continuous state spaces. In the following, if not mentioned otherwise, the discrete cases are regarded as special cases of the continuous case.

Stationary stochastic processes A stochastic process is said to be *stationary* if its joint probability distribution is invariant to a shift in time: Let $p_X(X_{t_1}, \ldots, X_{t_k})$ be the PDF of the joint distribution of $\{X_t\}$ at times $t_1, \ldots, t_k$. The stochastic process is called stationary, if for all $t_1, \ldots, t_k$:

$$p_X(X_{t_1}, \ldots, X_{t_k}) = p_X(X_{t_1+\tau}, \ldots, X_{t_k+\tau}) \,, \tag{2.46}$$

independent of $\tau \in \mathbb{R}$.

Second-order stochastic process A *second-order stochastic process* is completely defined by its *mean function* $\mu_X(t) = \mathrm{E}[X_t]$ and its *auto-covariance function* $C_X(t_1, t_2) = \mathrm{Cov}[X_{t_1}, X_{t_2}]$, with $t, t_1, t_2 \in T$. The *auto-covariance* function can be expressed in terms of the *standard deviation function* $\sigma_X(t)$ and the *auto-correlation coefficient function*

[1] In general, the set T is not restricted to $\mathbb{R}$, but can be any totally ordered set. However, in the context of this work we limit ourselves to $T \subseteq \mathbb{R}$. Note that typically the elements of totally ordered sets can be mapped to numerical values such that the ordering is preserved.

$\rho_X(t_1, t_2)$:

$$C_X(t_1, t_2) = \sigma_X(t_1) \cdot \sigma_X(t_2) \cdot \rho_X(t_1, t_2) \tag{2.47}$$

Mean stationary stochastic process The stochastic process is said to be *mean stationary*, if its *mean function* does not depend on time: $\mu_X(t) = \mu_X$; i.e., is constant.

Weak-sense stationary stochastic process A stochastic process is called *weak-sense stationary* (or *second-order stationary*) if its *mean function* $\mu_X(t)$ and *auto-covariance function* $C_X(t_1, t_2)$ are invariant to a shift in time; i.e., $\mu_X(t) = \mu_X(t + \tau) = \mu_X$ and $C_X(t, t + \tau) = C_X(0, \tau) = C_X(\tau)$, independent of $t \in T$ and $\tau \in \mathbb{R}$. In this case, the *auto-correlation coefficient function* is also expressed in terms of τ:

$$\rho_X(\tau) = \frac{C_X(\tau)}{{\sigma_X}^2} \tag{2.48}$$

Markov process A *Markov process* satisfies the *Markov property*; i.e.,

$$p_X(X_{t_{i+1}}, \ldots, X_{t_k} | X_{t_1}, \ldots, X_{t_i}) = p_X(X_{t_{i+1}}, \ldots, X_{t_k} | X_{t_i}), \quad k > i \tag{2.49}$$

Power spectral density The *power spectral density* $S_X(\omega)$ of a *second-order stationary* stochastic process is linked to the *auto-covariance function* $C_X(\tau)$ through a Fourier transform as:

$$S_X(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} C_X(\tau) e^{-i\omega\tau} \, \mathrm{d}\tau \,, \tag{2.50}$$

with $i^2 = -1$. The power spectral density is a real-valued function that is symmetric $S_X(\omega) = S_X(-\omega)$ and non-negative $S_X(\omega) \geq 0$. For a given power spectral density, the variance of the underlying stochastic process can be evaluated as

$${\sigma_X}^2 = C_X(0) = \int_{-\infty}^{\infty} S_X(\omega) \, \mathrm{d}\omega \tag{2.51}$$

The *auto-covariance function* can be obtained from the power spectral density by an inverse Fourier transform:

$$C_X(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_X(\omega) e^{i\omega\tau} \, \mathrm{d}\omega \tag{2.52}$$

White noise process In a *white noise* stochastic process all random variables are independent of each other, have *zero* mean and finite variance; i.e., the power spectral density is constant. The joint PDF can be expressed as:

$$p_X(X_{t_1}, \ldots, X_{t_k}) = \prod_{i=1}^{k} p_{X_i}(X_i) \tag{2.53}$$

Gaussian stochastic process A stochastic process whose joint distribution of $\{X_t\}$ at times $t_1, \ldots, t_k$ is a multivariate Normal distribution for all $t_1, \ldots, t_k$, is called a *Gaussian stochastic process*. One property of *Gaussian stochastic processes* is: the marginal PDF of X_t at time t, denoted $p_{X_t}(X_t)$, is also Gaussian.

2.3.5.2 Autoregressive models

Autoregressive models are used to represent discrete time stochastic processes. An autoregressive model is typically denoted by AR(m), where $m \in \{0, 1, \ldots\}$ is referred to as the order of the process. The order m defines the number of past states that the process remembers. The current state X_t, with $t \in T \subseteq \mathbb{Z}$, of a discrete time stochastic process X is expressed based on the m previous states $\{X_{t-1}, \ldots, X_{t-m}\}$:

$$X_t = c + \sum_{i=1}^{m} a_i \cdot X_{t-i} + b_t \,, \tag{2.54}$$

where b_t denotes white noise that has a standard deviation of σ_b, and $a_1, \ldots, a_m, c, \sigma_b$ are the parameters of the model.

Special case: AR(1) **process**

The AR(1) process is a Markov process, because only the present state is needed to generate future states of the process:

$$X_t = c + a \cdot X_{t-1} + b_t \tag{2.55}$$

For $|a| < 1$ the AR(1) model describes a weak-sense stationary stochastic process. If $a = 1$, the process has infinite variance. In the following, we will restrict ourselves to the case $|a| < 1$.

The mean μ_X, standard deviation σ_X and auto-correlation coefficient function $\rho_X(\tau) = \rho(t_2 - t_1)$ of the stationary stochastic process are (see Proof E.12):

$$\mu_X = \frac{c}{1-a} \tag{2.56}$$

$$\sigma_X = \frac{\sigma_b}{\sqrt{1-a^2}} \tag{2.57}$$

$$\rho_X(\tau) = a^{|\tau|} \tag{2.58}$$

The auto-correlation coefficient function given in Eq. (2.58) can also be written as:

$$\rho_X(\tau) = \exp\left(-\frac{|\tau|}{l}\right) \,, \tag{2.59}$$

where $l = -\frac{1}{\ln(a)}$, for $a > 0$. Thus, the auto-correlation coefficient function of the AR(1) process has an exponential correlation structure if $a \in (0, 1)$. Reversely, a second-order

stationary stochastic process with exponential auto-correlation coefficient function can be represented as a AR(1) process. For a given mean μ_X, standard deviation σ_X and correlation length l, the parameters of the AR(1) process can be obtained as:

$$a = \exp\left(-\frac{1}{l}\right) \tag{2.60}$$

$$\sigma_b = \sigma_X \cdot \sqrt{1 - a^2} \tag{2.61}$$

$$c = \mu_X \cdot (1 - a) \tag{2.62}$$

A proof of Eq. (2.56), Eq. (2.57) and Eq. (2.58) is given in Appendix E.3 (Proof E.12).

2.3.6 Stochastic fields

An overview of stochastic fields and their numerical treatment can be found in Appendix D.

2.4 Information theory and entropy

2.4.1 Introduction

Information theory studies the ...

quantification: e.g., what is the information content actually contained in a message?

storage: e.g., how can information be stored such that it is optimally compressed?

communication: e.g., how can information be made redundant such that it can be transmitted safely across noisy channels?

...of information. A detailed introduction to *information theory* can be found in e.g., [MacKay, 2003] or [Cover and Thomas, 2006].

2.4.2 Self-information

Self-information $S(\mathcal{D}|\mathcal{M})$ measures how surprising it is to observe $\mathcal{D}$ conditional on the chosen probabilistic model $\mathcal{M}$:

$$S(\mathcal{D}|\mathcal{M}) = -\ln\left(\Pr(\mathcal{D}|\mathcal{M})\right) \tag{2.63}$$

where $S(\mathcal{D}|\mathcal{M})$ is measured in *nats*, as the natural logarithm is used. If the logarithm with base two is used instead in Eq. (2.63), the unit is *bits*.

Example 2.2. *Throwing a fair die and a fair coin* :
The probability of observing "three" when throwing a fair die is 1/6. Consequently, the *self-information* of this observation is 1.79.

The probability of observing "heads" when flipping a fair coin is 1/2. Consequently, the *self-information* of this observation is 0.69. Thus, we are less surprised to observed "heads" when flipping a coin than observing "three" when throwing a die.

If we flip a coin and throw a die and observe the joint outcome, the self-information of observing "three" and "heads" is $1.79 + 0.69 = 2.48$.

2.4.3 Entropy

Entropy is a key quantity in *information theory*. Entropy is a measure to quantify the amount of uncertainty about the state of a stochastic variable or stochastic process. The larger the uncertainty about the state of a system, the larger the associated entropy.

Let X be a discrete stochastic variable with N possible states x_i, and $i \in N$. Entropy $H[X|\mathcal{M}]$ is defined as the expected value of the self-information of the states of X:

$$\begin{aligned} H[X|\mathcal{M}] &= \mathrm{E}\left[S(X|\mathcal{M})\right] = \mathrm{E}\left[-\ln(\Pr(X|\mathcal{M}))\right] \\ &= \sum_{i=1}^{N} -\ln\left(\Pr(x_i|\mathcal{M})\right) \cdot \Pr(x_i|\mathcal{M}) \end{aligned} \tag{2.64}$$

which is measured in *nats*. If the logarithm with base two is used instead in Eq. (2.64), the entropy is referred to as *Shannon entropy* and measured in *bits*.

2.4.4 Differential entropy

Differential entropy is a generalization of the notion of *entropy* to continuous quantities. However, from a mathematical point of view, *differential entropy* cannot be associated with *entropy* (as defined in Section 2.4.3). *Differential entropy* for random variable X is defined as:

$$h\left[X|\mathcal{M}\right] = -\int_{\mathbb{R}} \ln\left(p_X(x|\mathcal{M})\right) \cdot p_X(x|\mathcal{M})\,\mathrm{d}x \tag{2.65}$$

Contrary to *entropy*, *differential entropy* can become *negative*.

2.4.5 Kullback–Leibler divergence

The *Kullback-Leibler divergence* $D_{\mathrm{KL}}(B\|A)$ is a measure for the information gain of going from the probabilistic model underlying A to the probabilistic model underlying B, where A

and B are stochastic variables.

If A and B are discrete and have N states, the *Kullback-Leibler divergence* is defined as:

$$D_{\mathrm{KL}}(B\|A) = \sum_{i=1}^{N} \ln\left(\frac{\Pr(b_i)}{\Pr(a_i)}\right) \cdot \Pr(b_i) \tag{2.66}$$

where a_i and b_i denotes the ith state of A and B, respectively. $D_{\mathrm{KL}}(B\|A)$ requires that $\Pr(a_i) = 0$ implies $\Pr(b_i) = 0$. Moreover, $D_{\mathrm{KL}}(B\|A)$ is defined to be *zero* if $\Pr(b_i) = 0$.

If A and B are continuous, $D_{\mathrm{KL}}(B\|A)$ is defined as:

$$D_{\mathrm{KL}}(B\|A) = \mathrm{E}_B\left[\ln\left(\frac{p_B(x)}{p_A(x)}\right)\right] = \int_{\mathbb{R}} \ln\left(\frac{p_B(x)}{p_A(x)}\right) \cdot p_B(x)\,\mathrm{d}x \tag{2.67}$$

where $D_{\mathrm{KL}}(B\|A)$ is also referred to as the *relative entropy*. The quantity $D_{\mathrm{KL}}(B\|A)$ is non-negative. Furthermore, $D_{\mathrm{KL}}(B\|A)$ is *zero* if and only if $p_A(\cdot)$ is equivalent to $p_B(\cdot)$ almost everywhere.

Chapter 3

Generating samples of a distribution

3.1 Overview

Let Θ be a M-dimensional vector of stochastic variables, and let $\boldsymbol{\theta}$ be a realization of Θ. If an explicit distinction between Θ and $\boldsymbol{\theta}$ is not necessary, the symbol $\boldsymbol{\theta}$ is used to denote both the realization and the stochastic variable. Commonly, we have to distinguish between two different situations:

We can directly generate independent samples $\boldsymbol{\theta}$ from the specified target distribution:
The joint probability density function $p_\Theta(\boldsymbol{\theta})$ for Θ is given. If the individual components θ_i, $i = 1, \ldots, M$ are independent, a realization $\boldsymbol{\theta}$ can be obtained by sampling the components θ_i separately from the one-dimensional marginal densities $p_{\Theta_i}(\theta_i)$. However, when implementing stochastic analysis techniques in numerical codes, the following strategy is often advantageous instead of directly generating samples $\boldsymbol{\theta}$:

1. Generate a realization $\mathbf{u}$ of a M-dimensional vector of independent standard Normal random variables.

2. If the values of $\boldsymbol{\theta}$ are explicitly required: Transform this underlying vector $\mathbf{u}$ of independent standard Normal random variables to vector $\boldsymbol{\theta}$ that is a realization of Θ. The transformation from $\mathbf{u}$ to $\boldsymbol{\theta}$ is denoted by $\mathrm{T}^{-1} : \mathbf{u} \rightarrow \boldsymbol{\theta}$.

This strategy is particularly helpful if stochastic variables in Θ are dependent. Moreover, transforming the problem to the independent standard Normal space normalizes the joint

PDF of the the stochastic variables of the problem. This allows us to set-up importance sampling densities or Markov chain proposal distributions that achieve an acceptable performance for a wide range of problems – independent of the variance of the stochastic variables in Θ.

Sampling methods based on this principle are referred to as *transformation methods* and are discussed in Section 3.2.

It is not feasible/desirable to directly generate samples θ from the specified target distribution:
This is, for example, the case if (i) the joint PDF for Θ is only known up to a normalizing constant, or (ii) the joint distribution of Θ is not known explicitly. In this case, realizations θ of Θ can be generated by means of *rejection sampling* (see Section 3.3) or *Markov chain Monte Carlo* simulation (see Section 3.4).

3.2 Transformation methods

Transformation methods generate samples θ of Θ by (i) first generating a sample $\mathbf{u}$ from the independent standard Normal distribution, and (ii) thereafter applying transformation $\mathrm{T}^{-1} : \mathbf{u} \rightarrow \theta$ to obtain the desired realization θ.

Widely used transformation methods presented in this section are the *Rosenblatt transformation* (Section 3.2.1) and the *Nataf transformation* (Section 3.2.2).

3.2.1 Rosenblatt transformation

The Rosenblatt transformation [Rosenblatt, 1952; Hohenbichler and Rackwitz, 1981] can be used for the mapping T^{-1} if the joint PDF $p_{\Theta}(\theta)$ is known and can be written as $p_{\Theta}(\theta) = p_{\Theta_1}(\theta_1) \cdot p_{\Theta_2|\Theta_1}(\theta_2|\theta_1) \cdot \ldots \cdot p_{\Theta_N|\Theta_1,\Theta_2,\ldots,\Theta_{M-1}}(\theta_M|\theta_1,\theta_2,\ldots,\theta_{M-1})$. In this case, Θ_i can be expressed as: $\Theta_i = P^{-1}_{\Theta_i|\Theta_1,\ldots,\Theta_{i-1}}(\Phi(U_i))$, where $P^{-1}_{\Theta_i|\theta_1,\ldots,\theta_{i-1}}(\cdot)$ is the inverse CDF of stochastic variable Θ_i given the states of $\Theta_1,\ldots,\Theta_{i-1}$, and $\Phi(\cdot)$ is the CDF of the standard Normal distribution. It is straight-forward to apply the Rosenblatt transformation in numerical codes to express the mapping T^{-1} if the inverse CDF $P^{-1}_{\Theta_i|\Theta_1,\ldots,\Theta_{i-1}}(\cdot)$ of all Θ_i in Θ are known.

Example 3.1. *Modeling two correlated Normal random variables conditional on their sum*:
Let $\theta = [\theta_1, \theta_2]$, where θ_1 and θ_2 are Normal random variables with mean μ and standard deviation σ. The theory behind problems of this type is given in Section 2.3.4.2. θ_1 and θ_2

are correlated with coefficient of linear correlation ρ. Furthermore, let $Z = \theta_1 + \theta_2$. Thus, Z follows a Normal distribution with mean $\mu_Z = 2\mu$ and standard deviation $\sigma_Z = \sigma\sqrt{2(1+\rho)}$. The objective in this example is to model θ_1 and θ_2 conditional on realizations of Z.

The Rosenblatt transformation can be set-up as follows:

1. Generate a realization of Z, from a Normal distribution with mean μ_Z and standard deviation σ_Z.
2. Generate a realization of θ_1, from a Normal distribution with mean μ_* and standard deviation σ_*, according to Eqs. (2.37) and (2.38).
3. θ_2 can be expressed conveniently as $\theta_2 = Z - \theta_1$. Note that $\theta_2|\theta_1, Z$ is not an uncertain quantity.

3.2.2 Nataf transformation

If components of $\boldsymbol{\theta}$ are modeled as dependent, the joint PDF $p_\Theta(\boldsymbol{\theta})$ can sometimes not be expressed explicitly, because of incomplete dependence information. Instead, the dependent variables in $\boldsymbol{\theta}$ are assumed to be given in terms of their marginal distributions and the dependency structure is represented in terms of correlation coefficients. In this case, the Nataf distribution [Der Kiureghian and Liu, 1986] can be used to model the joint density $p_\Theta(\boldsymbol{\theta})$.

3.2.2.1 The Nataf distribution

Let $\mathbf{R}_\Theta$ be the given linear correlation matrix of stochastic vector $\boldsymbol{\theta}$. The Nataf distribution expresses $p_\Theta(\boldsymbol{\theta})$ as [Ditlevsen and Madsen, 2007]:

$$p_\Theta(\boldsymbol{\theta}) = \frac{p_{\Theta_1}(\theta_1) \cdot \ldots \cdot p_{\Theta_M}(\theta_M)}{\varphi(y_1) \cdot \ldots \cdot \varphi(y_N)} \cdot \varphi_N\left(\mathbf{y}|\mathbf{R}_\mathbf{Y}\right) \tag{3.1}$$

where $\mathbf{y} \in \mathbb{R}^M$ is a vector of correlated standard Normal variables, $p_{\Theta_i}(\cdot)$ is the marginal PDF of the ith component of $\boldsymbol{\theta}$, and $\varphi_M\left(\mathbf{y}|\mathbf{R}_\mathbf{Y}\right)$ is the multivariate standard Normal PDF that has correlation matrix $\mathbf{R}_\mathbf{Y}$. The relation between $\mathbf{y}$ and $\boldsymbol{\theta}$ is: $\theta_i = P_i^{-1}\left(\Phi(y_i)\right)$, where $P_i^{-1}(\cdot)$ is the inverse of the CDF of the ith marginal distribution and $\Phi(\cdot)$ is the CDF of the standard Normal distribution.

Let the components of $\mathbf{R}_\mathbf{X}$ and $\mathbf{R}_\mathbf{Y}$ be denoted by $\rho_{n,m}$ and $\rho'_{n,m}$, respectively. The relation between $\rho_{n,m}$ and $\rho'_{n,m}$ is [Der Kiureghian and Liu, 1986]:

$$\rho_{n,m} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \theta'_n\theta'_m\, \varphi_2(y_n, y_m, \rho'_{n,m})\, \mathrm{d}y_n\, \mathrm{d}y_m \tag{3.2}$$

where $\theta'_n = (\theta_n - \mathrm{E}[\theta_n])/\sqrt{\mathrm{Var}[\theta_n]}$, and $\varphi_2(y_n, y_m, \rho'_{n,m})$ is the PDF of the bivariate Normal

distribution of y_n and y_m with correlation coefficient $\rho'_{n,m}$. The values of θ_n and θ_m can be obtained from y_n and y_m by the mapping T^{-1}.

3.2.2.2 The mapping $\mathbf{T}^{-1}$

The transformation from $\mathbf{u}$ to $\boldsymbol{\theta}$ is done as follows:

1. First, the vector of independent standard Normal variables $\mathbf{u}$ is transformed to the vector of correlated standard Normal variables $\mathbf{y}$ with correlation matrix $\mathbf{R_Y}$. Let $\mathbf{A}$ be the Cholesky decomposition of $\mathbf{R_Y}$; i.e., $\mathbf{R_Y} = \mathbf{A}\mathbf{A}^{\mathrm{T}}$. The transformation can then be performed by $\mathbf{y} = \mathbf{A}\mathbf{u}$.

2. Second, the Θ_i are computed from the y_i by means of $\Theta_i = \mathrm{T}^{-1}(y_i) = P_i^{-1}(\Phi(y_i))$, where $P_i^{-1}(\cdot)$ is the inverse of the CDF of the ith marginal distribution, and y_i is the ith coefficient of $\mathbf{y}$.

3.2.2.3 Nataf transformation and Gaussian copulas

The Nataf transformation corresponds essentially to choosing a Gaussian copula to represent the dependency structure of $\boldsymbol{\theta}$ [Lebrun and Dutfoy, 2009], where the copula is parametrized using the linear correlation matrix $\mathbf{R}_\Theta$. Consequently, in the Nataf transformation, the joint distribution of $\boldsymbol{\theta}$ is modeled using a Gaussian copula.

3.2.2.4 Pitfalls of the Nataf transformation

1. Finding a $\rho'_{n,m}$ that belongs to a specified $\rho_{n,m}$ is not a trivial task, due to the double integral in Eq. (3.2). If the marginal distributions are Normal or log-normal, then an analytical relation between $\rho'_{n,m}$ and $\rho_{n,m}$ can be established [Ditlevsen and Madsen, 2007, Appendix 2]. Approximate relations for some marginal distributions and correlation coefficients $\rho_{n,m}$ are given in [Der Kiureghian and Liu, 1986]. In general, the $\rho'_{n,m}$ that belongs to a given $\rho_{n,m}$ can be found by constructing the inverse function of Eq. (3.2) numerically.

2. Not for all $\rho_{n,m}$ a corresponding $\rho'_{n,m}$ does exist [Der Kiureghian and Liu, 1986] (this issue is illustrated in Section 3.2.2.5). This holds in particular if $\rho_{n,m}$ is close to 1 or -1 [Ditlevsen and Madsen, 2007, Chapter 7]. Based on the relations derived for the two examples in Section 3.2.2.5, it seems likely that such constraints are imposed by the physics of the problem. Thus, if for a specific $\rho_{n,m}$ a corresponding $\rho'_{n,m}$ cannot be found, the correlation $\rho_{n,m}$ cannot be achieved with the imposed marginal distributions.

3. Moreover, even if all components $\rho'_{n,m}$ of $\mathbf{R_Y}$ are found, there is not guarantee that $\mathbf{R_Y}$ is positive definite [Ditlevsen and Madsen, 2007; Lebrun and Dutfoy, 2009] (such that the Cholesky decomposition can be applied). This problem arises in particular if the dimension of $\boldsymbol{\theta}$ is large [Lebrun and Dutfoy, 2009].

4. It is difficult to assemble the linear correlation matrix $\mathbf{R}_\Theta$ based on expert judgment, as a consequence of *(2)* and *(3)*: The expert would have to take the marginal distributions of $\boldsymbol{\theta}$ into account, to guarantee that a suitable transformation can be found [Lebrun and Dutfoy, 2009].

5. Sensitivity studies employing different marginal distributions and threshold exceedance studies can run into problems due to *(2)* and *(3)* [Lebrun and Dutfoy, 2009].

6. With the Gaussian copula that underlies the Nataf transformation, it is not possible to represent a positive tail dependence [Lebrun and Dutfoy, 2009].

3.2.2.5 Examples

The examples investigated in this section are:

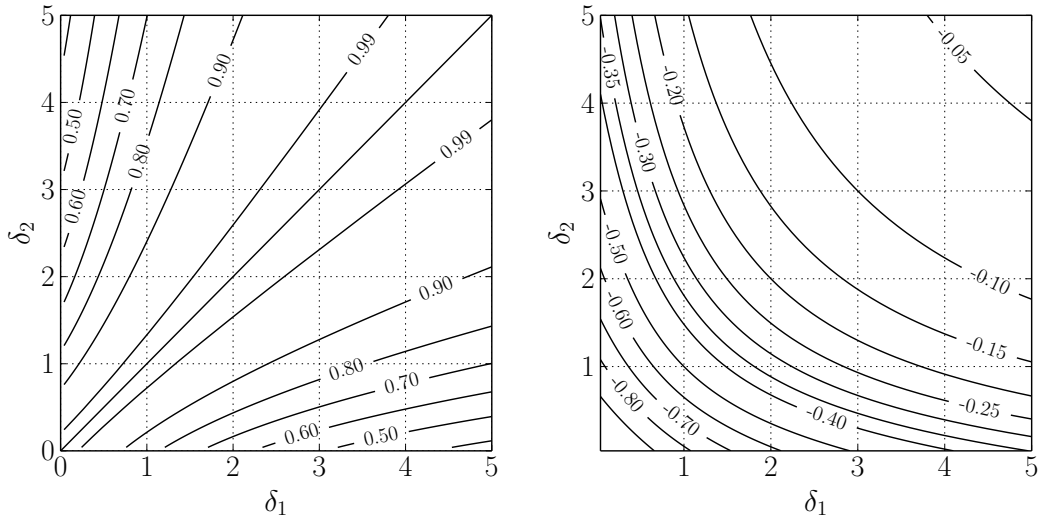
Example 3.2 Minimum and maximum correlation coefficients $\rho_{\min}$ and $\rho_{\max}$ are derived for two correlated random variables with log-normal marginal distributions.

Example 3.3 The minimum correlation coefficient $\rho_{\min}$ is derived for two correlated Bernoulli distributed random variables. The relation between ρ and ρ' is illustrated for different parameter combinations.

Example 3.2. *Modeling two correlated log-normal random variables with the Nataf distribution*: Let θ_1 and θ_2 be log-normal random variables that have coefficient of variation δ_1 and δ_2, respectively. The linear correlation coefficient between between θ_1 and θ_2 is denoted by ρ. For two correlated log-normal random variables, the correlation coefficient ρ' of the underlying standard Normal distributed random variables can be expressed explicitly as a function of ρ [Ditlevsen and Madsen, 2007, Appendix 2]:

$$\rho' = \frac{\ln\left(1+\rho\delta_1\delta_2\right)}{\sqrt{\ln\left(1+\delta_1^2\right)\left(1+\delta_2^2\right)}} \tag{3.3}$$

The maximum and minimum allowable ρ such that $-1 \leq \rho' \leq 1$ is maintained is denoted as $\rho_{\max}$ and $\rho_{\min}$, respectively. Based on Eq. (3.3), the equations for $\rho_{\max}$ and $\rho_{\min}$ can be derived



(a) $\rho_{\max}$ **(b)** $\rho_{\min}$

Figure 3.1: Two correlated log-normal random variables with correlation coefficient ρ. Upper bounds $\rho_{\max}$ and lower bounds $\rho_{\min}$ for ρ are depicted such that $-1 \leq \rho' \leq 1$ is maintained. The bounds are shown as contour lines for different coefficients of variation δ_1 and δ_2 of the log-normal random variables. (Example 3.2)

as:

$$\rho_{\max} = \frac{\exp\left(\sqrt{\ln\left(1+\delta_1^2\right)\ln\left(1+\delta_2^2\right)}\right) - 1}{\delta_1 \delta_2} \tag{3.4}$$

$$\rho_{\min} = \frac{\exp\left(-\sqrt{\ln\left(1+\delta_1^2\right)\ln\left(1+\delta_2^2\right)}\right) - 1}{\delta_1 \delta_2} \tag{3.5}$$

The two bounds $\rho_{\max}$ and $\rho_{\min}$ are shown in Fig. 3.1 for different coefficients of variation δ_1 and δ_2. For $\rho > 0$, a full correlation ($\rho = 1$) of θ_1 and θ_2 can only be achieved if the two random variables exhibit the same coefficient of variation. For $\rho < 0$, a $\rho = -1$ cannot be achieved with the Nataf model.

Note that the relation between ρ and ρ' stated in Eq. (3.3) is based on the theory of the problem; it can be derived from Eq. (B.29). Consequently, the two bounds $\rho_{\max}$ and $\rho_{\min}$ are imposed by theoretical constraints.

Example 3.3. *Modeling two correlated Bernoulli distributed random variables with the Nataf distribution*:
Let θ_1 and θ_2 be Bernoulli distributed random variables that can take states 1 and 0 and have common parameter p. In this case, the mapping T^{-1} is:

$$\theta_i = \mathrm{T}^{-1}(y_i) = \begin{cases} 1 & \text{if } \Phi(y_i) \leq p \\ 0 & \text{if } \Phi(y_i) > p \end{cases} \tag{3.6}$$

where $i \in \{1, 2\}$ and $\Phi(\cdot)$ is the CDF of the standard Normal distribution. Based on the discrete

nature of θ_1 and θ_2, Eq. (3.2) can be expressed as:

$$\begin{aligned}\rho = &\frac{1-p}{p}\int_{-\infty}^{\Phi^{-1}(p)}\int_{-\infty}^{\Phi^{-1}(p)}\varphi_2(y_1,y_2,\rho')\,\mathrm{d}y_1\,\mathrm{d}y_2\\ &+\frac{p}{1-p}\int_{\Phi^{-1}(p)}^{\infty}\int_{\Phi^{-1}(p)}^{\infty}\varphi_2(y_1,y_2,\rho')\,\mathrm{d}y_1\,\mathrm{d}y_2\\ &-2\int_{\Phi^{-1}(p)}^{\infty}\int_{-\infty}^{\Phi^{-1}(p)}\varphi_2(y_1,y_2,\rho')\,\mathrm{d}y_1\,\mathrm{d}y_2\end{aligned} \tag{3.7}$$

The double integral in the equation above can be reduced to a single integral:

$$\begin{aligned}\rho = &\frac{1-p}{p}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\Phi^{-1}(p)}\exp\left(-\frac{y^2}{2}\right)\cdot\Phi\left(\frac{\Phi^{-1}(p)-y\cdot\rho'}{\sqrt{1-(\rho')^2}}\right)\mathrm{d}y\\ &+\frac{p}{1-p}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\Phi^{-1}(p)}\exp\left(-\frac{y^2}{2}\right)\cdot\Phi\left(\frac{y\cdot\rho'-\Phi^{-1}(p)}{\sqrt{1-(\rho')^2}}\right)\mathrm{d}y\\ &-2\frac{1}{\sqrt{2\pi}}\int_{\Phi^{-1}(p)}^{\infty}\exp\left(-\frac{y^2}{2}\right)\cdot\Phi\left(\frac{\Phi^{-1}(p)-y\cdot\rho'}{\sqrt{1-(\rho')^2}}\right)\mathrm{d}y\end{aligned} \tag{3.8}$$

The relation between ρ and ρ' is illustrated in Fig. 3.2 for different values of p. Let $\rho'(\rho) : [\rho_{\min}, 1] \to [-1, 1]$, where $\rho_{\min}$ denotes the smallest value of ρ for which a corresponding ρ' can be found. Only for $p = 0.5$ we have $\rho_{\min} = -1$; otherwise, we have $\rho_{\min} > -1$. If the value of p approaches 0 or 1, $\rho_{\min}$ becomes 0.

Note that $\rho_{\min} > -1$ for p different from 0.5 is a limitation that comes from the physics of the problem – and is not a limitation due to the Nataf distribution: For the specified marginal distribution, the two random variables cannot exhibit a correlation coefficient smaller than $\rho_{\min}$. For example, let the mean of the Bernoulli distributed random variables be $p = 0.1$. A $\rho = -1$ can only be achieved if the mean of the second Bernoulli distributed random variable is 0.9 (and not 0.1).

The equation for $\rho_{\min}$ as a function of p can be derived explicitly based on the definition of the correlation coefficient. For the case of two Bernoulli distributed random variables with common parameter p, the correlation coefficient is defined as:

$$\rho = \frac{\mathrm{E}[\theta_1\theta_2] - p^2}{p(1-p)} \tag{3.9}$$

Note that for $p \leq 0.5$, the expectation $\mathrm{E}[\theta_1\theta_2]$ cannot be smaller than 0; and for $p \geq 0.5$, the expectation $\mathrm{E}[X_1X_2]$ cannot be smaller than $2p - 1$. The proof is omitted; it is based on assuming that with probability q we have $\theta_2 = 1 - \theta_1$, and with probability $1 - q$ the random variable θ_2 is a Bernoulli trial with rate r (while maintaining $E[\theta_2] = p$ independent of q, which gives $q = (p - r)/(1 - r - p)$). The smallest value that ρ can take for a given p is thus:

$$\rho_{\min} = \begin{cases} -\frac{p}{1-p} & \text{if } p \leq 0.5 \\ -\frac{1-p}{p} & \text{if } p > 0.5 \end{cases} \tag{3.10}$$

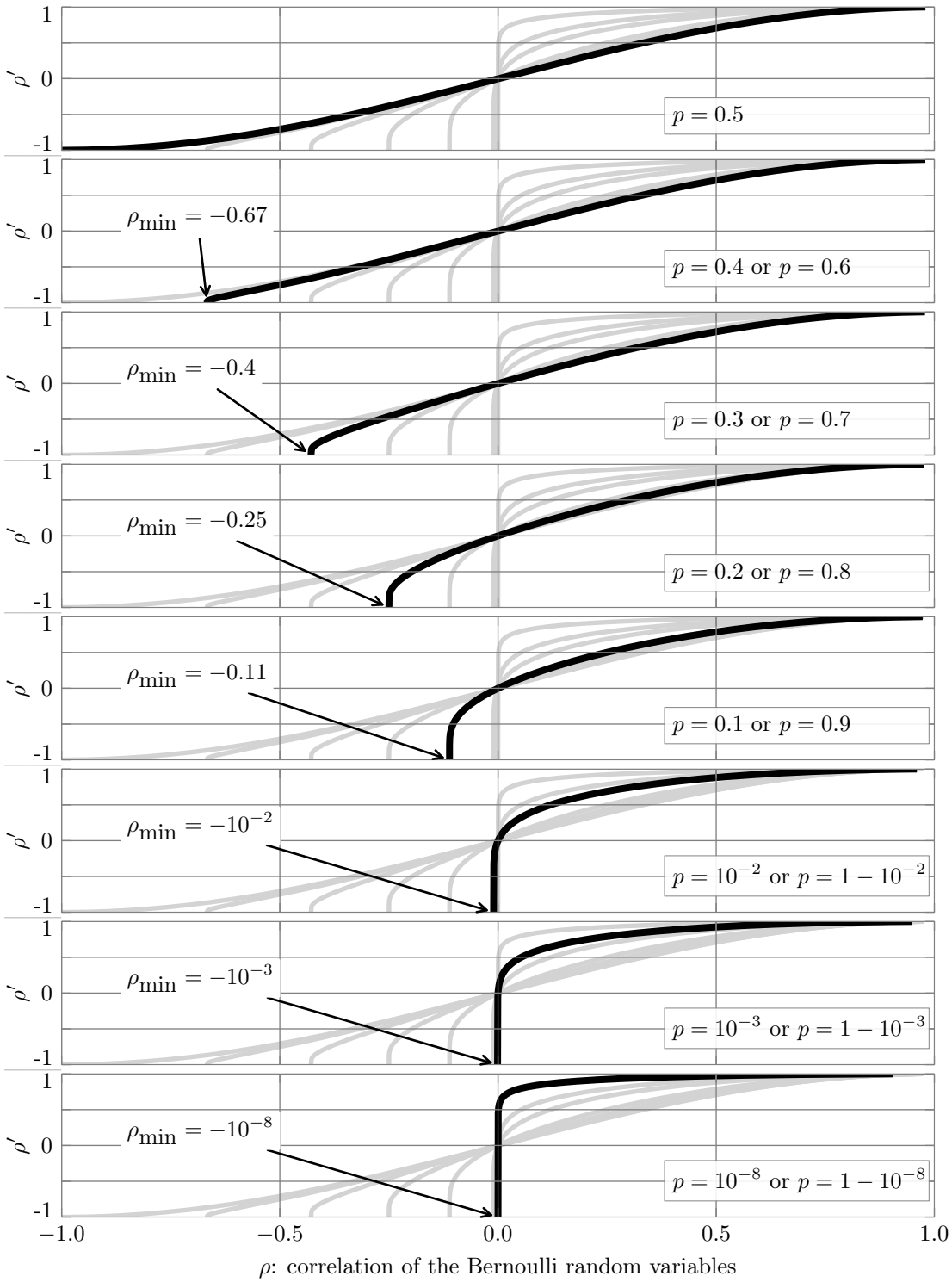


Figure 3.2: Relation between ρ and ρ' for different values of p, where p is the parameter of the Bernoulli distributed random variables (both random variables use the same p), ρ is the correlation between the Bernoulli distributed random variables, and ρ' is the correlation of the underlying standard Normal random variables. $\rho_{\min}$ denotes the smallest value of ρ for which a corresponding ρ' can be found; we have $\rho' = -1$ for $\rho_{\min}$. (Example 3.3)

3.3 Rejection sampling

Rejection sampling can generate independent samples of any distribution that is defined through its density, even if the density is not normalized or not known explicitly. Let the possibly unnormalized *target density* be $f(\boldsymbol{\theta})$; with $p_\Theta(\boldsymbol{\theta}) \propto f(\boldsymbol{\theta})$, where $p_\Theta(\boldsymbol{\theta})$ denotes the normalized target density, and $\boldsymbol{\theta} \in \mathbb{R}^M$. Furthermore, let $h(\boldsymbol{\theta})$ be a normalized density such that independent samples following this density can be generated directly; $h(\boldsymbol{\theta})$ is referred to as *proposal density*.

Rejection sampling postulates that we can pick a scaling factor $c \in \mathbb{R}_{>0}$ such that

$$c \cdot h(\boldsymbol{\theta}) \geq f(\boldsymbol{\theta}) \; \forall \; \boldsymbol{\theta} \in \mathbb{R}^M \tag{3.11}$$

The algorithm to generate an independent sample from target density $p_\Theta(\boldsymbol{\theta})$ is:

Algorithm 3.1. *Rejection sampling algorithm*:
This algorithm requires as input: (i) the possibly unnormalized target density $f(\boldsymbol{\theta})$, (ii) a proposal density $h(\boldsymbol{\theta})$, (iii) and a scaling factor c maintaining Eq. (3.11).

The algorithm returns an independent sample $\boldsymbol{\theta}$ from the target distribution.

1. Generate an independent sample $\boldsymbol{\xi}$ from density $h(\cdot)$.
2. Generate realization Γ from the uniform distribution with support $[0, 1]$.
3. Evaluate the ratio
$$r = \frac{f(\boldsymbol{\xi})}{c \cdot h(\boldsymbol{\xi})} \tag{3.12}$$
4. If $r \leq \Gamma$, set $\boldsymbol{\theta} = \boldsymbol{\xi}$; otherwise, go back to *(1)*.

The average computational costs to generate an independent sample from the target distribution are proportional to $1/\operatorname{E}[r]$, where r is the ratio defined in Eq. (3.12). Thus, for small $\operatorname{E}[r]$, the computational costs to generate independent samples from the target distribution can be large, rendering the *rejection sampling* algorithm inefficient.

3.4 Markov chain Monte Carlo

In the previous section, *rejection sampling* was introduced: It can generate independent samples from the target distribution at potentially large computational costs. Contrary to that, *Markov chain Monte Carlo* generates dependent samples from the target distribution at often considerably smaller computational costs.

Let $p_\Theta(\boldsymbol{\theta})$ be a distribution that we cannot sample from directly. *Markov chain Monte Carlo* (MCMC) is a technique that allows us to asymptotically generate dependent samples from $p_\Theta(\boldsymbol{\theta})$. The sampling process is called Markovian, as the next generated sample depends only on the current sample; all samples that were generated previously have no influence. Consequently, a Markov chain is memoryless. For a comprehensive introduction to MCMC, the reader is referred to [Gilks et al., 1996; Gelman et al., 2004a; Robert and Casella, 2004].

3.4.1 Formal introduction

Let $\mathbf{v}_0, \mathbf{v}_1, \ldots, \mathbf{v}_k$ be a sequence of samples, where $\mathbf{v}_k \in \mathbb{R}^M$ for all $k \in \{0, 1, \ldots\}$, and M is the dimension of the vectors $\mathbf{v}_k$. The sequence $\mathbf{v}_0, \mathbf{v}_1, \ldots, \mathbf{v}_k$ is called a *Markov chain* if the following rule holds for the joint PDF of any $\mathbf{v}_k$ with $k \geq 1$: $p(\mathbf{v}_k|\mathbf{v}_{k-1}, \ldots, \mathbf{v}_0) = p(\mathbf{v}_k|\mathbf{v}_{k-1})$. The conditional PDF $p(\mathbf{v}_k|\mathbf{v}_{k-1})$ is referred to as the *transition PDF* of the Markov chain. A Markov chain is defined through its transition PDF $p(\mathbf{v}_k|\mathbf{v}_{k-1})$ and the PDF of the initial state $p_0(\mathbf{v}_0)$.

We define the following properties that a Markov chain can have:

homogeneous A Markov chain is called *homogeneous* if the transition PDF is independent of the step k, i.e.: $p(\mathbf{v}_k|\mathbf{v}_{k-1}) = p(\mathbf{v}_{k+1}|\mathbf{v}_k)$ for all $k \geq 1$.

stationary distribution If $p_s(\mathbf{v}) = \int p(\mathbf{v}|\mathbf{w})p_s(\mathbf{w})\,\mathrm{d}\mathbf{w}$, then the density $p_s(\cdot)$ is termed a *stationary PDF* of a homogeneous Markov chain with transition PDF $p(\mathbf{v}|\mathbf{w})$. The stationary distribution is often also referred to as *invariant distribution*. Note that if we start a Markov chain with a sample from the stationary distribution, the chain will return a sample from the stationary distribution.

limiting distribution If the Markov chain converges to its stationary distribution independent of the starting point, the stationary distribution is called *limiting distribution*. If the chain has a limiting distribution, it has only one stationary distribution.

burn-in The number of steps until the Markov chain approximately reached its stationary distribution is called the *burn-in period*. Usually, it is virtually impossible to ensure that a chain has reached its stationary distribution [Tierney, 1994; Gilks et al., 1996, Section 1.4.6]. Different strategies to assess convergence are discussed in [Cowles and Carlin, 1996]. Another discussion can be found in [Robert and Casella, 2004, Section 12].

perfect sampling If the initial distribution is chosen as the stationary distribution: $p_0(\mathbf{v}_0) = p_s(\mathbf{v}_0)$, no burn-in is required. This case is referred to as *perfect sampling*. A discussion of the concept of perfect sampling can be found in, e.g., [Robert and Casella, 2004, Section 13].

detailed balance The following equation is called *detailed balance condition* or *reversibility condition*:

$$p(\mathbf{v}|\mathbf{w})p_s(\mathbf{w}) = p(\mathbf{w}|\mathbf{v})p_s(\mathbf{v}) \quad \forall\, \mathbf{w}, \mathbf{v} \in \mathbb{R}^M \tag{3.13}$$

If this condition hold, $p_s(\cdot)$ is a stationary distribution of the Markov chain. Moreover, if detailed balance hold, the chain is called *reversible*.

irreducible The chain is said to be *irreducible* if $\mathbf{v}_k$ can be in any subset of $\mathbb{R}^M$ that has non-zero probability in a finite number of steps, independent of the initial state $\mathbf{v}_0$ of the chain.

recurrent Recurrence is an extension of irreducibility: A chain is *recurrent* if any subset of $\mathbb{R}^M$ that has non-zero probability can be reached infinitely often for (almost) all starting points. The chain is said to be *positive recurrent* if it has a stationary distribution; and *null recurrent* otherwise.

aperiodic If there are portions of the state space that the chain can only visit in certain regularly spaced steps, the Markov chain is said to be periodic [Tierney, 1994]. If a Markov chain is not *periodic*, it is called *aperiodic*. A Markov chain is said to be *strongly aperiodic* if there is a non-zero probability that $\mathbf{v}_k = \mathbf{v}_{k+1}$. If a chain is strongly aperiodic, it implies that it is aperiodic.

If a Markov chain *has stationary distribution* $p_s(\cdot)$ *and is irreducible*, then: [Gilks et al., 1996, Section 4.3]

- $p_s(\cdot)$ is the unique invariant distribution of the chain.
- The chain is positive recurrent.
- One can prove asymptotical convergence in the average transition kernel and in sample path averages.

If, additional to that, the Markov chain is also *aperiodic*, then:

- One can prove asymptotical convergence of the transition kernel.

[Gilks et al., 1996, Section 4.3] point out that proving aperiodicity is usually of little importance, because we are typically interested only in convergence with respect to sample path averages.

3.4.2 Efficiency of MCMC sampling

Samples produced with MCMC are not independent, and thus the efficiency of MCMC sampling is reduced compared to independent samples generated directly from the target dis-

tribution. Loosely speaking, the efficiency of MCMC sampling decreases as the degree of dependency of the samples in the Markov chain increases. The degree of dependency of the samples in the Markov chain is determined by the choice of the transition PDF. All MCMC algorithms differ essentially in their choice of the transition PDF.

There is no unique measure to quantify the efficiency of MCMC sampling strategies (see [Gelman et al., 1996; Roberts et al., 1997, 2001]). For example, the efficiency measure defined in Eq. (F.15) can be employed. Similar to the measure defined in Eq. (F.15), the reciprocal of the asymptotic variance of the sample mean can be used to quantify the efficiency of MCMC sampling [Gelman et al., 1996]. However, both measures are ambiguous for problems with more than a single random variable. Another measure to quantify the MCMC efficiency is the *expected squared jumping distance* (ESJD). Maximizing the ESJD is equivalent to minimizing the first-order autocorrelation coefficient of the Markov chain [Thiéry, 2010]. The ESJD maximizes the MCMC efficiency if the higher order autocorrelation coefficients are monotonically increasing with respect to the first-order autocorrelation coefficient [Pasarica and Gelman, 2010]. The ESJD is monitored e.g. in [Beskos et al., 2009; Roberts and Rosenthal, 2009].

In general it is, however, difficult to directly optimize the efficiency of a MCMC sampling approach based on a moderate number of MCMC samples.

3.4.3 Metropolis-Hastings algorithm

A well-known MCMC algorithm is the *Metropolis-Hastings* algorithm [Metropolis et al., 1953; Hastings, 1970]. The algorithm can be used to sample from any specified target PDF $p_s(\cdot)$. In fact, the Metropolis-Hastings algorithm does not require the target distribution to be normalized. This is of particular relevance in Bayesian updating, because the posterior distribution is usually only known up to a scaling constant.

Algorithm 3.2. *Metropolis-Hastings algorithm*:
This algorithm requires as input: (i) the possibly unnormalized target density $p_s(\cdot)$, and (ii) a proposal density $q(\cdot|\cdot)$.

The algorithm generates state $\mathbf{v}_k$ from the previous state $\mathbf{v}_{k-1}$ of the Markov chain for any $k \geq 1$ as follows:

1. Draw a sample $\boldsymbol{\xi}$ from proposal distribution $q(\boldsymbol{\xi}|\mathbf{v}_{k-1})$.

2. Evaluate the ratio
$$r_{\mathbf{v}_{k-1}}(\boldsymbol{\xi}) = \frac{p_s(\boldsymbol{\xi})}{p_s(\mathbf{v}_{k-1})} \cdot \frac{q(\mathbf{v}_{k-1}|\boldsymbol{\xi})}{q(\boldsymbol{\xi}|\mathbf{v}_{k-1})} \tag{3.14}$$

3. Draw sample $\Gamma \in [0;1]$ from the uniform distribution.

4. Set

$$\mathbf{v}_k = \begin{cases} \boldsymbol{\xi} & \text{if } r_{\mathbf{v}_{k-1}}(\boldsymbol{\xi}) < \Gamma \\ \mathbf{v}_{k-1} & \text{otherwise} \end{cases} \tag{3.15}$$

Note that in the above algorithm, the density $p_s(\cdot)$ does not necessarily have to be normalized, because only its ratio appears in Eq. (3.14).

Acceptance rate Let quantity $a_{\mathbf{v}_{k-1}}(\boldsymbol{\xi})$ be defined as $a_{\mathbf{v}_{k-1}}(\boldsymbol{\xi}) = \min\left(r_{\mathbf{v}_{k-1}}(\boldsymbol{\xi}), 1\right)$, with $r_{\mathbf{v}_{k-1}}(\boldsymbol{\xi})$ according to Eq. (3.14). The acceptance rate p_{acr} is the long-run average of $a_{\mathbf{v}_{k-1}}(\boldsymbol{\xi})$. The acceptance rate is often closely coupled to the efficiency of the Metropolis-Hastings algorithm (see Section 3.5.8).

Transition PDF of going from state $\mathbf{w}$ to state $\mathbf{v}$ can be written as:

$$p(\mathbf{v}|\mathbf{w}) = q(\mathbf{v}|\mathbf{w}) \cdot r^*_{\mathbf{w}}(\mathbf{v}) + \delta_{\mathbf{w}}(\mathbf{v}) \cdot \int \left(1 - r^*_{\mathbf{w}}(\boldsymbol{\xi})\right) q(\boldsymbol{\xi}|\mathbf{w}) \, \mathrm{d}\boldsymbol{\xi} \tag{3.16}$$

where $\delta_{\mathbf{w}}(\mathbf{v})$ is the Dirac mass at $\mathbf{w}$, and $r^*_{\mathbf{w}}(\mathbf{v})$ is the acceptance ratio for going from $\mathbf{w}$ to $\mathbf{v}$ defined as

$$r^*_{\mathbf{w}}(\mathbf{v}) = \min\left(1, r_{\mathbf{w}}(\mathbf{v})\right) = \min\left(1, \frac{p_s(\mathbf{v})}{p_s(\mathbf{w})} \cdot \frac{q(\mathbf{w}|\mathbf{v})}{q(\mathbf{v}|\mathbf{w})}\right) \tag{3.17}$$

where $r_{\mathbf{w}}(\mathbf{v})$ is defined according to Eq. (3.14).

Proofs related to the transition PDF of the Metropolis-Hastings algorithm are given in Appendix E.4.

Metropolis algorithm: The Metropolis-Hastings algorithm is a generalization of the *Metropolis* algorithm [Metropolis et al., 1953]: In the Metropolis algorithm, the proposal distribution is assumed to be symmetric, and, thus, the acceptance ratio given in Eq. (3.14) reduces to:

$$r(\boldsymbol{\xi}|\mathbf{v}_{k-1}) = \frac{p_s(\boldsymbol{\xi})}{p_s(\mathbf{v}_{k-1})} \tag{3.18}$$

This "variant" of the Metropolis-Hastings algorithm is also referred to as symmetric *random walk Metropolis* (RWM) algorithm.

Choice of the proposal distribution: The efficiency of the Metropolis-Hastings algorithm is closely coupled to the choice of the proposal distribution. In general, there are two primary choices to be made when selecting the proposal distribution: (i) The *type* of the proposal distribution and (ii) the *spread* of the proposal distribution have to be chosen. Here spread is to be understood as a measure of dispersion; e.g., standard deviation or variance.

The influence of the proposal distribution on the efficiency of the MCMC sampling is discussed in Section 3.5.8.

3.5 MCMC sampling with focus on a special category of target distributions

In reliability analysis (see Section 4.4) and in combination with the BUS approach for Bayesian inference (see Chapter 6), special types of target distributions arise (see Section 3.5.1). This section introduces MCMC algorithms specifically designed to tackle target distributions in reliability analysis and in the BUS approach.

3.5.1 Type of target distributions focused on

3.5.1.1 Definition

The focus in this thesis is on target densities $p_g(\cdot)$ that are proportional to:

$$p_g(\boldsymbol{\theta}) \propto I_g(\boldsymbol{\theta}) \cdot p_\Theta(\boldsymbol{\theta}) \tag{3.19}$$

where $p_\Theta(\cdot)$ is a joint probability density function, and $I_g(\boldsymbol{\theta})$ is the indicator function defined as

$$I_g(\boldsymbol{\theta}) = \begin{cases} 1 & \text{if } g(\boldsymbol{\theta}) \leq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.20}$$

where $g(\cdot)$ is a function that defines the support of the target distribution; i.e., the support of the target distribution is the region in which $g(\cdot) \leq 0$. The sample $\boldsymbol{\theta}$ is expressed through M-dimensional vector $\mathbf{u}$ of independent standard Normal random variables and transformation $\mathrm{T}^{-1} : \mathbf{u} \to \boldsymbol{\theta}$; i.e., $\boldsymbol{\theta} = \mathrm{T}^{-1}(\mathbf{u})$ (see Section 3.2). Thus, Eq. (3.19) can be rewritten as:

$$p_{G(\mathbf{u})}(\mathbf{u}) \propto I_G(\mathbf{u}) \cdot \prod_{i=1}^{M} \varphi(u_i) \tag{3.21}$$

where $G(\mathbf{u}) = g(\boldsymbol{\theta}) = g\left(\mathrm{T}^{-1}(\mathbf{u})\right)$, $\varphi(\cdot)$ denotes the PDF of the standard Normal distribution, and $I_G(\mathbf{u})$ denotes the indicator function defined as

$$I_G(\mathbf{u}) = \begin{cases} 1 & \text{if } G(\mathbf{u}) \leq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.22}$$

3.5.1.2 Nomenclature

Following the nomenclature employed in *reliability analysis*[1], the following expressions are used:

$g(\cdot)$ is referred to as *limit-state function.*

$G(\cdot)$ is referred to as *limit-state function* in standard Normal space.

$P_f = \mathrm{E}_\Phi\left(I_G(\mathbf{u})\right)$ is the expectation of the indicator function where the problem is expressed in terms of the underlying independent standard Normal distribution. The quantity P_f is called the *probability of failure.*

$\{\mathbf{u}|G(\mathbf{u}) \leq 0\}$ is referred to as the *support domain* or *failure domain.*

$\mathbf{u}^*$ is the point in the support domain $\mathbf{U}_f = \{\mathbf{u}|G(\mathbf{u}) \leq 0\}$ that is closest to the origin; i.e., $\mathbf{u}^* = \arg\min_{\mathbf{u}\in\mathbf{U}_f}\left(\|\mathbf{u}\|\right)$. This point is referred to as *design point* in standard Normal space.

$\boldsymbol{\theta}^*$ is defined as $\mathrm{T}^{-1} : \mathbf{u}^* \rightarrow \boldsymbol{\theta}^*$, and called the *design point.*

3.5.2 Two-stage approach for MCMC

For the special type of target densities $p_g(\boldsymbol{\theta})$ introduced in Section 3.5.1, a sample $\boldsymbol{\theta}$ from $p_g(\boldsymbol{\theta})$ can be generated using the following two-stage approach:

Algorithm 3.3. *Two-stage approach for MCMC in reliability analysis*:
As input, the algorithm requires the current state of the Markov chain. The algorithm returns the next state of the Markov chain.

1. A (dependent[2]) sample $\mathbf{y}$ from the multivariate independent standard Normal distribution is generated using one of the investigated MCMC algorithms; e.g., by means of *Algorithm (3.5.6).* The generated sample is then transformed[3] as $\mathrm{T}^{-1} : \mathbf{y} \rightarrow \mathbf{v}$, where $\mathbf{v}$ is a sample from the joint density $p_\Theta\left(\mathbf{v}\right)$ used in Eq. (3.19).
2. The generated sample $\mathbf{v}$ is accepted if $g(\mathbf{v}) \leq 0$ and rejected otherwise, where $g(\mathbf{v})$ is the limit-state function that belongs to target density $p_g(\boldsymbol{\theta})$ (see Eq. (3.19)). If $\mathbf{v}$ is rejected, the sample from the previous MCMC step is re-used.

---

[1] The theory of *reliability analysis* is presented in Section 4.4.
[2] The generated sample $\mathbf{y}$ is conditional on the current state of the Markov chain.
[3] The transformation is performed according to Section 3.2.

The MCMC algorithms introduced in Section 3.5 are specialized MCMC variants that are particularly efficient in generating dependent samples $\mathbf{y}$ form the multivariate independent standard Normal distribution. The subsequent transformation $\mathrm{T}^{-1} : \mathbf{y} \rightarrow \mathbf{v}$ (see Section 3.2) and the acceptance/rejection step based on $g(\mathbf{v}) \leq 0$ are standard steps that are equivalent for all MCMC variants discussed in Section 3.5.

Typically, the computationally demanding step is to evaluate the limit-state function $g(\mathbf{v})$ for the proposed sample $\mathbf{v}$. Therefore, the objective is to find a MCMC strategy that – for a fixed number of MCMC steps – maximizes the number of effectively independent[1] samples that follow density $p_g(\boldsymbol{\theta})$.

Remark: If the target density $p_{\Theta}(\cdot)$ can be written as the product of one-dimensional densities $p_i(\cdot)$; i.e., as $p_{\Theta}(\mathbf{v}) = \prod_{i=1}^{M} p_i(v_i)$, then sample $\mathbf{v}$ in the first stage of the two-stage approach can, in principle, be generated directly – without first generating an underlying standard Normal sample. This can be achieved by the MCMC algorithms presented in Section 3.5.4 and Section 3.5.5. However, typically it is numerically more convenient to work in the underlying standard Normal space and to perform transformation T^{-1} subsequently (see also Section 3.1).

3.5.3 Exemplary target distributions studied

Based on the type of target distributions introduced in the Section 3.5.1, in this section exemplary target distributions are defined that are studied throughout this thesis. The target distributions presented in the following differ essentially with respect to their associated limit-state function (see Section 3.5.1).

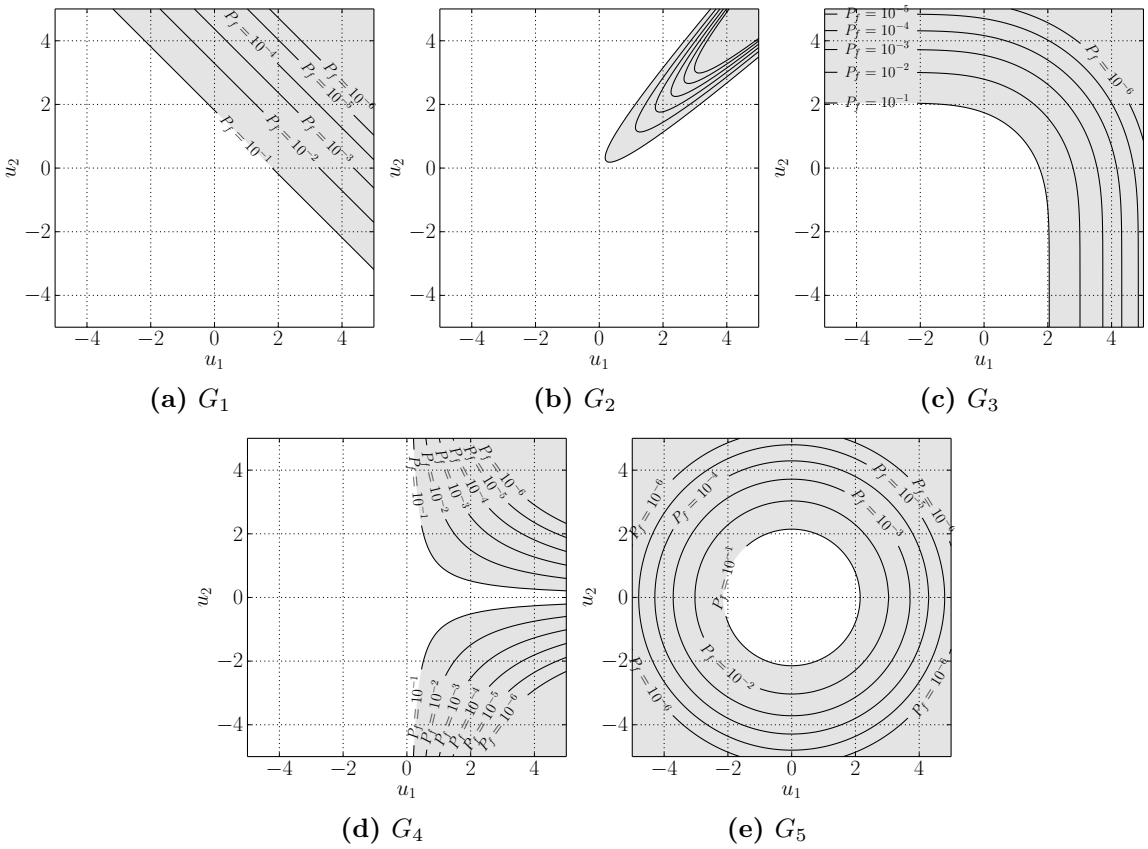
3.5.3.1 Linear Gaussian

The probability that the sum of standard Normal random variables (i.e., $\mathbf{u} = \boldsymbol{\theta}$) is larger than a specified threshold is assessed:

$$g_1(\boldsymbol{\theta}) = G_1(\mathbf{u}) = \beta_1 - \frac{1}{\sqrt{M}} \sum_{i=1}^{M} u_i \tag{3.23}$$

where the $\theta_i = u_i$ are standard Normal random variables, M is the dimension of the problem, and $\beta_1 = \Phi^{-1}(-P_{f,1})$ with $P_{f,1}$ as the target probability of failure. The shape of the failure domain is illustrated in Fig. 3.3a. The design point $\boldsymbol{\theta}^*$ of the problem is located at $\theta_1^* = \ldots = \theta_M^* = \frac{\beta_1}{\sqrt{M}}$.

[1] The *number of effectively independent samples* is an ambiguous measure. One measure based on the sample variance is introduced in Section F.1.2.4. The performance of different MCMC strategies is discussed in Section 3.5.8 for different efficiency measures.



(a) G_1 **(b)** G_2 **(c)** G_3

(d) G_4 **(e)** G_5

Figure 3.3: Shape of the support domains of target densities $p_{g_1}(\cdot)$, $p_{g_2}(\cdot)$, $p_{g_3}(\cdot)$, $p_{g_4}(\cdot)$ and $p_{g_5}(\cdot)$. The type of the associated target distributions is described in Section 3.5.1. The support domains are illustrated in standard Normal space (i.e., the functions $G_1(\cdot)$, $G_2(\cdot)$, $G_3(\cdot)$, $G_4(\cdot)$ and $G_5(\cdot)$ are illustrated) for two random variables (i.e., $M = 2$). For limit-state functions g_1, g_2, g_4 and g_5, the problem can directly be formulated in standard Normal space. In case of g_3, the limit-state function is linear in the original random variable space, but non-linear in standard Normal space. Limit-state functions g_2 and g_5 are shown for $\kappa = 10$ and $m = 0$, respectively. The depicted failure domains correspond to probabilities of failure 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5} and 10^{-6}. The failure domain corresponding to a target probability of failure $P_f = 10^{-1}$ is colored gray. (see Section 3.5.3)

The target density that employs limit-state function g_1 is denoted by $p_{g_1}(\boldsymbol{\theta})$.

3.5.3.2 Gaussian with quadratic term

Extending the previously introduced limit-state function, a quadratic term is added to the sum of standard Normal variables [Papaioannou et al., 2015]:

$$g_2(\boldsymbol{\theta}) = G_2(\mathbf{u}) = \alpha_2 - \frac{1}{\sqrt{M}} \sum_{i=1}^{M} u_i + \frac{\kappa}{4} (u_1 - u_2)^2 \tag{3.24}$$

where the $\theta_i = u_i$ are standard Normal stochastic variables, $M \geq 2$ is the dimension of the problem, κ is the principal curvature at the design point, and α_2 denotes the distance of the design point to the origin. The design point $\boldsymbol{\theta}^*$ of the problem is located at $\theta_1^* = \ldots = \theta_M^* = \frac{\alpha_2}{\sqrt{M}}$. The target probability of failure $P_{f,2}$ is linked to α_2 and κ as:

$$P_{f,2} = \int_{y=-\infty}^{-\alpha_2} \int_{z=-\infty}^{\infty} \varphi\left(-y + \frac{1}{2}\kappa z^2\right) \cdot \varphi(z) \, \mathrm{d}z \, \mathrm{d}y \tag{3.25}$$

The shape of the failure domain is illustrated in Fig. 3.3b for $\kappa = 10$. For $\kappa = 0$, the limit-state function g_2 is equivalent to g_1. For large κ, the problem reduces essentially to a two-dimensional problem of u_1 and u_2.

The target density that employs limit-state function g_2 is denoted by $p_{g_2}(\boldsymbol{\theta})$.

3.5.3.3 Sum of exponentials

Another limit-state function that has an analytical solution is the sum of stochastic variables with an exponential distribution:

$$g_3(\boldsymbol{\theta}) = \alpha_3 - \frac{1}{\sqrt{M}} \sum_{i=1}^{M} \theta_i \tag{3.26}$$

where θ_i are exponential stochastic variables with a mean of *one*, and target probability of failure $P_{f,3}$ is linked to α_3 through the CDF of the gamma distribution as:

$$P_{f,3} = \frac{\gamma(N, \alpha_3)}{\Gamma(N)} \tag{3.27}$$

where $\gamma(\cdot,\cdot)$ is the lower incomplete gamma function $\gamma(w,z) = \int_0^z \nu^{w-1} e^{-\nu} \, \mathrm{d}\nu$, and $\Gamma(\cdot)$ denotes the gamma function $\Gamma(w) = \gamma(w, \infty)$. The shape of the failure domain is illustrated in Fig. 3.3c.

The target density that employs limit-state function g_3 is denoted by $p_{g_3}(\boldsymbol{\theta})$.

3.5.3.4 Limit-state function with two design points

The following two-dimensional limit-state function has two design points:

$$g_4(\boldsymbol{\theta}) = G_4(\mathbf{u}) = \frac{\alpha_4}{|u_2|} - u_1 \tag{3.28}$$

where the $u_1 = \theta_1$ and $u_2 = \theta_2$ are standard Normal random variables, and α_4 is a positive scalar constant. The target probability of failure is:

$$P_{f,4} = 2 \cdot \int_0^\infty \Phi\left(-\frac{\alpha_4}{u}\right) \cdot \varphi(u)\, \mathrm{d}u \tag{3.29}$$

The two design points of limit-state function g_4 are located at $u_1^{*,1} = u_2^{*,1} = \sqrt{\alpha_4}$ and at $u_1^{*,2} = -u_2^{*,2} = \sqrt{\alpha_4}$. For example, with $\alpha_4 = 10$, we have $P_{f,4} = 5.42 \cdot 10^{-6}$ and $u_1^* = \pm u_2^* = \sqrt{10} \approx 3.16$. The shape of the failure domain is illustrated in Fig. 3.3d.

The target density that employs limit-state function g_4 is denoted by $p_{g_4}(\boldsymbol{\theta})$.

3.5.3.5 Points outside of hypersphere

Let $\boldsymbol{\theta} = \mathbf{u}$ be a M-dimensional vector of independent standard Normal random variables. The failure domain is defined as all points $\mathbf{u}$ located outside of a hypersphere that has radius $r \in \mathbb{R}_{>0}$ (see Fig. 3.3e). The following limit-state function is used to represent the failure domain:

$$g_5(\boldsymbol{\theta}) = G_5(\mathbf{u}) = 1 - \frac{\|\mathbf{u}\|^2}{r^2} - \frac{u_1}{r} \cdot \frac{1 - (\|\mathbf{u}\|/r)^m}{1 + (\|\mathbf{u}\|/r)^m} \tag{3.30}$$

where for $m \in [0, 4]$ the failure region[1] $G_5(\mathbf{u}) \leq 0$ is independent of m. The coefficient m modifies the gradient of the limit-state function in u_1-direction. The shape of G_5 is depicted in Fig. 3.4 for different values of m, with $M = 2$ and $P_{f,5} = 10^{-6}$. The parameter m will be modified to assess the performance of Subset Simulation[2]. In case $m = 0$, the limit-state function reduces to $g_5(\boldsymbol{\theta}) = G_5(\mathbf{u}) = 1 - \|\mathbf{u}\|^2/r^2$.

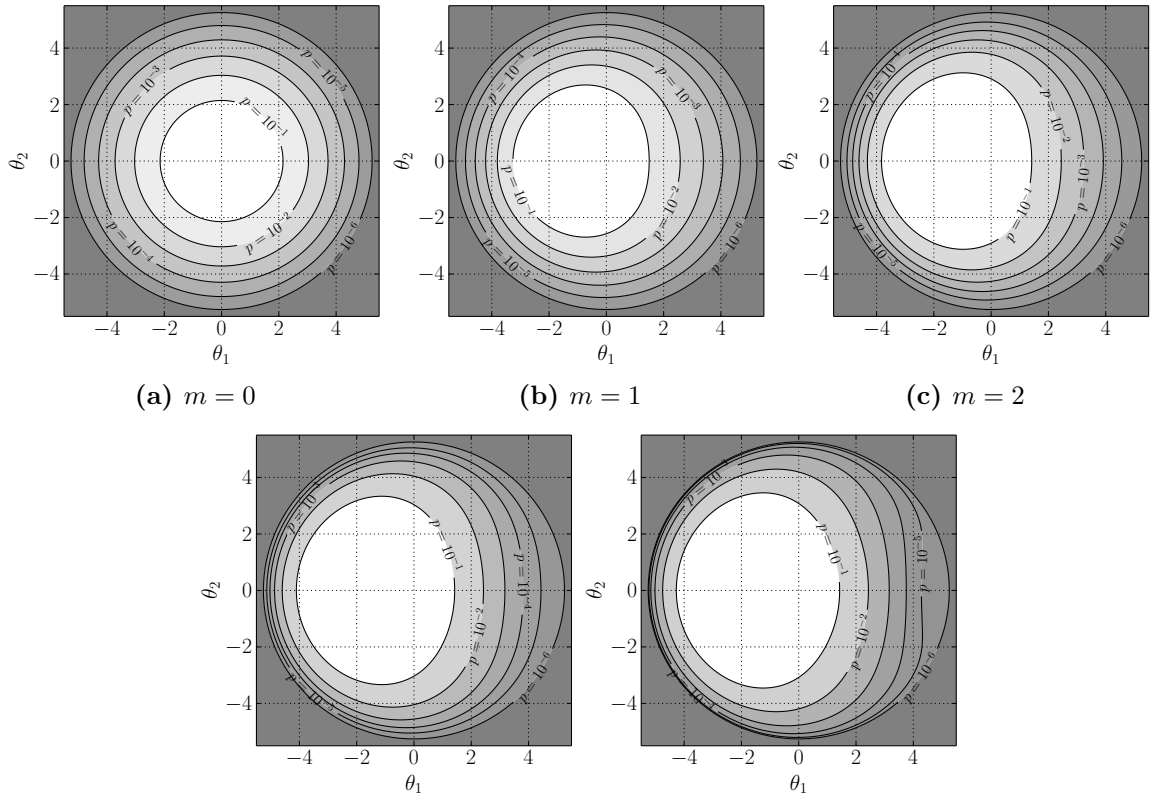
The sum of squared independent standard Normal random variables follows a chi-squared distribution. Therefore, the analytical solution of the probability of failure $P_{f,5}$ can be expressed by means of the CDF of a chi-squared distribution:

$$P_{f,5} = 1 - \frac{\gamma\left(M/2, r^2/2\right)}{\Gamma\left(M/2\right)} = \frac{\Gamma\left(M/2, r^2/2\right)}{\Gamma\left(M/2\right)} \tag{3.31}$$

where $\gamma(\cdot,\cdot)$ is the lower incomplete gamma function $\gamma(w, z) = \int_0^z \nu^{w-1} e^{-\nu}\, \mathrm{d}\nu$, $\Gamma(\cdot,\cdot)$ is the

[1] $m = 4$ is the largest integer number for which the shape of the failure region does not depend on m.

[2] *Subset Simulation* is a method to approximately estimate the probability of failure in a reliability problem. Subset Simulation is properly introduced in Section 5.3.



(a) $m = 0$ **(b)** $m = 1$ **(c)** $m = 2$

(d) $m = 3$ **(e)** $m = 4$

Figure 3.4: The shape of limit-state function g_5 is depicted for a two-dimensional problem and different values of m. The target probability of failure $P_{f,5}$ is set to 10^{-6}. The parameter r in g_5 can be evaluated based on $P_{f,5}$ as the inverse of the upper regularized incomplete gamma function; i.e., $r(P_{f,5} = 10^{-6}) = 5.26$. Let $\Theta_t = \{\boldsymbol{\theta} \in \Theta | g_5(\boldsymbol{\theta}) \leq t\}$, where t is chosen based on a value p such that $\Pr(g_5 \leq t) = p$. Additional to the final failure domain, the contour lines associated with domains Θ_t are shown for $p = \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$.

upper incomplete gamma function $\Gamma(w, z) = \int_z^\infty \nu^{w-1} e^{-\nu} \, \mathrm{d}\nu$, and $\Gamma(\cdot)$ denotes the gamma function $\Gamma(w) = \gamma(w, \infty)$.

The target density that employs limit-state function g_5 is denoted by $p_{g_5}(\boldsymbol{\theta})$.

3.5.4 Component-wise Metropolis-Hastings

Au and Beck [Au and Beck, 2001] observed and proved that the standard Metropolis-Hastings algorithm rejects almost all samples if the dimension M of $\mathbf{u}$ is large; i.e., the acceptance rate decreases with increasing M, for a fixed spread s_q of the proposal. This renders the algorithm inapplicable for moderate to large M [Au and Beck, 2001; Au and Wang, 2014]. For target distributions where the PDF $p_\Theta(\boldsymbol{\theta})$ can be expressed as $p_\Theta(\boldsymbol{\theta}) \propto \prod_{i=1}^{M} p_{\Theta_i}(\theta_i)$, [Au and Beck, 2001] propose to apply the Metropolis-Hastings algorithm separately to each

component[1]. This algorithm is referred to as *component-wise Metropolis-Hastings* (cwMH).

In the cwMH algorithm, the Markov chain moves on if a proposed sample in at least one component is accepted. Let a_1 be the average probability that a proposed sample in a component is accepted. As long a $a_1 > 0$, the probability $a = 1 - (1 - a_1)^M$ that the chain moves on converges asymptotically to *one* as $M \to \infty$ (even if a_1 is small). Consequently, the performance of the component-wise Metropolis-Hastings algorithm is (nearly) independent of the dimension M of the problem. Therefore, this family of MCMC algorithms is efficient for application in Subset Simulation [Au and Beck, 2001] (see Section 5.3).

Algorithm 3.4. *Component-wise Metropolis-Hastings (cwMH) algorithm*:
This algorithm requires as input: (i) the current state $\mathbf{w}$ of the Markov chain, and (ii) the one-dimensional proposal density[2] $q(\cdot|\cdot)$, and (iii) the possibly unnormalized one-dimensional target densities $p_{\Theta_i}(\cdot)$.

The algorithm returns $\mathbf{v}$, the next state of the Markov chain.

1. For each i in $\{1, \ldots, M\}$ do:
   (a) Propose sample ξ from proposal distribution $q(\xi|w_i)$.
   (b) Calculate
   $$r = \frac{p_{\Theta_i}(\xi)}{p_{\Theta_i}(w_i)} \cdot \frac{q(w_i|\xi)}{q(\xi|w_i)} \tag{3.32}$$
   (c) Draw sample $\Gamma \in [0; 1]$ from the uniform distribution.
   (d) Set $v_i = \xi$ if $\Gamma \leq r$; otherwise set $v_i = w_i$.

Usually, the one-dimensional proposal distribution $q(\xi|w_i)$ in the cwMH algorithm is chosen to be symmetric, i.e.: $q(\xi|w_i) = q(w_i|\xi)$. Au and Beck [Au and Beck, 2001] noted that the type of the symmetric proposal distribution does not have a large influence on the efficiency of the algorithm. However, the spread (e.g., the standard deviation) of the proposal distribution does. One could for example choose the Normal distribution as proposal distribution:

$$q(\xi|w_i) = \varphi\left(\frac{\xi - w_i}{s_q}\right) \tag{3.33}$$

where $s_q > 0$ denotes the spread.

[1] In this contribution the algorithm is specifically applied to generate dependent samples from the multivariate independent standard Normal distribution; i.e., for target distributions that can be written as $\prod_{i=1}^{M} \varphi(\theta_i)$, where $\varphi(\cdot)$ denotes the PDF of the standard Normal distribution.

[2] In general, for each of the M components of $\mathbf{w}$, denoted w_i, a separate proposal distribution can be specified. To keep the problem simple, the same proposal distribution is used in *Algorithm (3.4)* for each component of $\mathbf{w}$. However, in general, it is straight-forward to extent *Algorithm (3.4)* to proposal distributions that depend on the index $i = 1, \ldots, M$ of the respective component w_i of vector $\mathbf{w}$.

3.5.5 Conditional Metropolis-Hastings

If the joint target distribution of $\boldsymbol{\theta}$, denoted p_s, is known explicitly and has the form $p_s(\boldsymbol{\theta}) = p_\Theta(\boldsymbol{\theta}) \propto \prod_{i=1}^{M} p_{\Theta_i}(\theta_i)$ (i.e., i.e., the components of $\boldsymbol{\theta}$ are required to be independent), we can construct a Metropolis-Hastings proposal distribution that leads to an acceptance rate of *one*. This particular variant of Metropolis-Hastings is referred to as CMH in the following (for *conditional Metropolis-Hastings*). The CMH algorithm is a generalized variant of the CS algorithm (see Section 3.5.6).

Let $\mathbf{w}$ and $\mathbf{v}$ be M-dimensional random vectors that follow distribution p_s. Note that the components w_i, $i = 1, \ldots, M$ of $\mathbf{w}$ are independent – as are the components v_i of $\mathbf{v}$. Furthermore, let $\mathbf{z}$ and $\mathbf{y}$ be the standard Normal transformations of $\mathbf{w}$ and $\mathbf{v}$, respectively. Consequently, the components z_i of $\mathbf{z}$ can be expressed as $z_i = \Phi^{-1}(P_s(w_i))$, where $P_s(\cdot)$ is the CDF of the target distribution, and $\Phi^{-1}(\cdot)$ is the inverse CDF of the standard Normal distribution. The components v_i of $\mathbf{v}$ can be expressed as $v_i = P_s^{-1}(\Phi(y_i))$, where $P_s^{-1}(\cdot)$ is the inverse CDF of the target distribution. The random vectors $\mathbf{w}$ and $\mathbf{v}$ are dependent. The dependency is specified using the Nataf model: The dependency between $\mathbf{w}$ and $\mathbf{v}$ is expressed in terms of an equivalent correlation between $\mathbf{z}$ and $\mathbf{y}$.

In a Markov chain that moves from state $\mathbf{w}$ to the next state $\mathbf{v}$, the state $\mathbf{w}$ is conditionally fixed. Describing the correlation between the corresponding vectors $\mathbf{z}$ and $\mathbf{y}$ in terms of a single correlation coefficient ρ, the state $\mathbf{y}$ can be sampled from a conditional Normal distribution that has mean $\rho \cdot \mathbf{z}$ and isotropic standard deviation $\sqrt{1-\rho^2}$.

Algorithm 3.5. *Conditional Metropolis-Hastings (CMH) algorithm*:
This algorithm requires as input: (i) the current state $\mathbf{w}$ of the Markov chain, and (ii) the correlation coefficient $\rho \in [0; 1)$ that specifies the correlation between the transformed standard Normal states of the Markov chain, and (ii) the CDF $P_s(\cdot)$ of the target distribution and its inverse $P_s^{-1}(\cdot)$.

The algorithm returns $\mathbf{v}$, the next state of the Markov chain.

1. Transform $\mathbf{w}$ to standard Normal space:

$$z_i = \Phi^{-1}(P_s(w_i)), \qquad \text{for each } i \in \{1, \ldots, M\} \tag{3.34}$$

2. Obtain a realization of $\mathbf{y}$ conditionally on the state of $\mathbf{z}$:

$$\mathbf{y} = \rho \cdot \mathbf{z} + \sqrt{1-\rho^2} \cdot \mathbf{u} \tag{3.35}$$

where $\mathbf{u}$ is a realization of a M-dimensional vector of independent standard Normal random variables.

3. Transform $\mathbf{y}$ to $\mathbf{v}$:

$$v_i = P_s^{-1}\left(\Phi(\, y_i\,)\right), \qquad \text{for each } i \in \{1, \ldots, M\} \tag{3.36}$$

where $\mathbf{v}$ is the next state of the Markov chain.

Note that in the algorithm above, $\Pr(\mathbf{v} = \mathbf{w}) = 0$. Thus, in each MCMC step, a state different from the current state is generated – which means that the CMH algorithm does not reject any sample and has an acceptance rate of *one*. However, this holds only for the first stage of the two-stage approach explained in Section 3.5.2; the overall acceptance rate is smaller than *one* due to the acceptance/rejection step based on $g(\mathbf{v}) \leq 0$ in the second stage.

Proof 3.1. Algorithm (3.5) can be interpreted as a Metropolis-Hastings algorithm with an acceptance rate of *one*. As the components of both $\mathbf{w}$ and $\mathbf{v}$ are assumed to be independent, the joint PDF $p(\cdot)$ of both $\mathbf{w}$ and $\mathbf{v}$ can be expressed as the product of the probability densities of the individual components. Therefore, it is sufficient to conduct the proof for the special case of $M = 1$.
The PDF of y is:

$$p_{Y|Z}(y|z) = \varphi\left(\frac{y - \rho \cdot z}{\sqrt{1-\rho^2}}\right)$$

where $\varphi(\cdot)$ denotes the PDF of the standard Normal distribution. The conditional PDF of the proposal distribution $p_{V|W}(v|w)$ can be derived based on $p_{Y|Z}(y|z)$ and the relation between y and v:

$$\begin{aligned}
p_{V|w}(v|w) &= \varphi\left(\frac{y - \rho \cdot z}{\sqrt{1-\rho^2}}\right) \cdot \left|\frac{\mathrm{d}\Phi^{-1}\left(P_s(v)\right)}{\mathrm{d}v}\right| \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y - \rho \cdot z}{\sqrt{1-\rho^2}}\right)^2\right) \cdot \left|\frac{\mathrm{d}\Phi^{-1}\left(P_s(v)\right)}{\mathrm{d}P_s(v)} \cdot \frac{\mathrm{d}P_s(v)}{\mathrm{d}v}\right| \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y - \rho \cdot z)^2}{2 \cdot (1-\rho^2)}\right) \cdot \left|\frac{\mathrm{d}y}{\mathrm{d}\Phi(y)} \cdot p_s(v)\right| \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2 - 2\rho yz + \rho^2 z^2}{2 \cdot (1-\rho^2)}\right) \cdot \frac{p_s(v)}{\varphi(y)} \\
&= \exp\left(-\frac{y^2 - 2\rho yz + \rho^2 z^2}{2 \cdot (1-\rho^2)} + \frac{y^2}{2}\right) \cdot p_s(v) \\
&= \exp\left(-\frac{\rho^2 y^2 - 2\rho yz + \rho^2 z^2}{2 \cdot (1-\rho^2)}\right) \cdot p_s(v)
\end{aligned}$$

The Metropolis-Hastings acceptance ratio can be expressed as:

$$\begin{aligned}
r\,(v|w) &= \frac{p_s(v) \cdot p_{W|v}(w|v)}{p_s(w) \cdot p_{V|w}(v|w)} \\
&= \exp\left(-\frac{\rho^2 y^2 - 2\rho yz + \rho^2 z^2}{2 \cdot (1-\rho^2)} + \frac{\rho^2 y^2 - 2\rho yz + \rho^2 z^2}{2 \cdot (1-\rho^2)}\right) = \exp(0)
\end{aligned}$$

$$= 1$$

Consequently, the acceptance ratio r is *one*, independent of ρ. □

3.5.6 Conditional sampling in standard Normal space

3.5.6.1 Algorithm

The *conditional sampling in standard Normal space* (CS) algorithm was proposed in [Papaioannou et al., 2015; Au and Patelli, 2016; Au, 2016] to generate MCMC samples from an independent standard Normal target distribution. The CS algorithm is an efficient MCMC algorithm for application in Subset Simulation (see Section 5.3).

Let the target distribution take the form $p_s(\boldsymbol{\theta}) = p_\Theta(\boldsymbol{\theta}) \propto \prod_{i=1}^{M} \varphi(\theta_i)$, where $\varphi(\cdot)$ denotes the PDF of the standard Normal distribution. In this special case, the CMH algorithm (*Algorithm (3.5)*) reduces to the CS algorithm:

Algorithm 3.6. *Conditional sampling in standard Normal space (CS) algorithm*:
This algorithm requires as input: (i) the current state $\mathbf{z}$ of the Markov chain, and (ii) the correlation coefficient $\rho \in [0; 1)$ that specifies the correlation between the proposed sample and the seed. The algorithm returns $\mathbf{y}$, the next state of the Markov chain.

1. Sample random vector $\mathbf{u}$ as a realization of a M-dimensional vector of independent standard Normal random variables.

2. Obtain a realization of $\mathbf{y}$ conditionally on the state of $\mathbf{z}$:

$$\mathbf{y} = \rho \cdot \mathbf{z} + \sqrt{1 - \rho^2} \cdot \mathbf{u} \tag{3.37}$$

The parameter ρ in *Algorithm (3.6)* is linked to the spread s_q of the proposal distribution as $s_q = \sqrt{1 - \rho^2}$. From a numerical point of view, *Algorithm (3.6)* should ideally directly be implemented using parameter s_q to control the spread of the proposal, instead of employing ρ; i.e., Eq. (3.37) is expressed as

$$\mathbf{y} = \sqrt{1 - s_q^2} \cdot \mathbf{z} + s_q \cdot \mathbf{u} \tag{3.38}$$

Eq. (3.38) is less prone to floating-point errors than Eq. (3.37) if the value of s_q is small.

3.5.6.2 Efficiency of CS in high-dimensional problems

Note: The relations derived in the following are only valid as $M \to \infty$.

For large M, the samples $\mathbf{y}$ and $\mathbf{z}$ are essentially located on the surface of a hypersphere with radius $\sqrt{M}$ (see Section 2.3.4.3.2). Moreover, based on Section 2.3.4.3.3, $\mathbf{u} \cdot \mathbf{z} \to 0$ as $M \to \infty$, i.e., $\mathbf{u}$ is orthogonal to $\mathbf{z}$ for large M. Thus, based on Eq. (3.37), we can deduce: (i) The *scalar projection*[1] of $\mathbf{y}$ on $\mathbf{z}$ is $\rho\sqrt{M}$. (ii) The *scalar rejection*[2] of $\mathbf{y}$ on $\mathbf{z}$ is $\sqrt{(1-\rho^2)M}$. (iii) The angle ω between $\mathbf{y}$ and $\mathbf{z}$ is $\omega = \cos^{-1}(\rho)$. (iv) The distance between $\mathbf{y}$ and $\mathbf{z}$ is $\sqrt{2M(1-\rho)}$. Consequently, for $\rho < 1$, the samples $\mathbf{y}$ are located on the "ring" that is obtained by intersecting the hypersphere that has radius $\sqrt{(1-\rho^2)M}$ and is centered around $\rho\mathbf{y}$ with the hypersphere that has radius $\sqrt{M}$ and is centered around the origin.

Moreover, for very large M, the parameter ρ does effectively represent the cosine of the angle between $\mathbf{y}$ and $\mathbf{z}$. This behavior is particularly interesting, as in high dimensions the prior samples in independent standard Normal space are asymptotically located on the hypersphere with radius $\sqrt{M}$ centered around the origin.

3.5.6.3 Generalized CS variant

In *Algorithm (3.6)*, the same ρ is used for each of the M components; i.e., the covariance matrix of the proposal distribution is a diagonal matrix with diagonal entries $1-\rho^2$. A generalized variant of the CS algorithm in which general covariance matrices can be specified, is proposed in [Papaioannou et al., 2015]. Especially for low-dimensional problems (i.e., small M), the generalized CS algorithm can outperform the standard isotropic CS algorithm. The generalized CS variant is difficult to apply in the following situations:

- Typically, the covariance matrix of the proposal distribution is estimated from samples of the target distribution. However, for problems with many random variables (i.e., if M is large), it is difficult to estimate the covariance matrix based on a reasonable number of samples. If the number of samples is not chosen large enough, the estimated covariance matrix might not be positive definite.

- If the shape of the target distribution on support domain $\{\mathbf{u}|G(\mathbf{u}) \leq 0\}$ cannot be described well by means of a linear dependency structure.

[1] The *scalar projection* is defined as $\mathbf{yz}/\|\mathbf{z}\|$.

[2] The *scalar rejection* is defined as $\sqrt{\|\mathbf{y}\|^2 - (\mathbf{yz}/\|\mathbf{z}\|)^2}$.

3.5.7 Directional conditional sampling

The Metropolis-Hastings strategy proposed in the following requires the target distribution to be proportional to the multivariate independent standard Normal distribution.

3.5.7.1 Background

Let $\mathbf{u}$ be a M-dimensional vector of independent standard Normal random variables. The standard way to generate a realization of $\mathbf{u}$ is to independently generate a standard Normal sample for each component of $\mathbf{u}$. Alternatively, a realization of $\mathbf{u}$ can also be obtained using the following strategy (see e.g. [Katafygiotis and Zuev, 2008]):

Algorithm 3.7. *One viable strategy to generate a M-dimensional vector of independent standard Normal random variables.*:

1. Assign to l^2 the realization of a chi-squared distribution with M degrees of freedom.
2. Generate realizations of two M-dimensional vectors $\mathbf{d}_1$ and $\mathbf{d}_2$ of independent standard Normal random variables – by independently generating a standard Normal sample for each component of $\mathbf{d}_1$ and $\mathbf{d}_2$.[1]
3. Scale $\mathbf{d}_1$ and $\mathbf{d}_2$ such that they have a length of *one*.
4. Ensure that $|\mathbf{d}_1 \cdot \mathbf{d}_2| < 1$; otherwise[2] go back to *step (2)*.
5. In the hyperplane spanned by $\mathbf{d}_1$ and $\mathbf{d}_2$, select vector $\mathbf{d}_3$ such that $\mathbf{d}_1 \cdot \mathbf{d}_3 = 0$; i.e.:

$$\mathbf{d}_3 = \mathbf{d}_2 - (\mathbf{d}_1 \cdot \mathbf{d}_2) \cdot \mathbf{d}_1$$

6. Scale the length of $\mathbf{d}_3$ to *one*.
7. Assign to ϕ the realization of a uniform distribution with support $[0, 2\pi]$.
8. Set $\mathbf{u} = \sqrt{l^2} \cdot (\mathbf{d}_1 \cos\phi + \mathbf{d}_3 \sin\phi)$

The above algorithm can be split in three principal steps:

1. Sample the length l of vector $\mathbf{u}$.

[1] This step might seem tautologous: Two M-dimensional vectors of independent standard Normal random variables are required in order to generate a single M-dimensional vector of independent standard Normal random variables. However, the appeal of this approach will become apparent in the remainder of this section.

[2] i.e., if $\mathbf{d}_1$ and $\mathbf{d}_1$ are parallel

2. Sample the orientation of a hyperplane that contains the origin.

3. Generate a sample uniformly distributed on the circle obtained by intersection of the obtained hyperplane with the surface of a hypersphere around the origin that has radius l.

If the goal is to obtain a sample $\mathbf{u}$ that depends on the state of $\mathbf{w}$ (where both $\mathbf{u}$ and $\mathbf{w}$ follow a multivariate independent standard Normal distribution), the dependency between $\mathbf{u}$ and $\mathbf{w}$ can be modeled separately in each of the three principal steps. This allows a finer control over the dependency structure compared to using a single parameter ρ in the CS algorithm (Section 3.5.5).

3.5.7.2 Algorithm

Algorithm 3.8. *Directional conditional sampling (DCS).*:
This algorithm requires as input (i) the current state $\mathbf{w}$ of the Markov chain, (ii) correlation coefficient ρ_{r} that controls the dependency between the length of the proposed sample vector and the length of the seed sample vector, and (iii) correlation coefficient ρ_{ω} that controls the angle between the seed and the proposed sample vector.

The algorithm generates the next state $\mathbf{v}$ of the Markov chain.

1. Let l_{s}^2 denote the squared length of $\mathbf{w}$; i.e., $l_{\mathrm{s}}^2 = \|\mathbf{w}\|^2$. Note that l_{s}^2 follows a chi-squared distribution with M degrees of freedom. Apply the CMH algorithm (*Algorithm (3.5)*) to generate sample l^2, where l_{s}^2 is the current state of the chain, ρ_{r} is the required correlation coefficient, and $P_s(\cdot)$ in *Algorithm (3.5)* is the CDF of the chi-squared distribution with M degrees of freedom.

2. Generate random vector $\mathbf{z}$, where each component of $\mathbf{z}$ is a standard Normal random variable.

3. Ensure that $\mathbf{wz} < \|\mathbf{w}\| \cdot \|\mathbf{z}\|$ and $\|\mathbf{z}\| > 0$; otherwise, go back to the previous step.

4. Modify vector $\mathbf{z}$ as follows:
$$\mathbf{z} = \mathbf{z} - \frac{\mathbf{z} \cdot \mathbf{w}}{l_{\mathrm{s}}^2} \cdot \mathbf{w}$$
Note that $\mathbf{z}$ is now perpendicular to $\mathbf{w}$.

5. Normalize vector $\mathbf{z}$:
$$\mathbf{z} = \frac{1}{\|z\|} \cdot \mathbf{z}$$

6. Propose angle ω using the CMH algorithm (*Algorithm (3.5)*):

(a) Generate u as a sample from the standard Normal distribution.

(b) Transform u to ω:

$$\omega = \text{erf}\left(\frac{u \cdot \sqrt{1-\rho_\omega^2}}{\sqrt{2}}\right) \cdot \pi$$

7. Generate the next state $\mathbf{v}$ of the Markov chain as:

$$\mathbf{v} = \sin(\omega) \cdot \mathbf{z} + \frac{\cos(\omega)}{l_\text{s}^2} \cdot \mathbf{w}$$

3.5.7.3 Practical applicability

The cwMH, CMH and the CS algorithm require only a single parameter to control the spread of the proposal distribution. In contrast, two parameters (ρ_r and ρ_ω) are used to control the spread in the DCS algorithm (*Algorithm (3.8)*). On the one hand, two parameter allow finer control over the spread of the proposal distribution. On the other hand, finding appropriate values for two parameters is more challenging than tuning the spread with respect to only a single parameter.

For some low-dimensional problems, application of the DCS algorithm can be appealing. For example, for target distribution defined in terms of of limit-state function $g_5(\cdot)$ (see Section 3.5.3.5) with $m = 0$, the optimal spread is controlled through parameter ρ_r, and the parameter ρ_ω is ideally set to 0. In this example, the DCS algorithm reduces the problem to a one-dimensional problem and performs rather efficiently.

For general high dimensional problems (i.e., for large M), the CS algorithm (Section 3.5.6) is more appealing than the DCS algorithm: In high-dimensions the samples are efficiently located on the surface of a hypersphere (see Section 2.3.4.3.2). Thus, the influence of the parameter ρ_r diminishes as M increases. As a consequence, the DCS is similar to the CS algorithm for large M.

3.5.8 Numerical performance investigations

3.5.8.1 Overview

In the Metropolis-Hastings algorithm, the efficiency of MCMC sampling is often closely coupled to the acceptance rate [Roberts et al., 2001]. For example, for a one-dimensional Normal target distribution and a symmetric proposal distribution, the optimal spread is 0.44 [Gelman et al., 1996]. For problems with many random variables, a often near-optimal acceptance rate is 1/4 under quite general conditions [Roberts et al., 1997].

Instead of optimizing a measure for the efficiency of the MCMC sampling, it is often simpler to tune the acceptance rate to a specified target acceptance rate (see e.g., [Roberts and Rosenthal, 2009; Papaioannou et al., 2015]).

3.5.8.2 Numerical investigations

In this section, the efficiency of the CS and cwMH Metropolis-Hastings algorithms is investigated for different proposal spreads and different target distributions by means of numerical examples. The type of the investigated target density is selected according to Eq. (3.19) in Section 3.5.1.

The following efficiency measures are used to assess the MCMC performance:

$$\text{eff}_{\text{MH},1} = \frac{\text{ESJD}}{\text{ESJD}_{\text{opt}}} \tag{3.39}$$

$$\text{eff}_{\text{MH},2} = \text{eff}_{\gamma, I_{G_1}(\boldsymbol{\theta})} \tag{3.40}$$

$$\text{eff}_{\text{MH},3} = \text{eff}_{\gamma, I_{G_2}(\boldsymbol{\theta})} \tag{3.41}$$

$$\text{eff}_{\text{MH},4} = \text{eff}_{\frac{1}{M}\sum_{i=1}^{M}\theta_i} \tag{3.42}$$

$$\text{eff}_{\text{MH},5} = \text{eff}_{\theta_1} \tag{3.43}$$

where ESJD_{opt} is the ESJD in case of independent samples from the target distribution, M denotes the number of random variables in the problem, and θ_i is the ith component (out of M) of sample $\boldsymbol{\theta}$. The efficiency measures employed in Eqs. (3.42) and (3.43) are defined according to Eq. (F.15). The efficiency measure $\text{eff}_{\gamma, I_{G_k}(\boldsymbol{\theta})}$ employed in Eqs. (3.40) and (3.41) is defined according to Eq. (F.20). With $I_{G_k}(\boldsymbol{\theta})$ the indicator function of function $G_k(\boldsymbol{\theta}) = G(\boldsymbol{\theta}) - t_k$. The coefficients t_1 and t_2 are chosen such that $\text{E}\left[I_{G_k}(\boldsymbol{\theta})\right]$ equals 10% and 50%, respectively.

The following MCMC algorithms are investigated: (a) CMH in standard Normal space (Section 3.5.5) – which is in this special case equivalent to the CS algorithm proposed in [Papaioannou et al., 2015], (b) cwMH with a standard Normal proposal distribution, and (c) cwMH with a uniform proposal distribution. For all three proposal distributions, the spread s is interpreted as the standard deviation.

The MCMC sampling in the examples is done as follows: A Markov chain of length 100 is run repeatedly for each investigated algorithm, starting from a seed that already follows the target distribution. Based on a large number of such chains, the efficiency measures introduced in Eqs. (3.39) – (3.43) are evaluated.

The examples investigated in this section are:

Example 3.4 The performance of the CS algorithm and two cwMH algorithms is investi-

gated for two truncated standard Normal distributions (that have different support) as target distributions.

Example 3.5 The optimal performance of the CS algorithm is investigated for a one-dimensional truncated standard Normal distribution with decreasing support $[t, \infty)$.

Example 3.6 The performance of the CS algorithm and two cwMH algorithms is investigated for a 10-dimensional truncated standard Normal distribution. The surface of the domain of support has a non-linear shape. The employed target density is of type $p_{g_2}(\boldsymbol{\theta})$, defined according to Eqs. (3.19) and (3.24).

Example 3.7 The optimal performance of the CS algorithm and two cwMH algorithms is investigated for a 10-dimensional truncated standard Normal distribution with decreasing support $[t, \infty)$. The surface of the domain of support has a non-linear shape. The employed target density is of type $p_{g_2}(\boldsymbol{\theta})$, defined according to Eqs. (3.19) and (3.24). The maximum efficiencies obtained with the three investigated Metropolis-Hastings algorithms are compared.

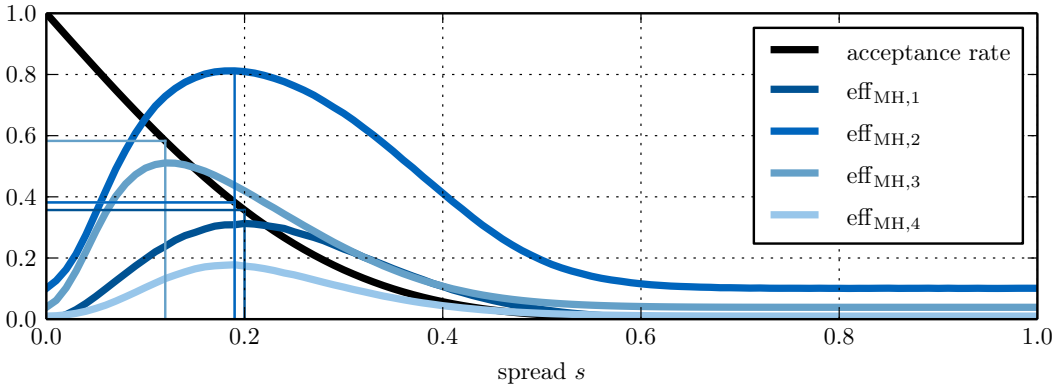
Example 3.8 The acceptance rate of the CS algorithm that leads to the best performance is investigated. The target distribution is chosen as a truncated standard Normal domain, where the surface of the domain of support has non-linear shape. The employed target density is of type $p_{g_2}(\boldsymbol{\theta})$, defined according to Eqs. (3.19) and (3.24). The dimension M of the target distribution and the domain of support of the target distribution are modified.

Example 3.9 Same as Example 3.8 with one exception: The surface of the domain of support of the target distribution is linear. The employed target density is of type $p_{g_1}(\boldsymbol{\theta})$, defined according to Eqs. (3.19) and (3.23).

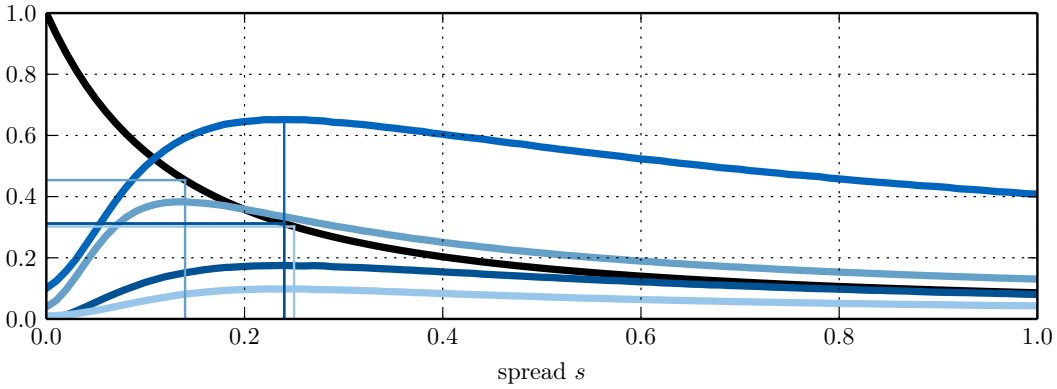
Example 3.10 Same as Example 3.8 and Example 3.9, with yet another shape of the domain of support of the target distribution. The employed target density is of type $p_{g_3}(\boldsymbol{\theta})$, defined according to Eqs. (3.19) and (3.26).

Example 3.4. *MCMC sampling of a one-dimensional truncated Normal distribution*:
A truncated Normal distribution that has support $[\beta_1; \infty)$ is investigated, where the underlying Normal distribution is a standard Normal distribution. Thus, the employed target density is of type $p_{g_1}(\boldsymbol{\theta})$, defined according to Eqs. (3.19) and (3.23), with $G_1(u) = \beta_1 - u$. Two cases are considered: $\beta_1 = 9$ and $\beta_1 = 3$. The performance of the MCMC algorithms CS, cwMH with normal proposal, and cwMH with uniform proposal is investigated for proposal spreads between *zero* and *one*.

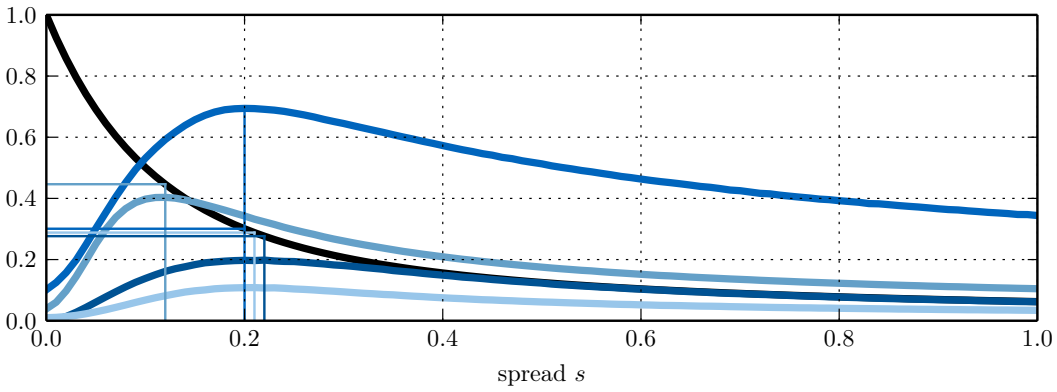
The performance of the investigated Metropolis-Hastings algorithms with respect to the efficiency measures Eqs. (3.39) – (3.43) is illustrated in Fig. 3.5 for $\beta_1 = 9$ and in Fig. 3.6 for $\beta_1 = 3$



(a) Conditional sampling in standard Normal space (standard deviation s)

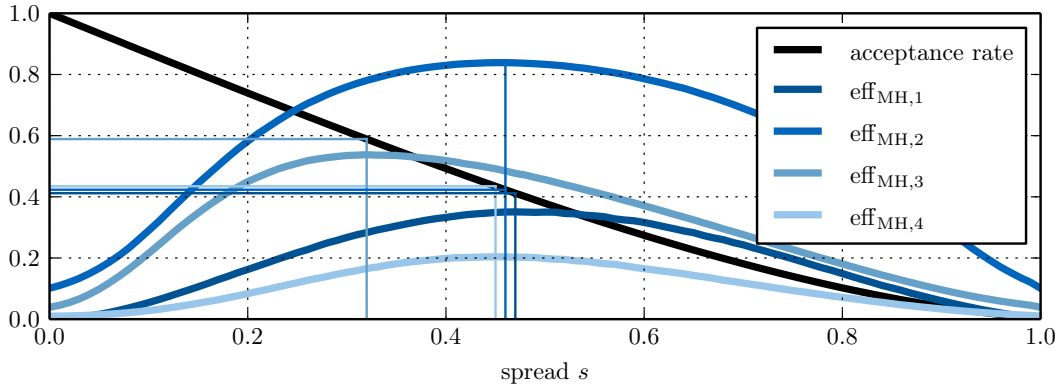


(b) cwMH with normal proposal and standard deviation s

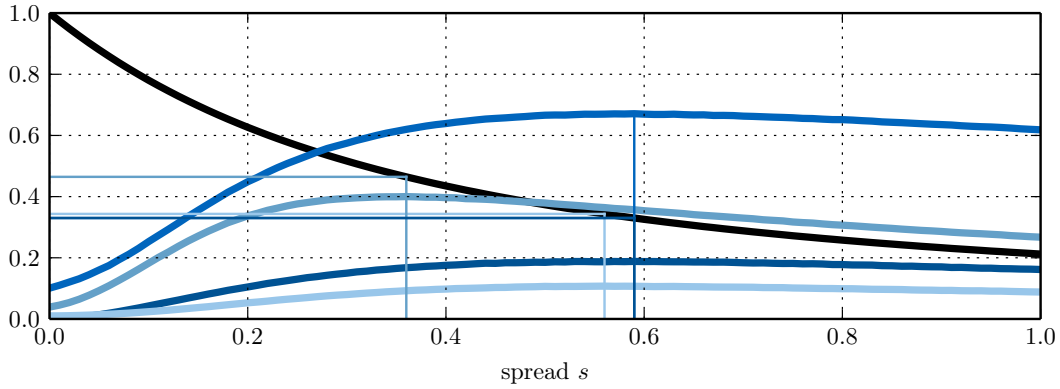


(c) cwMH with uniform proposal and standard deviation s

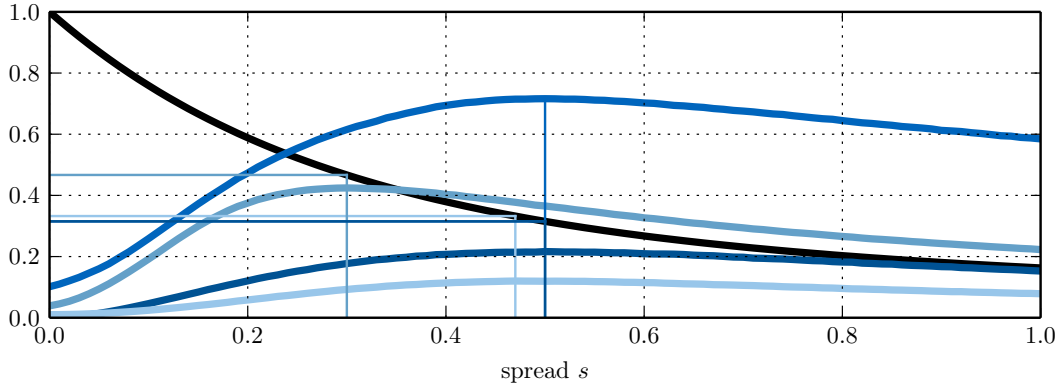
Figure 3.5: MCMC samples that follow a one-dimensional truncated standard Normal distribution ($u \geq 9$) are generated using different Metropolis-Hastings algorithms. The performance of different efficiency measures is monitored for proposal spreads between *zero* and *one*. (Example 3.4)



(a) Conditional sampling in standard Normal space (standard deviation s)



(b) cwMH with normal proposal and standard deviation s



(c) cwMH with uniform proposal and standard deviation s

Figure 3.6: MCMC samples that follow a one-dimensional truncated standard Normal distribution ($u \geq 3$) are generated using different Metropolis-Hastings algorithms. The performance of different efficiency measures is monitored for proposal spreads between *zero* and *one*. (Example 3.4)

as a function of the proposal spread. The acceptance rates and proposal spreads that are optimal with respect to the monitored efficiency measures are indicated. The efficiency measure $\text{eff}_{\text{MH},5}$ is not shown in the plots, as it coincides with $\text{eff}_{\text{MH},4}$ for one-dimensional problems.

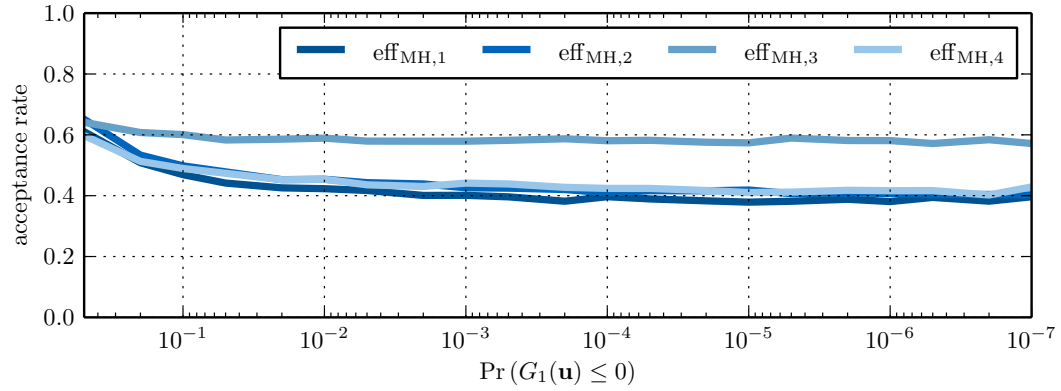
The CS algorithm clearly outperforms the two investigated cwMH algorithms: The optima of all efficiency measures are largest in the CS algorithm (this holds for $\beta_1 = 9$ as well as for $\beta_1 = 3$). For all three investigated MCMC algorithms and for the two cases $\beta_1 = 9$ and $\beta_1 = 3$, the acceptance rates that optimize the efficiency measures are approximately the same for efficiency measures $\text{eff}_{\text{MH},1}$, $\text{eff}_{\text{MH},2}$ and $\text{eff}_{\text{MH},4}$. The corresponding optimal acceptance rate is around 0.4 for CS, and around 0.3 for the two cwMH algorithms, for both $\beta_1 = 9$ and $\beta_1 = 3$. However, the acceptance rate that optimizes measure $\text{eff}_{\text{MH},3}$ is larger for both $\beta_1 = 9$ and $\beta_1 = 3$.

Example 3.5. *MCMC sampling of a one-dimensional truncated Normal distribution (cont'd)*: As in Example 3.4, a truncated standard Normal distribution on support $[\beta_1; \infty)$ is investigated; i.e., the employed target density is of type $p_{g_1}(\boldsymbol{\theta})$, defined according to Eqs. (3.19) and (3.23), with $G_1(u) = \beta_1 - u$. The optimal spread is identified for the CS algorithm and different values of β_1. The acceptance rate and the efficiency corresponding to the identified optimal spread are shown in Fig. 3.7a and Fig. 3.7b, respectively.

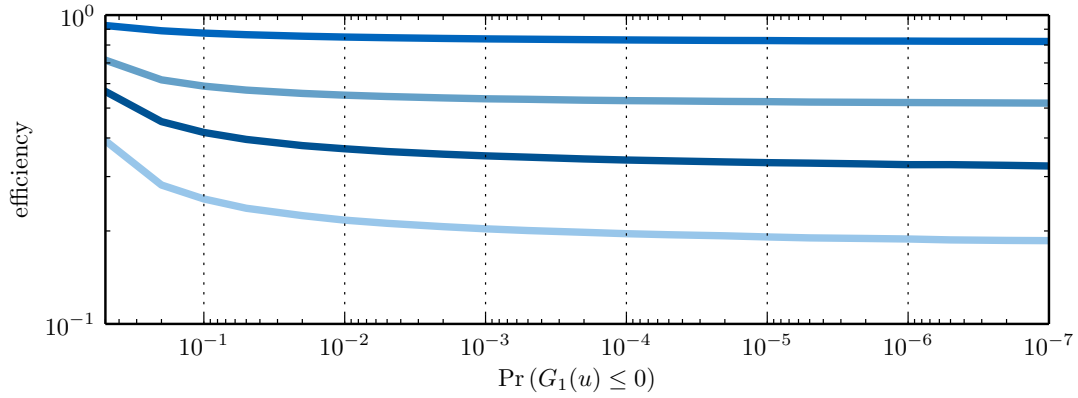
In accordance with Example 3.4, efficiency measures $\text{eff}_{\text{MH},1}$, $\text{eff}_{\text{MH},2}$ and $\text{eff}_{\text{MH},4}$ become optimal for similar acceptance rates: For $\Pr(G_1(\mathbf{u}) \leq 0) < 10^{-2}$, the optimal acceptance rate is around 0.4 in this example. In case of efficiency measure $\text{eff}_{\text{MH},3}$, the optimal acceptance rate is close to 0.6, independent of $\Pr(G_1(\mathbf{u}) \leq 0)$. Note that for $\beta_1 = \infty$, the CS algorithm generates independent samples from the target distribution if the proposal spread is set to *one*. Thus, the optimal acceptance rate is *one*, independent of the employed efficiency measure. The maximum efficiency of all investigated efficiency measures decreases with decreasing $\Pr(G_1(\mathbf{u}) \leq 0)$. The decrease in efficiency is steeper for larger $\Pr(G_1(\mathbf{u}) \leq 0)$.

Example 3.6. *MCMC sampling of a* 10*-dimensional truncated Normal distribution*:
The employed target density in this example has type $p_{g_2}(\boldsymbol{\theta})$, defined according to Eqs. (3.19) and (3.24). The parameters in Eq. (3.24) are selected as: $\alpha_2 = 3.09$, $M = 10$ and $\kappa = 10$; i.e., $\Pr(G_2(\mathbf{u}) \leq 0) \approx 10^{-3}$. As in the previous example, the performance of the MCMC algorithms CS, cwMH with normal proposal, and cwMH with uniform proposal is investigated for proposal spreads between *zero* and *one*. Again, performance is measured in terms of the efficiency measures Eqs. (3.39) – (3.43).

The results are depicted in Fig. 3.8. Contrary to Example 3.4, the optimal acceptance rates that correspond to the different investigated efficiency measures do not coincide well: Optimal acceptance rates vary between 0.2 and 0.5. Looking at individual efficiency measures, there is no distinct best MCMC algorithm for this example: At least two out of the three investigated MCMC algorithms exhibit a similar optimal performance ($\text{eff}_{\text{MH},1}$: CS and normal cwMH; $\text{eff}_{\text{MH},2}$: all three algorithms have similar optimal performance; $\text{eff}_{\text{MH},3}$: CS and normal cwMH; $\text{eff}_{\text{MH},4}$: all three algorithms have similar optimal performance; $\text{eff}_{\text{MH},5}$: CS and

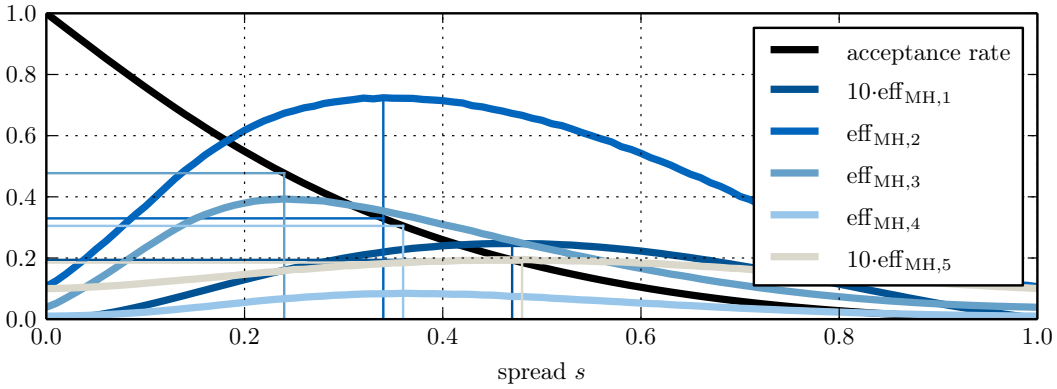


(a) Optimal acceptance rates

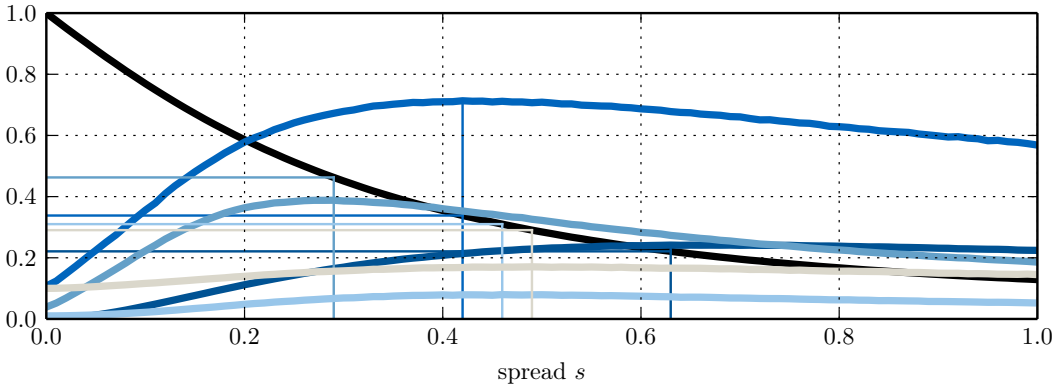


(b) Maximum efficiencies

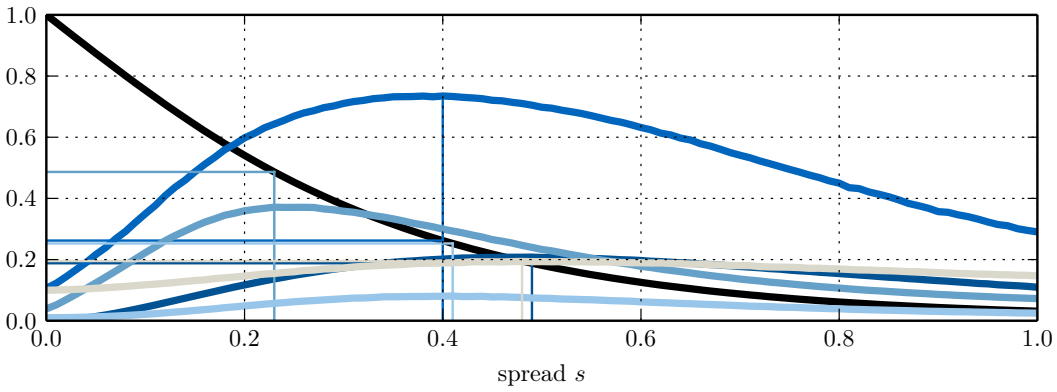
Figure 3.7: The underlying target distribution is a one-dimensional truncated standard Normal distribution. The support $[\beta_1, \infty)$ is modified by changing β_1. $\Pr(G(\mathbf{u}) \leq 0)$ denotes the probability that an independent standard Normal sample will be in $[\beta_1, \infty)$. The optimal acceptance rates and the corresponding efficiencies are shown that maximize the efficiency measures Eqs. (3.39) – (3.43) in the conditional sampling (CS) algorithm. (Example 3.5)



(a) Conditional sampling in standard Normal space (standard deviation s)



(b) cwMH with normal proposal and standard deviation s



(c) cwMH with uniform proposal and standard deviation s

Figure 3.8: MCMC samples that follow a 10-dimensional truncated standard Normal distribution (target density has type $p_{g_2}(\boldsymbol{\theta})$) are generated using different Metropolis-Hastings algorithms. The performance of different efficiency measures is monitored for proposal spreads between *zero* and *one*. (Example 3.6)

uniform cwMH). However, the CS algorithm performs most stable in all investigated efficiency measures.

As in Example 3.4, the optimal $\text{eff}_{\text{MH},3}$ / $\text{eff}_{\text{MH},1}$ exhibits the largest / smallest acceptance rate among the investigated efficiency measures, respectively.

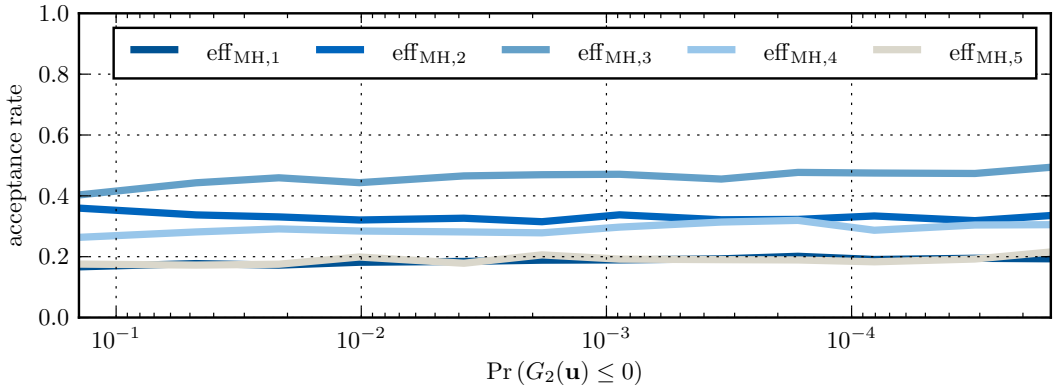
Example 3.7. *optimal MCMC spread (10-dimensional truncated Normal distribution)*:
Again, we take target density $p_{g_2}(\boldsymbol{\theta})$ defined according to Eqs. (3.19) and (3.24) with $M = 10$ and $\kappa = 10$. The spreads that maximize the efficiency measures Eqs. (3.39) – (3.43) are determined conditional on α_2. The parameter α_2 is varied and the corresponding $\Pr(G_2(\mathbf{u}) \leq 0)$ is calculated. The optimal performance of the MCMC algorithms CS, cwMH with normal proposal, and cwMH with uniform proposal is assessed.

The optimal acceptance rates corresponding to the spreads that maximize efficiency measures Eqs. (3.39) – (3.43) are shown in Fig. 3.9 for the three investigated Metropolis-Hastings algorithms. *CS algorithm:* The optimal acceptance rate of efficiency measures $\text{eff}_{\text{MH},1}$, $\text{eff}_{\text{MH},2}$ and $\text{eff}_{\text{MH},5}$ is approximately independent of $\Pr(G_2(\mathbf{u}) \leq 0)$. For efficiency measures $\text{eff}_{\text{MH},1}$ and $\text{eff}_{\text{MH},5}$, the optimal acceptance rate is approximately 0.2. For efficiency measure $\text{eff}_{\text{MH},2}$, the optimal acceptance rate is close to 0.35. The optimal acceptance rate of both efficiency measures $\text{eff}_{\text{MH},3}$ and $\text{eff}_{\text{MH},4}$ increases with decreasing $\Pr(G_2(\mathbf{u}) \leq 0)$. For efficiency measure $\text{eff}_{\text{MH},3}$, the optimal acceptance rate increases from 0.25 at $\Pr(G_2(\mathbf{u}) \leq 0) \approx 0.1$ to 0.35 at $\Pr(G_2(\mathbf{u}) \leq 0) \approx 10^{-5}$. For efficiency measure $\text{eff}_{\text{MH},4}$, the optimal acceptance rate increases from 0.4 at $\Pr(G_2(\mathbf{u}) \leq 0) \approx 0.1$ to 0.5 at $\Pr(G_2(\mathbf{u}) \leq 0) \approx 10^{-5}$. *cwMH with Normal proposal:* The optimal acceptance rates of all investigated efficiency measures depend on $\Pr(G_2(\mathbf{u}) \leq 0)$. At $\Pr(G_2(\mathbf{u}) \leq 0) = 0.14$, the optimal acceptance rate is 0.42 for all investigated efficiency measures. For efficiency measures $\text{eff}_{\text{MH},1}$, $\text{eff}_{\text{MH},2}$, $\text{eff}_{\text{MH},4}$ and $\text{eff}_{\text{MH},5}$, the optimal acceptance rate decreases with decreasing $\Pr(G_2(\mathbf{u}) \leq 0)$; where for $\Pr(G_2(\mathbf{u}) \leq 0) < 5 \cdot 10^{-3}$, the optimal acceptance rate remains constant. For efficiency measure $\text{eff}_{\text{MH},3}$, the optimal acceptance rate increases with decreasing $\Pr(G_2(\mathbf{u}) \leq 0)$. *cwMH with uniform proposal:* For $\Pr(G_2(\mathbf{u}) \leq 0) > 5 \cdot 10^{-2}$, the efficiency measures $\text{eff}_{\text{MH},1}$, $\text{eff}_{\text{MH},2}$ and $\text{eff}_{\text{MH},5}$ seem to depend on $\Pr(G_2(\mathbf{u}) \leq 0)$. However, for $\Pr(G_2(\mathbf{u}) \leq 0) < 5 \cdot 10^{-2}$, all investigated efficiency measures are approximately independent of $\Pr(G_2(\mathbf{u}) \leq 0)$.

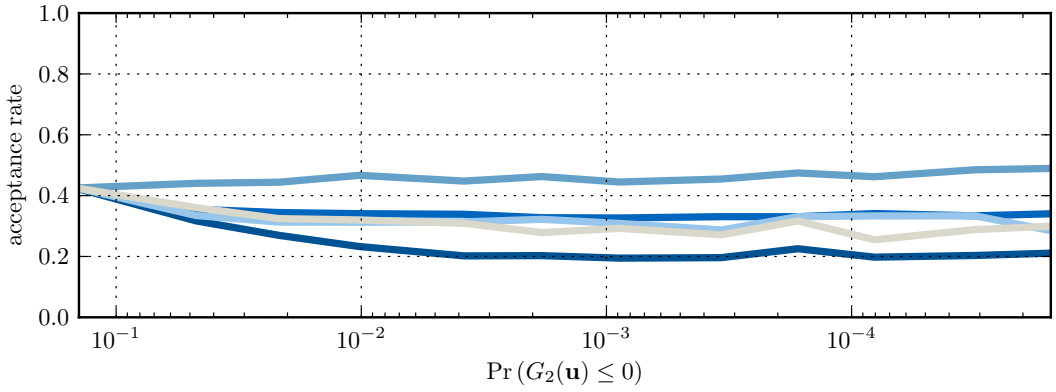
For all three investigated Metropolis-Hastings algorithms, the optimal acceptance rate of efficiency measure $\text{eff}_{\text{MH},1}$ is the smallest optimal acceptance rate among all investigated efficiency measures; whereas efficiency measure $\text{eff}_{\text{MH},3}$ has the largest optimal acceptance rate.

The maximum efficiencies obtained with the three investigated Metropolis-Hastings algorithms are compared in Fig. 3.10 as a function of $\Pr(G_2(\mathbf{u}) \leq 0)$. The efficiency of the CS algorithm is always at least as large as the efficiency of the two investigated cwMH algorithms. For efficiency measures $\text{eff}_{\text{MH},2}$, $\text{eff}_{\text{MH},3}$ and $\text{eff}_{\text{MH},4}$, the optimal efficiency is almost independent of $\Pr(G_2(\mathbf{u}) \leq 0)$. For efficiency measures $\text{eff}_{\text{MH},1}$ and $\text{eff}_{\text{MH},5}$, the optimal efficiency decreases with decreasing $\Pr(G_2(\mathbf{u}) \leq 0)$.

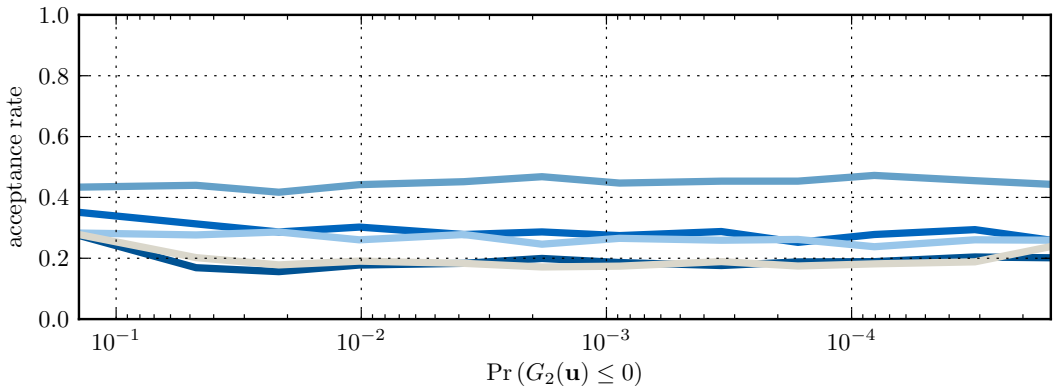
In the previous MCMC examples, the CS algorithm exhibited the most stable performance. The remaining examples in this section will focus on the CS algorithm.



(a) Conditional sampling in standard Normal space (standard deviation s)



(b) cwMH with normal proposal and standard deviation s



(c) cwMH with uniform proposal and standard deviation s

Figure 3.9: Optimal acceptance rates as a function of $\Pr(G_2(\mathbf{u}) \leq 0)$ for three different Metropolis-Hastings algorithms. The underlying target distribution is a 10-dimensional truncated standard Normal distribution that has type $p_{g_2}(\boldsymbol{\theta})$. (Example 3.7)

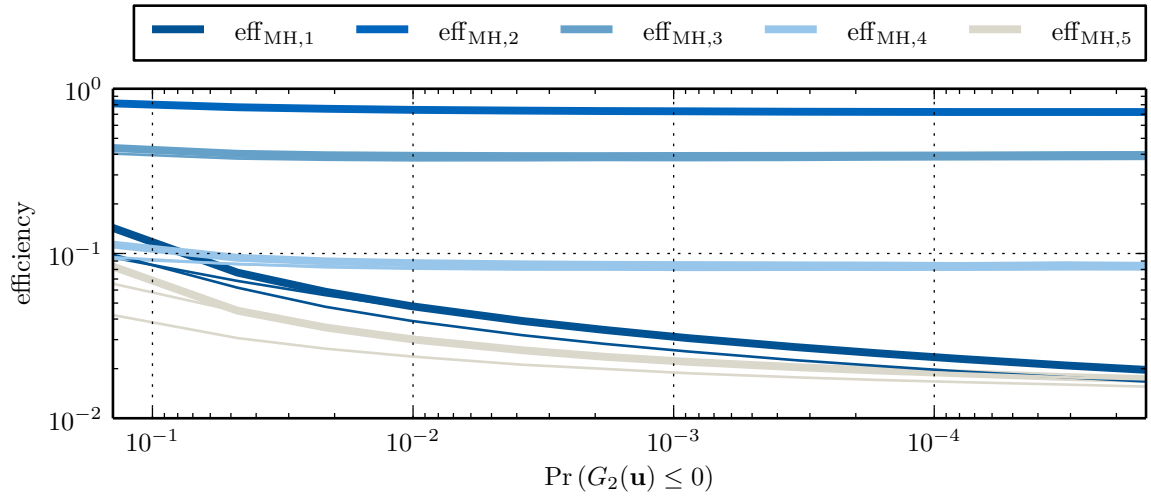


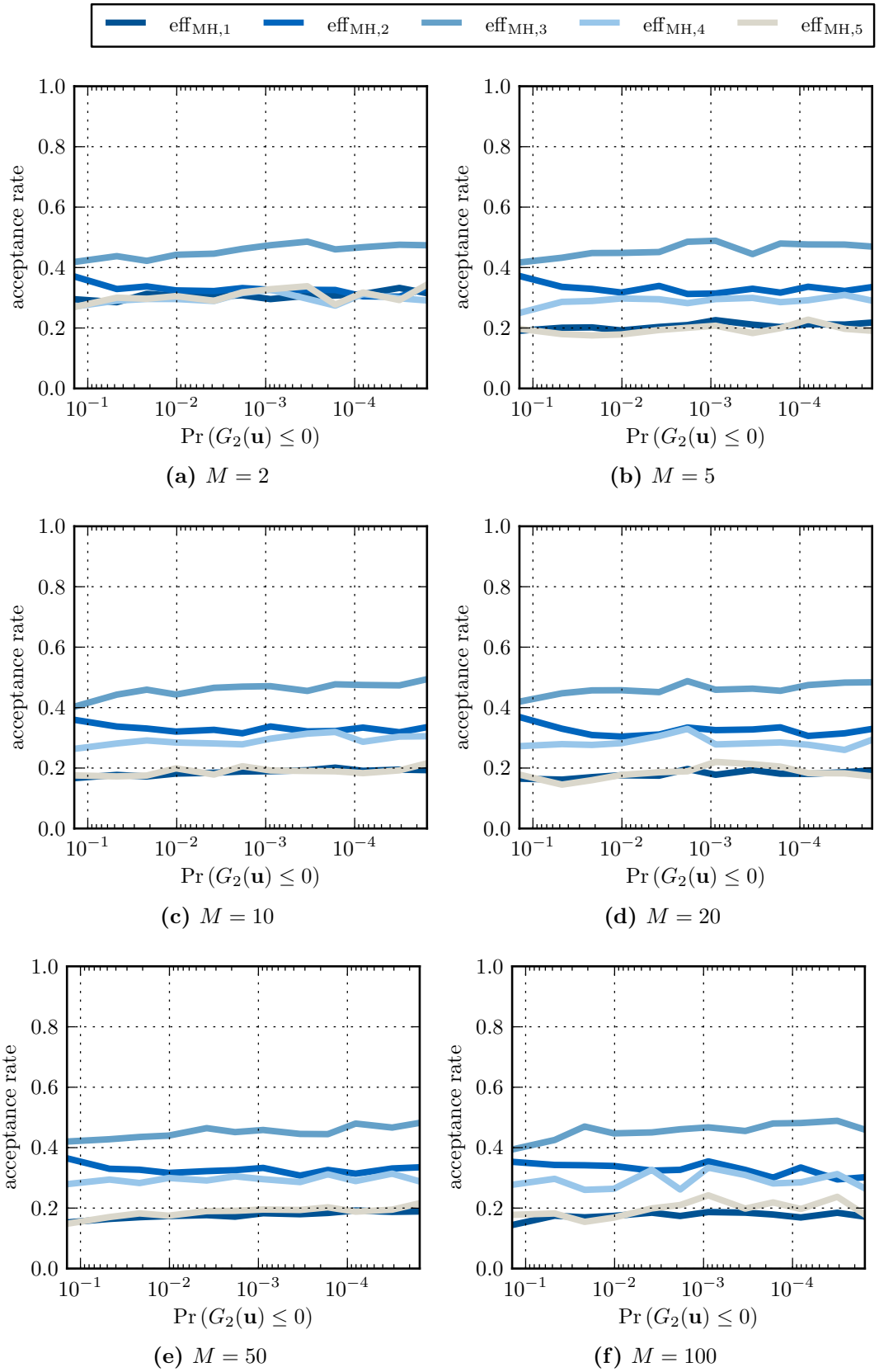
Figure 3.10: Maximum efficiencies for MCMC proposal spreads in $[0, 1]$ and target densities of type $p_{g_2}(\boldsymbol{\theta})$ with $M = 10$ and $\kappa = 10$. The bold lines correspond to the CS Metropolis-Hastings algorithm. The thin lines in lighter shades correspond to cwMH with a Normal and uniform proposal distribution. (Example 3.7)

Example 3.8. *optimal MCMC spread (M-dimensional truncated Normal distribution)*:
The previous example (Example 3.7) is slightly modified: Instead of investigating different Metropolis-Hastings algorithms for a fixed dimension ($M = 10$), only the CS algorithm is investigated for different dimensions M. The employed type of the target density is once more $p_{g_2}(\boldsymbol{\theta})$ defined according to Eqs. (3.19) and (3.24) with $\kappa = 10$.

The optimal acceptance rates that lead to the largest efficiency of the CS algorithm are illustrated in Fig. 3.11 for different M as functions of $\Pr(G_2(\mathbf{u}) \leq 0)$. For $M = 2$, the optimal acceptance rates of all efficiency measures except $\text{eff}_{\text{MH},3}$ are close together – the optimal acceptance rate is 0.3. The optimal acceptance rate for efficiency measure $\text{eff}_{\text{MH},3}$ increases slightly for decreasing $\Pr(G_2(\mathbf{u}) \leq 0)$ and varies between 0.4 and 0.5. For all M with $M \geq 5$, the behavior of the optimal acceptance rate for different $\Pr(G_2(\mathbf{u}) \leq 0)$ does not depend on M. The optimal acceptance rate of efficiency measure $\text{eff}_{\text{MH},3}$ is largest among all investigated efficiency measures – the optimal acceptance rate of $\text{eff}_{\text{MH},3}$ increases for decreasing $\Pr(G_2(\mathbf{u}) \leq 0)$ from 0.4 to 0.5. The optimal acceptance rate of efficiency measures $\text{eff}_{\text{MH},1}$ and $\text{eff}_{\text{MH},5}$ is smallest among all investigated efficiency measures; the optimal acceptance rate is slightly smaller than 0.2, independent of $\Pr(G_2(\mathbf{u}) \leq 0)$. The optimal acceptance rate of efficiency measures $\text{eff}_{\text{MH},2}$ and $\text{eff}_{\text{MH},4}$ varies between 0.25 and 0.35.

Example 3.9. *optimal MCMC spread (M-dimensional truncated Normal distribution)*:
The same study as in Example 3.8 is performed, however, the shape of the target distribution is chosen as $p_{g_1}(\boldsymbol{\theta})$ according to Eqs. (3.19) and (3.23).

The optimal acceptance rates that lead to the largest efficiency of the CS algorithm are il-



(a) $M = 2$ **(b)** $M = 5$

(c) $M = 10$ **(d)** $M = 20$

(e) $M = 50$ **(f)** $M = 100$

Figure 3.11: Optimal acceptance rates as a function of $\Pr(G_2(\mathbf{u}) \leq 0)$ for the M-dimensional truncated standard Normal distribution of type $p_{g_2}(\boldsymbol{\theta})$. The CS algorithm is employed to generate MCMC samples. (Example 3.8)

lustrated in Fig. 3.12 for different M as functions of $\Pr(G_1(\mathbf{u}) \leq 0)$. Except for $M = 1$, the optimal acceptance rates can be considered independent of M. However, contrary to Example 3.8, the optimal acceptance rates of the different efficiency measures are not constant for different $\Pr(G_1(\mathbf{u}) \leq 0)$. The optimal acceptance rates decrease for decreasing $\Pr(G_1(\mathbf{u}) \leq 0)$. For $\Pr(G_1(\mathbf{u}) \leq 0) < 10^{-3}$, the optimal acceptance rates remain approximately constant. The optimal acceptance rates of efficiency measures $\text{eff}_{\text{MH},1}$ and $\text{eff}_{\text{MH},5}$ behave similar: The optimal acceptance rate of both efficiency measures converges to a value between 0.25 and 0.3. The optimal acceptance rates of $\text{eff}_{\text{MH},1}$ and $\text{eff}_{\text{MH},5}$ are the smallest acceptance rates among all investigated efficienciey measures. Contrary to that, the optimal acceptance rate of $\text{eff}_{\text{MH},3}$ is the largest among all investigated efficiency measures. The optimal acceptance rate of $\text{eff}_{\text{MH},3}$ is slightly smaller than 0.6. The optimal acceptance rate of measures $\text{eff}_{\text{MH},2}$ and $\text{eff}_{\text{MH},4}$ converges to a value slightly larger than 0.4.

Example 3.10. *optimal MCMC spread (M-dimensional truncated Normal distribution)*:
The same study as in Example 3.8 and in Example 3.9 is performed, but with yet another shape of the domain of the truncated standard Normal distribution: The shape of the target distribution is chosen as $p_{g_3}(\boldsymbol{\theta})$, defined according to Eqs. (3.19) and (3.26).

The optimal acceptance rates that lead to the largest efficiency of the CS algorithm are illustrated in Fig. 3.13 for different M as functions of $\Pr(G_3(\mathbf{u}) \leq 0)$. The behavior of all investigated efficiency measures can be considered as independent of M. Efficiency measures $\text{eff}_{\text{MH},1}$ and $\text{eff}_{\text{MH},5}$ behave similar and exhibit the smallest optimal acceptance rate: For $\Pr(G_3(\mathbf{u}) \leq 0) < 10^{-1}$, the optimal acceptance rate is approximately 0.25 and can be considered independent of $\Pr(G_3(\mathbf{u}) \leq 0)$. Efficiency measure $\text{eff}_{\text{MH},3}$ has clearly the largest optimal acceptance rate with a value slightly smaller than 0.6. For efficiency measures $\text{eff}_{\text{MH},2}$ and $\text{eff}_{\text{MH},4}$, the optimal acceptance rate decreases with decreasing $\Pr(G_3(\mathbf{u}) \leq 0)$. The optimal acceptance rate of measure $\text{eff}_{\text{MH},2}$ converges to 0.4, and the optimal acceptance rate of measure $\text{eff}_{\text{MH},4}$ converges to a value around 0.3.

3.5.8.3 Summary

The performance of the CS algorithm and the cwMH algorithm with uniform and Normal proposal distribution was investigated for different example problems. The performance was evaluated with respect to different efficiency measures. For the investigated example problems, spread values in the neighborhood of the optimal spread[1] yield chain efficiencies close to the maximum chain efficiency; i.e., the peak of the optimum was found to be flat. The average acceptance rate that is optimal[2] is different for each efficiency measure. In all investigated example problems, the acceptance rate that maximizes efficiency measure $\text{eff}_{\text{MH},1}$ was smallest and the acceptance rate that maximizes efficiency measure $\text{eff}_{\text{MH},3}$ was largest.

[1] The *optimal spread* is the spread that optimizes the corresponding efficiency measure.
[2] The *optimal acceptance rate* is the acceptance rate for which the chain efficiency is maximized.

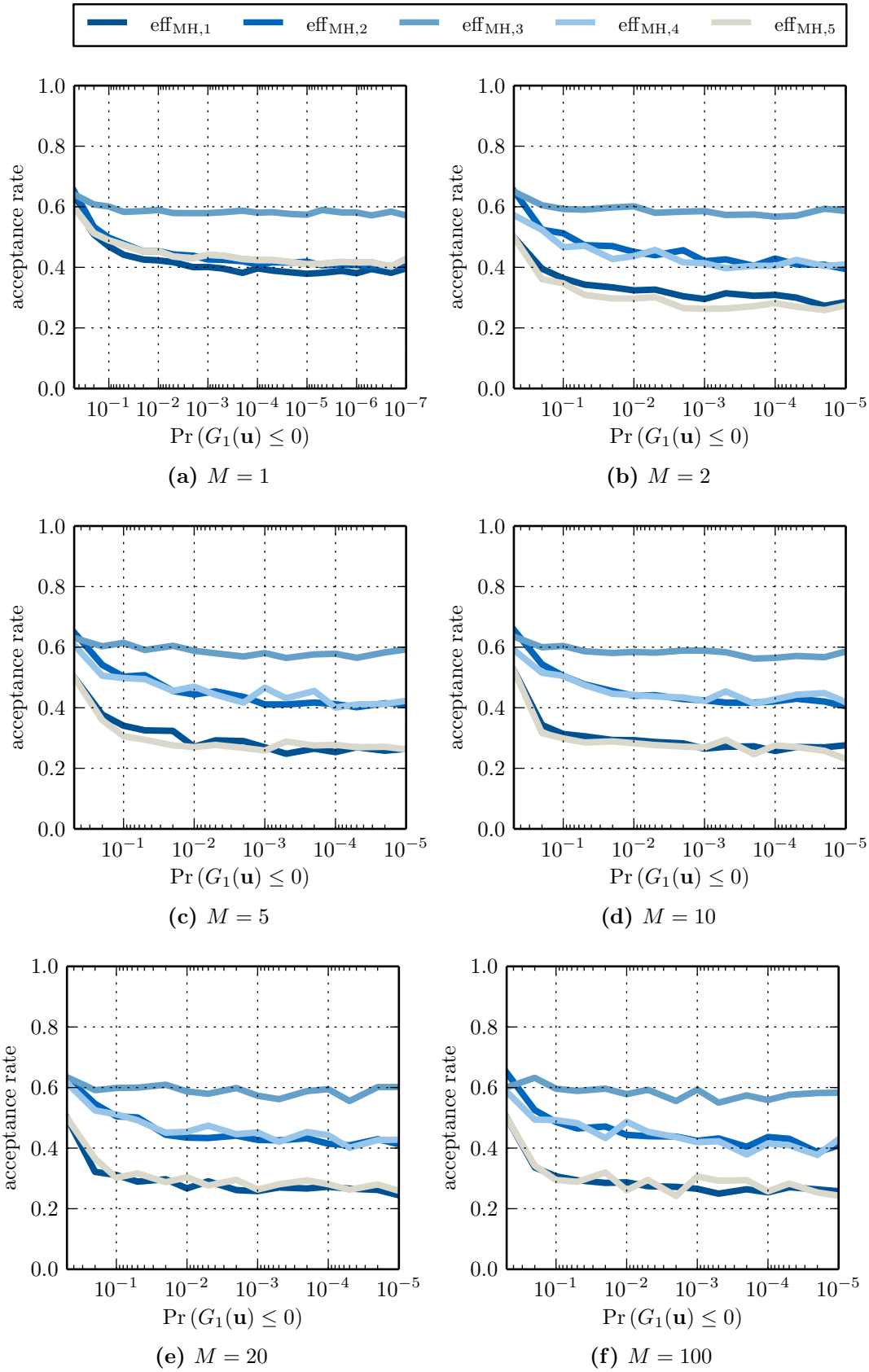
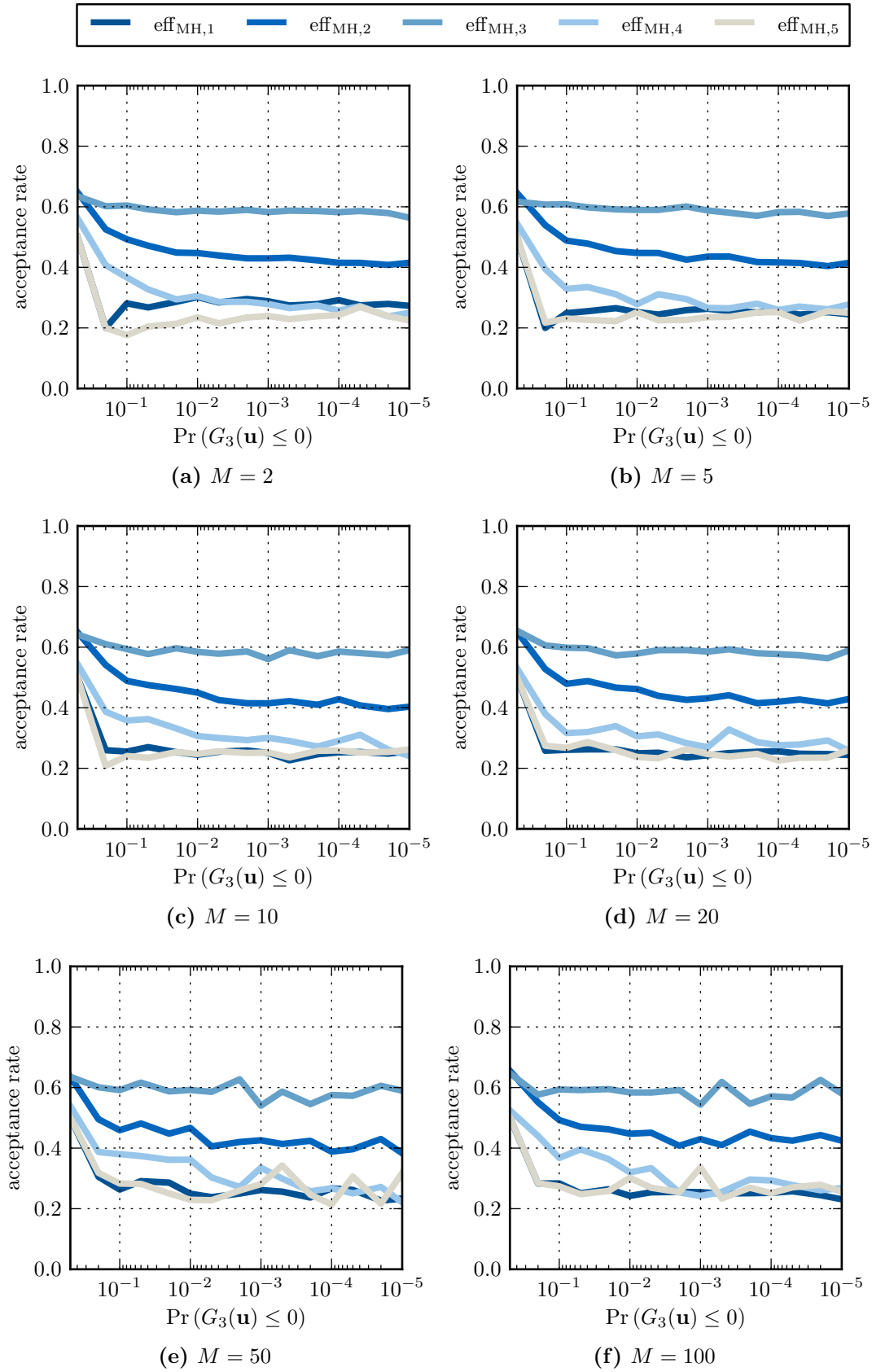


Figure 3.12: Optimal acceptance rates as a function of $\Pr(G_1(\mathbf{u}) \leq 0)$ for the M-dimensional truncated standard Normal distribution of type $p_{g_1}(\boldsymbol{\theta})$. The CS algorithm is employed to generate MCMC samples. (Example 3.9)



(a) $M = 2$ **(b)** $M = 5$

(c) $M = 10$ **(d)** $M = 20$

(e) $M = 50$ **(f)** $M = 100$

Figure 3.13: Optimal acceptance rates as a function of $\Pr(G_3(\mathbf{u}) \leq 0)$ for the M-dimensional truncated standard Normal distribution of type $p_{g_3}(\boldsymbol{\theta})$. The CS algorithm is employed to generate MCMC samples. (Example 3.10)

In Examples 3.4 – 3.7, the CS algorithm performed at least as good as the cwMH algorithm with uniform and Normal proposal distribution. In Examples 3.8 – 3.10, only the performance of the CS algorithm was assessed. For all investigated example problems, the optimal acceptance rate does not decrease with increasing dimension M.

3.5.9 Adaptive learning of the spread of the proposal

As is shown in Section 3.5.8, the spread of the proposal distribution influences the efficiency of MCMC sampling. On the one hand, if the spread is too large, the proposed sample $\mathbf{v}$ will relatively often be outside of the support domain, i.e. $g(\mathbf{v}) > 0$, and the proposed sample must be rejected often. On the other hand, if the spread is too small, the proposed sample $\mathbf{v}$ is in the vicinity of the current state of the Markov chain and is, thus, very likely inside the support domain, i.e. $g(\mathbf{v}) \leq 0$; which means the proposed sample is accepted often. In both cases the Markov chain produces highly dependent samples, and the efficiency of the chain is decreased. The efficiency of the MCMC sampling can be optimized by choosing the spread s_q of the proposal distribution such that the dependency between the generated samples is minimized. However, it is difficult to find proposal spreads s_q that lead to a minimal dependency of the generated samples from just a few Markov chains. Finding optimal proposal spreads is challenging, because:

1. There is no unique measure to quantify the efficiency of MCMC sampling (see Section 3.4.2). For different target quantities, different efficiency measures can be decisive. For example: (i) The average behavior of samples $\boldsymbol{\theta}$ might be of interest. (ii) The behavior with respect to a specific component θ_i of samples vector $\boldsymbol{\theta}$ might be relevant. (iii) The performance in terms of function $h(\boldsymbol{\theta})$ is of interest.

2. Even if the relevant efficiency measure can be formulated, it is difficult to find the spread that maximizes this efficiency measure, based on a few MCMC samples: (i) Most efficiency measures cannot reliably be estimated from just a few MCMC steps. (ii) The chain efficiency for the current spread is estimated based on the generated samples, thus, the spread that optimizes the efficiency can only be approximated conditional on the generated samples.

As a consequence, suboptimal adaption of the proposal spread through a proxy is usually employed instead [Papaioannou et al., 2015]. Two such proxies are investigated: The *expected acceptance rate* (Section 3.5.9.1) and the *expected squared jumping distance* (Section 3.5.9.2).

Some care is needed when designing algorithms to adaptively learn the spread of the proposal distribution. This algorithms use information from past chains to adapt the spread of future chains. Strictly speaking, this violates the Markovian property (not within a single chain, but looking at all the chains generated). The algorithms presented in the following try to

avoid the problem by adapting the spread primarily in the beginning, and by reducing the adaption rate the longer the simulation runs.

3.5.9.1 Adaption based on the acceptance rate

The key idea is to adapt the proposal spread such that a pre-specified *target acceptance rate* t_{acr} is maintained. The *acceptance rate* p_{acr} is defined as the average probability that the proposed state $\mathbf{v}$ of a Markov chain is accepted; i.e., $\Pr[g(\mathbf{v} \leq 0)]$. This quantity can be easily estimated even from a small number of MCMC steps: p_{acr} is approximated as the number of proposed candidates accepted divided by the total number of candidates proposed.

As the shape of the chain efficiency as a function of the proposal spread is flat around the optimal efficiency, it is not crucial to exactly know the target acceptance rate t_{acr} that results in the optimal proposal spread. Instead, it is sufficient to work with target acceptance rates t_{acr} that lead to proposal spreads close to the optimal value. [Zuev et al., 2012] recommend to select the spread of the proposal distribution such that p_{acr} is between 30% and 50% in order to minimize the dependency of the Markov chain samples. [Papaioannou et al., 2015] propose to select the spread such that $p_{\text{acr}} \approx 0.44$ is maintained; i.e., $t_{\text{acr}} = 0.44$.

An algorithm to learn a near-optimal spread of the proposal distribution adaptively during Subset Simulation is proposed in [Papaioannou et al., 2015]:

Algorithm 3.9. *Adaptive learning of the spread of the proposal [Papaioannou et al., 2015]*:
The algorithm requires as input (i) the target acceptance rate t_{acr}, (ii) the initial spread of the proposal distribution (e.g., $s_q = 1$), and (iii) the required number of samples N_q between modifying the spread adaptively (e.g., $N_q = 100$).

The algorithm modifies the spread s_q of the proposal distribution.

At the beginning of the MCMC sampling procedure, set $N_{\text{adpt}} = 1$.

After each Markov chain[1], perform the following algorithm:

1. Denote the number of samples since the spread was last modified by n. Proceed only if $n \geq N_q$.
2. Compute an estimate for the average acceptance rate p_{acr} of the MCMC sampling as the number of proposed candidates accepted divided by the total number of candidates proposed so far.
3. Only if cwMH is used[2]:

[1] This algorithms assumes that the length of the individual Markov chains is short; e.g., the length is 10. This is a relevant setting for MCMC in Subset Simulation (Section 5.3). Note: It is important not to modify the spread within a Markov chain. Modifying the spread within a Markov chain based on samples from the chain would violate the Markov property of the chain.

[2] This step is only relevant if M is small.

(a) Compute an estimate $p_{\text{acr},1}$ for the probability that a single component of the sample vector is accepted in the first stage of the two-stage approach presented in Section 3.5.2.

(b) If $p_{\text{acr},1} < p_{\text{acr}}$ and $p_{\text{acr},1} < t_{\text{acr}}$, then set $p_{\text{acr}} = p_{\text{acr},1}$.

4. Compute coefficient

$$c_q = \frac{p_{\text{acr}} - t_{\text{acr}}}{\sqrt{N_{\text{adpt}}}} \tag{3.44}$$

5. Adapt the spread as: $s_q = \exp(c_q) \cdot s_q$.

6. Increase N_{adpt} by one.

Some notes regarding *Algorithm (3.9)*:

- In [Papaioannou et al., 2015], the above algorithm was specifically proposed in combination with the CS algorithm (Section 3.5.6). However, the algorithm can directly be used to adopt the spread of arbitrary proposal distributions.
- *Algorithm (3.9)* does not destroy the Markovian property as the spread of the proposal distribution is adopted only between different Markov chains [Zuev et al., 2012]; i.e., within a single Markov chain, the spread is fixed.
- The performance of *Algorithm (3.9)* is independent of the dimension M of the target distribution.

3.5.9.2 Adaption based on the ESJD

The *expected squared jumping distance* (ESJD) (see section Section 3.4.2) can be easily estimated from previous MCMC steps: It is the sum of the squared jumping distances[1] of all previous MCMC steps divided by the total number of MCMC steps. Maximizing the ESJD minimizes the first-order autocorrelation of the chain [Thiéry, 2010].

3.5.9.2.1 Importance sampling strategy

[Pasarica and Gelman, 2010] propose to select the spread based on importance weights such that the ESJD is maximized. By successively increasing the number of MCMC steps while simultaneously adapting the spread of the proposal, the importance sampling estimate of the spread that maximizes the ESJD improves gradually. As MCMC steps with different spreads[2]

[1] The jumping distance is defined as the Euclidean distance between the current state of the Markov chain and the previous state of the Markov chain. The jumping distance is *zero*, if the proposed state of the Markov chain is rejected.

[2] Note that for a single Markov chain, the spread is kept constant.

are employed in the importance sampling estimate, [Pasarica and Gelman, 2010] propose to apply *multiple importance sampling* [Hesterberg, 1995]. In *multiple importance sampling*, a mixture distribution is employed as importance sampling distribution.

This strategy to optimize the proposal spread can be applied relatively straight-forward in combination with the CS algorithm (Section 3.5.6). In the CS algorithm, the multivariate proposal density is given explicitly. Thus, the importance weights can be evaluated easily. For the cwMH algorithm (Section 3.5.4), the procedure is more involved, as the multivariate proposal density is not stated explicitly.

The adaptive optimization of the proposal spread was implemented in combination with the CS algorithm. The following observations were made:

- The optimization of the ESJD is performed using importance sampling. The importance sampling distribution is based on the previously employed spreads. Especially for a small number of MCMC steps, the estimated spread for which the ESJD is approximately maximized can fluctuate considerably.
- The quality of the importance sampling approximation decreases the farther away the spread of interest is from the previously employed spread(s). Especially for spread values close to *zero* or *one*, the quality of the approximation can be poor. One can try to stabilize the approximation by incorporating the information that the ESJD is bound to be *zero* if the spread is *one*, and close to *zero* for a spread of *zero* and small *failure probabilities*.
- Even though the ESJD is to be maximized, it can be helpful to monitor the associated acceptance rate. Especially for few previous MCMC steps, the initially estimated maximum can be in regions where the associated acceptance rate is rather small. To avoid that the sampling gets stuck because of too large spreads, it can be beneficial to specify a bound for the smallest allowable acceptance rate.
- Such an importance sampling based strategy cannot work for large M. This can be easily demonstrated by specifically looking at the CS algorithm: For large M, the proposed samples lie effectively on the surface of a hypersphere with radius $s_q \cdot \sqrt{M}$ centered around $\sqrt{1 - s_q^2} \cdot \mathbf{w}$, where s_q is the employed proposal spread and $\mathbf{w}$ is the current state of the chain. For a proposal spread s different from s_q (for $s = s_q$, the importance ratio is *one*), the importance ratios converge asymptotically to *zero* as $M \to \infty$.

3.5.9.2.2 Curve fitting

The previously presented importance sampling strategy has problems if M is large. As an alternative, a curve fitting can be performed to approximate the ESJD as a function of

the proposal spread. The data used to learn the best fit is increased successively during the simulation: The data consists of estimated acceptance rates from the already simulated Markov chains with the previously employed spreads. For the curve fitting, appropriate function types can be selected based on the plots shown in Examples 3.4 – 3.7.

In the opinion of the author, such an adaptive variant has potential; however, it has not yet been implemented and tested at the time of writing this report.

3.5.9.3 Adaptive directional conditional sampling

The *directional conditional sampling* (DCS) Metropolis-Hastings algorithm (*Algorithm (3.8)*) requires two parameters to control the spread of the proposal distribution: ρ_{r} and ρ_{ω}. In principal, an importance sampling based strategy similar to Section 3.5.9.2 can be applied also in this case. However, in this case a two-dimensional instead of a one-dimensional optimization must be solved. This means that the problems with the importance sampling approach listed in Section 3.5.9.2 are amplified, and the practical applicability is limited.

Chapter 4

Forward Analysis

In *forward analysis* we perform a stochastic analysis conditional on our current state of knowledge. This is contrary to *Bayesian Analysis* (Chapter 6), where the goal is to incorporate new information into the stochastic analysis. The probabilistic model employed in forward analysis becomes the prior probability model for Bayesian analysis if new observations/information about the system of interest become available.

4.1 Stochastic Model Class

4.1.1 Introduction

Let $\boldsymbol{\theta} \in \Gamma,\ \Gamma \subseteq \mathbb{R}^M$ be a M-dimensional stochastic vector of *model parameters* that are uncertain in the analysis. Furthermore, we introduce $\mathcal{M}$ as a class that contains all assumptions made either explicitly or implicitly (also the ones that we are not aware of). In the following we will refer to $\mathcal{M}$ as the underlying stochastic model class of the performed stochastic analysis.

4.1.2 Definition of a stochastic model class

The concept of a *stochastic model class* was introduced by [Beck, 2010] to emphasize that no deterministic model can make perfect predictions of a real system. In a forward analysis, a stochastic model class $\mathcal{M}$ is composed of the following fundamental probability models:

(1) A *prior probability model* that assigns a relative plausibility to each state of the parameter vector $\boldsymbol{\theta}$, i.e. $p(\boldsymbol{\theta}|\mathcal{M})$.

(2) A *stochastic forward model* $p(\mathbf{r}|\mathbf{f}, \boldsymbol{\theta}, \mathcal{M})$ that expresses our belief in the plausibility that

the real system generates output $\mathbf{r}$, conditioned on a given vector of model parameters $\boldsymbol{\theta}$ and on a given model input $\mathbf{f}$. The system output $\mathbf{r}$ can for example be the outflow of water from the hydrological catchment, the inter-story accelerations of the building excited by ground motion, or the occurring settlements at the geotechnical site. Examples for the model input $\mathbf{f}$ are: the precipitation and evapotranspiration in the catchment area, the seismic ground motion, or the loading conditions at the geotechnical site. Model parameters can be curve number and soil moisture for the hydrological model, stiffness, mass and damping matrices for the structural model, Young's modulus, Poisson's ratio and cohesion for the geotechnical model.

(3) Additionally, if the model input $\mathbf{f}$ is uncertain, a *stochastic input model* $p(\mathbf{f}|\mathbf{s}, \boldsymbol{\theta}_\mathbf{f}, \mathcal{M})$ is required to quantify our belief in the input uncertainty, where $\boldsymbol{\theta}_\mathbf{f}$ is the parameter vector of the input model and $\mathbf{s}$ comprises the information available about the model input ($\boldsymbol{\theta}_\mathbf{f}$ is a vector that contains selected components of $\boldsymbol{\theta}$).

Note that the probability models that define the stochastic model class $\mathcal{M}$ represent the state of plausible knowledge about the system conditional on the available (incomplete) information, and, thus, they are not inherent properties of the system [Beck, 2014]. All predictions are conditional on the selected stochastic model class $\mathcal{M}$. Consequently, the quality of the predictions with respect to the real system is also conditional on $\mathcal{M}$, and, thus, the predictions can only be as good as the validity of the assumptions in $\mathcal{M}$.

4.1.3 Stochastic embedding

The system of interest can, for example, be a structure (e.g., a building, a bridge or a tunnel), a machine (e.g., a car, a plane or a ship) or a part of a machine, or an environmental system (e.g., a hydrological catchment or an ecosystem). Commonly, we cannot directly express the response of the targeted system. In such cases, the system of interest is represented by a model. Typically, a parametrized deterministic model is available that approximates the behavior of the real system. As deterministic model, often a numerical model (e.g., a finite element model) is employed in the forward analysis.

A parametrized deterministic model can be used as a basis to derive a stochastic model class; a procedure that is referred to as *stochastic embedding* [Beck, 2010]: The stochastic forward model $p(\mathbf{r}|\mathbf{f}, \boldsymbol{\theta}, \mathcal{M})$ is expressed as: $\mathbf{r} = \mathbf{q}(\mathbf{f}, \boldsymbol{\theta}_\mathbf{q}) + \mathbf{v}$, where $\mathbf{r}$ is the (unknown) real system output, $\mathbf{q}(\mathbf{f}, \boldsymbol{\theta}_\mathbf{q})$ denotes the output of the deterministic model with $\boldsymbol{\theta}_\mathbf{q}$ as the parameter vector of the deterministic model ($\boldsymbol{\theta}_\mathbf{q}$ contains selected components of $\boldsymbol{\theta}$) and $\mathbf{f}$ as the model input, and $\mathbf{v}$ is the *output prediction-error* that is uncertain. Thus, the PDF of the stochastic forward model can be expressed as $p(\mathbf{r}|\mathbf{q}(\mathbf{f}, \boldsymbol{\theta}_\mathbf{q}), \boldsymbol{\theta}_\mathbf{v}, \mathcal{M})$, where $\boldsymbol{\theta}_\mathbf{v}$ is the parameter vector of the *prediction-error model*. Alternatively, instead of the additive error structure, a multiplicative error structure could be selected, i.e., $\mathbf{r} = \mathbf{q}(\mathbf{f}, \boldsymbol{\theta}_\mathbf{q}) \cdot \mathbf{v}$. The *output*

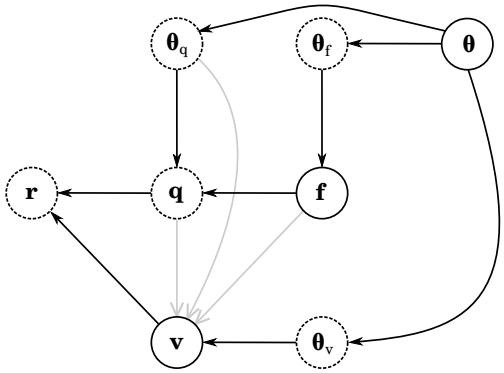


Figure 4.1: Assumed dependency structure in a stochastic model class (as discussed in Section 4.1) represented as a Bayesian network. The dependencies are represented as arrows. The gray arrows denote dependencies that exist in the real world but that are usually not considered explicitly in modeling. The nodes with a continuous border denote random variables. The nodes with a dashed border can be computed deterministically if all input quantities are conditionally fixed. The parameter vector $\boldsymbol{\theta}$ is composed of $\boldsymbol{\theta} = \left[\boldsymbol{\theta}_\mathtt{q}^\mathsf{T}, \boldsymbol{\theta}_\mathtt{f}^\mathsf{T}, \boldsymbol{\theta}_\mathtt{v}^\mathsf{T}\right]^\mathsf{T}$, where $\boldsymbol{\theta}_\mathtt{q}$ is the vector of uncertain parameters of the deterministic model, $\boldsymbol{\theta}_\mathtt{f}$ is the parameter vector of the *input model*, and $\boldsymbol{\theta}_\mathtt{v}$ is the parameter vector of the *prediction-error model*.

prediction-error $\mathbf{v}$ quantifies the inability of the deterministic model to predict the response of the true system perfectly; it rates how plausible a particular response $\mathbf{r}$ of the true system is given model output $\mathbf{q}(\mathbf{f}, \boldsymbol{\theta}_\mathtt{q})$.

An exemplary dependency structure of a stochastic forward model is represented as a Bayesian network in Fig. 4.1.

4.1.4 Imperfect models

No deterministic model produces a perfect representation of reality. The mismatch between the model output and the actual system response is referred to as the *output prediction-error* (see Section 4.1.3). The smaller this error, the better the model can predict the response of the actual underlying real-world system. An exact probabilistic quantification of this error is, however, virtually impossible: (i) It depends on the history of actual system inputs (which can usually not be observed exactly) and on the considered model inputs (even system input regarded as negligible influences the modeling error). (ii) It is coupled to a specific system/-model and cannot directly be transferred to similar systems/models. (iii) It has to consider all assumptions and simplifications of the model (even the ones that were not done on purpose). (iv) It is tightly linked to a specific output quantity of interest. Consequently, different output quantities of interest have different *output prediction-errors*. (v) The probabilistic structure depends on both time and space.

4.1.5 Unknown unknowns

The stochastic model class $\mathcal{M}$ contains all underlying assumptions and is based on our knowledge and expertise. Naturally, uncertainties that we are not aware of or deemed negligible, are not incorporated in the model and are, thus, not considered. It is important to keep in mind: The stochastic model class $\mathcal{M}$ represents our uncertainty about the state of the represented system, and, as such, is not immune to human error.

In this context, one typically speaks of *Black Swan* events that are defined as follows [Taleb, 2007]: (i) It is an event that could not have been expected based on past experiences (*rarity*). (ii) The event has an *extreme impact*. (iii) In *retrospective* we can come up with reasons that make it explainable and predictable. It is important to understand that also *Black Swan* events are conditional on our current state of knowledge. For somebody else, such a surprising event of large impact might not be that surprising after all. A discussion of *Black Swan* events can be found e.g., in [Taleb, 2007; Aven, 2014].

4.2 Simulating the probabilistic model response

Typically, in *forward analysis*, independent samples $\boldsymbol{\theta}$ of the stochastic vector of model parameters can be generated. This allows the use of *Monte Carlo simulation* to generate samples $\mathbf{r}$ of the system output:

Algorithm 4.1. *Monte Carlo Simulation to probabilistically simulate the model response*: This algorithm generate N samples $\{\mathbf{r}_i\}_{i=1,\dots,N}$ of the system output.

For $i = 1, \dots, N$, do:

1. Generate an independent sample $\boldsymbol{\theta}_i$ of the stochastic vector of model parameters.
2. If the model input $\mathbf{f}$ is uncertain, generate an independent realization $\mathbf{f}_i$ of $\mathbf{f}$ (conditional on $\boldsymbol{\theta}_i$).
3. Evaluate the response of the model $\mathbf{q}_i = \mathbf{q}(\mathbf{f}_i, \boldsymbol{\theta}_{\mathbf{q}_i})$ conditional on $\boldsymbol{\theta}_i$ and $\mathbf{f}_i$.
4. Generate a sample $\mathbf{v}_i$ of the *output prediction-error* conditional on $\boldsymbol{\theta}_i$, $\mathbf{q}_i$ and $\mathbf{f}_i$.
5. A sample of the system response is: $\mathbf{r}_i = \mathbf{q}_i + \mathbf{v}_i$

Note that samples of the system response $\mathbf{r}_i$ represent our belief about the potential (true) system response.

Based on the generated N samples $\{\mathbf{r}_i\}_{i=1,\dots,N}$ the e.g., mean, variance and quantiles of the response can be estimated.

On the one hand, *Monte Carlo simulation* is a simple and robust method to generate independent and unweighted samples in *forward analysis*. On the other hand, if higher moments or very small/large quantiles are to be estimate, many samples are needed to get a good estimate. If the number of required samples becomes large, *Monte Carlo simulation* becomes inefficient: Typically, it is computationally demanding to evaluate the model response $\mathbf{q}(\mathbf{f}_i, \boldsymbol{\theta}_{\mathbf{q}_i})$. In such cases, more advanced sampling methods might be more efficient; e.g., *importance sampling*, which generates weighted samples based on a proposal distribution.

4.3 Credible intervals

4.3.1 Definition

Let X be the stochastic quantity of interest. X can be for example a stochastic model parameter of interest, the response of the forward model, or a function that depends on the response of the forward model. Furthermore, let $p = \Pr(a \leq X \leq b|\mathcal{M})$, with $a \leq b$. If X is a scalar quantity, $[a, b]$ is referred to as the p credible interval for X. If X is a vector quantity, $[a, b]$ is referred to as the p credible region for X.

The p credible interval (or region) $[a, b]$ for X states that conditional on $\mathcal{M}$, the *true* value of X is with probability p within $[a, b]$. Note that there is not one unique p credible interval $[a, b]$ for X: For a fixed p, the relation $p = \Pr(a \leq X \leq b|\mathcal{M})$ can be met with different intervals $[a, b]$. Commonly, one of the following p credible intervals is used:

equal-tailed interval The *equal-tailed* p credible interval $[a, b]$ is defined as $a = P_X^{-1}(\frac{1-p}{2})$ and $b = P_X^{-1}(1 - \frac{1-p}{2})$, where $P_X^{-1}(\cdot)$ is the inverse CDF for X. This credible interval is used most-often, as it is straight-forward to compute, given $P_X^{-1}(\cdot)$.

lower tail interval The p credible interval $(-\infty, b]$ for the lower tail is defined as $b = P_X^{-1}(p)$. This credible interval is employed to state that conditional on $\mathcal{M}$, the *true* value of X is with probability p not larger than b.

upper tail interval The p credible interval $[a, \infty)$ for the upper tail is defined as $a = P_X^{-1}(1-p)$. This credible interval is employed to state that conditional on $\mathcal{M}$, the *true* value of X is with probability p not smaller than a.

highest density interval Amongst all viable p credible intervals, the interval that contains the highest density values. For univariate unimodal densities, it is the interval for which the length $b - a$ is minimized.

Table 4.1: Ambiguity of p credible intervals (see Example 4.1).

	p credible interval		
	$p = 95\%$	$p = 90\%$	$p = 50\%$
equal-tailed interval	$[0.14, 3.62]$	$[0.18, 2.78]$	$[0.40, 1.24]$
lower-tail interval	$(0, 2.78]$	$(0, 2.06]$	$(0, 0.71]$
upper-tail interval	$[0.18, \infty)$	$[0.24, \infty)$	$[0.71, \infty)$
highest density interval	$[0.04, 2.79]$	$[0.06, 2.07]$	$[0.16, 0.77]$

Example 4.1. *Ambiguity of p credible intervals.*:
Let X be a log-Normal stochastic variable that has mean *one* and standard deviation *one*. Different 95%, 90% and 50% credible intervals are given in Table 4.1.

4.3.2 Comparison to confidence intervals

In frequentist statistics, typically *confidence intervals* are employed; whereas in a Bayesian framework, *credible intervals* are used. For some simple problems, credible and confidence intervals can coincide. However, the meaning of credible and confidence intervals is fundamentally different:

A p *credible interval* $[a, b]$ for stochastic quantity X states that the unknown *true* value of X will be with probability p within the specified interval. This statement is meaningful for a single simulation (e.g., N generated samples in a forward analysis). It quantifies the uncertainty about X based on the available information.

A p *confidence interval* $[a, b]$ for stochastic quantity X states that if a simulation (e.g., N generated samples in a forward analysis) is repeated many times and each time the confidence interval $[a, b]$ is evaluated anew, the *true* value will lie in $p \cdot 100\%$ of the obtained intervals. This implicitly assumes that the simulation can be repeated arbitrarily often: The confidence interval is varying in each simulation, the frequency of the true value being inside of $[a, b]$ is p.

For a discussion of *credible* and *confidence intervals* see [Jaynes and Kempthorne, 1976].

4.4 Reliability analysis

4.4.1 Introduction

Reliability analysis aims at evaluating the probability of failure of a system of interest. In this context, failure is defined as the system being in an undesired state, e.g.: admissible

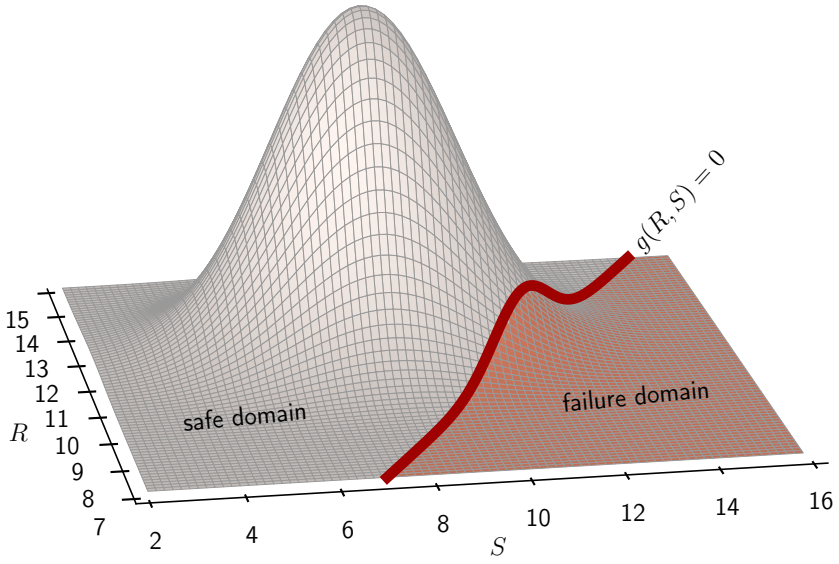


Figure 4.2: Failure domain and safe domain for demand S (Normal, $\mu = 8$, $\sigma = 2$) and capacity R (Normal, $\mu = 12$, $\sigma = 1$). Performance of the system is expressed in terms of limit-state function $g(R, S) = R - S$.

stresses are exceeded, the stability of a structure is no longer maintained, or water levels in a river that exceed a certain threshold will result in a flood event. Let such an undesired system response be denoted as proposition $\mathcal{F}$, where the probability of failure P_f is defined as the probability that $\mathcal{F}$ occurs, i.e., $P_f = \Pr(\mathcal{F}|\mathcal{M})$. Note that the outcome of the reliability analysis is conditional on the stochastic model class $\mathcal{M}$.

For a comprehensive treatise on structural reliability, the reader is referred to [Ditlevsen and Madsen, 2007; Melchers, 1999].

4.4.2 Limit-state function – formulation of the reliability problem

Let $g(\boldsymbol{\theta})$ be a function such that $g(\boldsymbol{\theta}) \leq 0$ if and only if the system is in an undesired state, and $g(\boldsymbol{\theta}) > 0$ otherwise, where $\boldsymbol{\theta}$ is the stochastic vector of uncertain model parameters. The function $g(\boldsymbol{\theta})$ is known as *limit-state function* or *performance function* in the literature. The probability of failure P_f can then be expressed as:

$$P_f = \Pr\left[g(\boldsymbol{\theta}) \leq 0\right] = \int_{\Gamma_f} p(\boldsymbol{\theta}|\mathcal{M})\, \mathrm{d}\boldsymbol{\theta} \tag{4.1}$$

where Γ_f is the *failure domain* defined as $\Gamma_f = \{\boldsymbol{\theta} \in \Gamma | g(\boldsymbol{\theta}) \leq 0\}$.

The exact state of vector $\boldsymbol{\theta}$ is uncertain, due to our imperfect knowledge/understanding of the real world. The task of the person conducting the reliability analysis is to assign a plausibility

to each state of $\boldsymbol{\theta}$ by means of the joint probability density function $p(\boldsymbol{\theta}|\mathcal{M})$.

The formulation of the limit-state function in case demand and capacity of the investigated system can be separated is discussed in Section 4.4.5.

For the definition of the undesired system response, one typically distinguishes between the ultimate, damage and the serviceability limit-state [Melchers, 1999], where *ultimate* refers to at least partial collapse of the structure, and *serviceability* means disruption of normal use.

4.4.3 Reliability index

The probability of failure P_f is a measure for the degree of safety (i.e., the reliability) of the investigated structure. The smaller the probability of failure, the larger is the reliability. The probability of failure P_f usually ranges from 10^{-1} to 10^{-7} [Kiureghian, 1989][1]. An often more convenient measure to express the reliability of the investigated structure is provided by the reliability index. The *generalized reliability index* β is defined as [Ditlevsen, 1979; Ditlevsen and Madsen, 2007, Chapter 6]:

$$\beta = -\Phi^{-1}(P_f) \tag{4.2}$$

where Φ^{-1} is the inverse of the CDF of the standard Normal distribution. The reliability index β increases with decreasing probability of failure; i.e., the larger β the larger is the reliability of the investigated structure. For $P_f = 10^{-1}$ the reliability index is approximately 1, and for $P_f = 10^{-7}$ the reliability index is approximately 5. The relation between the reliability index β and the probability of failure P_f is illustrated in Fig. 4.3.

4.4.4 Design point

The *design point* $\boldsymbol{\theta}^*$ is the point in the failure domain $\Gamma_f = \{\boldsymbol{\theta} \in \Gamma | g(\boldsymbol{\theta}) \leq 0\}$ that has the smallest distance to the origin in standard Normal space[2]; i.e.,

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta} \in \Gamma_f}{\arg\min} \left(\| \mathrm{T}(\boldsymbol{\theta}) \| \right) \tag{4.3}$$

[1] In [Kiureghian, 1989], the uncertainty in estimating the reliability index is discussed. Different formulations for the reliability index are presented. However, the interpretation of uncertainty taken in [Kiureghian, 1989] is different from the one adopted in this thesis. [Kiureghian, 1989] consider the probability of failure and the reliability index as actual properties of the investigated system. [Kiureghian, 1989] impose that under a perfect state of knowledge, this quantities can be estimated directly. For practical problems, the state of knowledge is, however, invariably imperfect [Kiureghian, 1989]. Thus, the *true* probability of failure and the *true* reliability index can only be estimated under uncertainty.

In the view adopted in this thesis, a perfect state of knowledge would result in a probability of failure of either *zero* or *one*. Under a perfect state of knowledge, there is no intrinsic uncertainty anymore.

[2] The concept of the underlying standard Normal space is explained in Section 3.2.

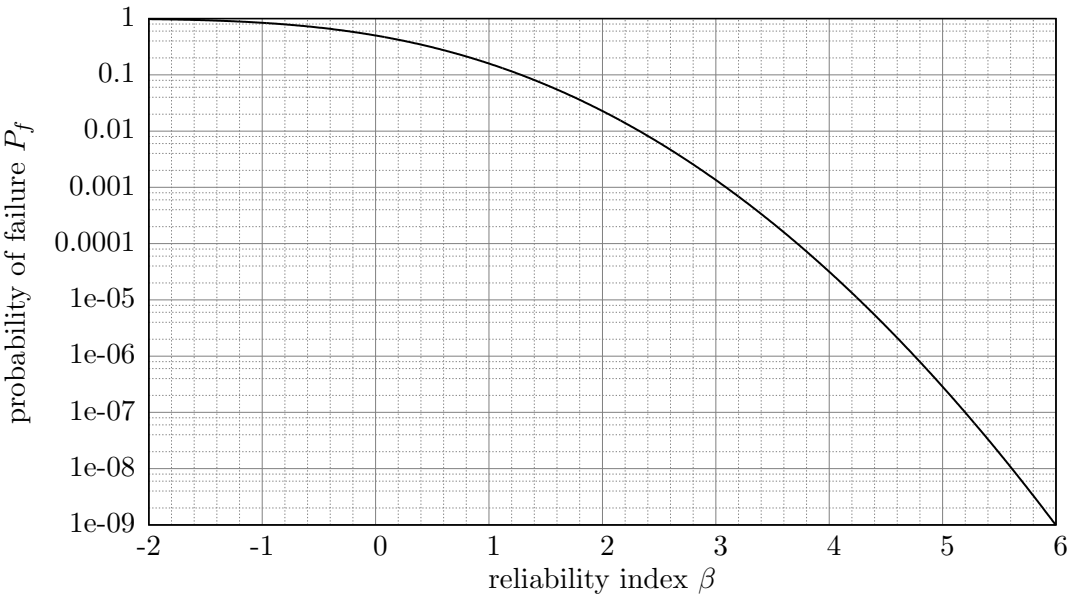


Figure 4.3: Relation between the reliability index β and the probability of failure P_f.

which can equivalently be expressed as

$$\mathbf{u}^* = \underset{\mathbf{u}\in\mathbf{U}_f}{\arg\min} \left(\|\mathbf{u}\|\right) \tag{4.4}$$

with $\boldsymbol{\theta}^* = \mathrm{T}^{-1}(\mathbf{u}^*)$, and $\mathbf{U}_f = \{\mathbf{u} \in \mathbb{R}^N | g^*(\mathbf{u}) \leq 0\}$.

4.4.5 Formulation of the limit-state function in terms of demand and capacity

For some problems the limit-state function can be expressed in terms of demand S and capacity R of the system of interest. Failure occurs if the demand exceeds the capacity, i.e.:

$$g(\boldsymbol{\theta}) = R(\boldsymbol{\theta}) - S(\boldsymbol{\theta}) \tag{4.5}$$

The difference $R - S$ is also referred to as *safety margin*. If the distribution of both R and S is known, the reliability problem stated in Eq. (4.1) can equivalently be expressed as:

$$P_f = \Pr\left[g(\boldsymbol{\theta}) \leq 0\right] = \int_\Gamma P_R(\boldsymbol{\theta}) \cdot p_S(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \tag{4.6}$$

Different ways to express the limit-state function in terms of R and S are shown in [Melchers, 1999, Section 1.4.2].

For some systems, a separation of demand and capacity is usually not feasible (e.g., due to soil-structure interaction in tunneling). If demand and capacity cannot be separated easily,

the limit-state function is often defined as the difference between some threshold value and the corresponding model output; e.g., the displacement at the tip of a cantilever beam versus the maximum allowed displacement for that system.

Example 4.2. *Demand and capacity follow a Normal distribution*:
If both the demand S and capacity R in Eq. (4.5) follow a Normal distribution, $g(\boldsymbol{\theta})$ is a random variable that follows a Normal distribution as well. Let μ_S and σ_S be the mean and standard deviation of S, μ_R and σ_R be the mean and standard deviation of R. The mean and standard deviation of $g(\boldsymbol{\theta})$, denoted μ_g and σ_g, is $\mu_g = \mu_R - \mu_S$ and $\sigma_g = \sqrt{\sigma_R^2 + \sigma_S^2}$ if R and S are independent. In this special case, the probability of failure can be expressed explicitly:

$$P_f = \Pr\left(g(\boldsymbol{\theta}) \leq 0\right) = \Phi\left(\frac{\mu_S - \mu_R}{\sqrt{\sigma_R^2 + \sigma_S^2}}\right) \tag{4.7}$$

where $\Phi(\cdot)$ denotes the CDF of the standard Normal distribution (Section B.2.1). The associated reliability index is:

$$\beta = \frac{\mu_R - \mu_S}{\sqrt{\sigma_R^2 + \sigma_S^2}} \tag{4.8}$$

Example 4.3. *Demand and capacity follow a log-normal distribution*:
If both the demand S and capacity R in Eq. (4.5) follow a log-normal distribution and are independent, $g(\boldsymbol{\theta})$ can be expressed as $g(\boldsymbol{\theta}) = R(\boldsymbol{\theta})/S(\boldsymbol{\theta}) - 1$. In this case we can write $\ln(g+1) = \ln(R) - \ln(S)$, where $\ln(g+1)$ clearly follows a Normal distribution. Let λ_S and ζ_S be the parameters of S, and λ_R and ζ_R be the parameters of R (compare Section B.2.4). Note that R/S has a log-normal distribution with parameters $\lambda = \lambda_R - \lambda_S$ and $\zeta = \sqrt{\zeta_R^2 + \zeta_S^2}$. In this special case, the probability of failure can be expressed explicitly:

$$P_f = \Phi\left(\frac{\lambda_S - \lambda_R}{\sqrt{\zeta_R^2 + \zeta_S^2}}\right) \tag{4.9}$$

where $\Phi(\cdot)$ denotes the CDF of the standard Normal distribution (Section B.2.1). The associated reliability index is:

$$\beta = \frac{\lambda_R - \lambda_S}{\sqrt{\zeta_R^2 + \zeta_S^2}} \tag{4.10}$$

4.4.6 Reference time period

The probability of failure P_f and the associated reliability index β are always in reference to a period of time. Often an annual reference frame is picked. However, also a 50- or 100-year reference frame could be chosen. When formulating the uncertain model parameters, it is important to be clear about the reference period. The chosen reference period is most important for the parameters describing the uncertainties about the applied loading. It makes

a difference whether the uncertainty is expressed in terms of the largest annual load or in terms of the maximum load within a 50-year reference period.

If independence between the individual years is assumed, one can compute the reliability index $\beta_{1,\text{independence}}$ in an annual reference frame from the reliability index β_{50} in a 50-year reference period:

$$\beta_{1,\text{independence}} = \Phi^{-1}\left[(\Phi(\beta_{50}))^{\frac{1}{50}}\right] \tag{4.11}$$

where $\Phi(\cdot)$ denotes the CDF of the standard Normal distribution, and Φ^{-1} its inverse function. Typically it can be assumed that the maximum annual loads are statistically independent. However, for the resistance model, it is typically implicit assumed that the resistance does not change during service life.

Eq. (4.11) is valid if the coefficient of variation of the uncertainties in the resistance model is small compared to the coefficient of variation of the uncertainties about the loading; i.e., $\beta_1 = \beta_{1,\text{independence}}$. If the uncertainties in the resistance model are not small, then $\beta_1 < \beta_{1,\text{independence}}$, where in general β_1 is bounded by $\beta_{50} \leq \beta_1 \leq \beta_{1,\text{independence}}$. This is why the required target reliability index[1] $\beta_{1,\text{target}}$ can be set a bit larger than $\beta_{1,\text{independence,target}}$ obtained from the required target reliability index $\beta_{50,\text{target}}$. Required target reliability indices for an annual and a 50-year reference period are specified in [Eurocode 0, 2015, Annex B] for different reliability classes[2].

4.4.7 Imperfect models

Due to the *output prediction-error* (see Section 4.1), if the state of the model is considered acceptable, it does not necessarily mean that the system is actually in a desirable state – and vice versa. A viable strategy to probabilistically resolve this issue is to take the (uncertain) *output prediction-error* $\mathbf{v}$ explicitly into account. However, it is far from trivial to express this error probabilistically: Our information about the accuracy of the employed models is commonly very limited.

For example, for civil engineering structures, [Eurocode 0, 2015] requires that modeling uncer-

[1] *Target reliability index* refers to the reliability index that must at least be maintained for a structure to be considered safe.

[2] This is about [Eurocode 0, 2015, Table B2]:

reliability class RC1 $\beta_{50,\text{target}} = 3.3 \Rightarrow \beta_{1,\text{independence,target}} = 4.3 > \beta_{1,\text{target}} = 4.2$. Thus, there is some kind of dependency assumed; possibly due to the uncertainties in the resistance model. However, if the underlying $\beta_{50,\text{target}}$ is actually between 3.25 and 3.27, this could also be attributed to round-off errors.

reliability class RC2 $\beta_{50,\text{target}} = 3.8 \Rightarrow \beta_{1,\text{independence,target}} = \beta_{1,\text{target}} = 4.7$. This means that statistical independence is assumed between different years. Possibly because the uncertainties in the loading are assumed to dominate compared to the uncertainties about the resistance for models in this class.

reliability class RC3 $\beta_{50,\text{target}} = 4.3 \Rightarrow \beta_{1,\text{independence,target}} = 5.1 < \beta_{1,\text{target}} = 5.2$. According to theoretical considerations, $\beta_{1,\text{target}}$ should not be larger than 5.1. It is not clear why a value of 5.2 was specified. It seems unlikely that this can be attributed to round-off errors.

tainties are considered in the analysis. If the modeling uncertainty is smaller than the degree of conservativeness of the model (a conservative model has a larger probability of failure than the underlying system; i.e., a conservative model represent the state of the actual system on the safe side), modeling uncertainties could be neglected. However, careful interpretation is required if the reliability of such a model is compared with the reliability of other models.

A discussion of modeling uncertainties in the context of structural reliability can also be found in [Ditlevsen and Madsen, 2007, Chapter 3]. In general, probabilistic modeling is difficult in reliability analysis, because the behavior in the tails of the distributions employed to represent our uncertainty about $\boldsymbol{\theta}$ has a considerable influence on the reliability of the investigated problem.

4.4.8 Interpretation of the probability of failure and the reliability index

The probability of failure P_f is conditional on the assumed stochastic model class $\mathcal{M}$. The quantity P_f should not be viewed as the actual probability of failure of the system of interest. This is also pointed out in [Eurocode 0, 2015, Annex C]. During it's service live, a structure will either fail or it wont fail. The quantity P_f expresses our belief about the failure event based on the knowledge we have. Obviously, unimaginable events, also known as *Black Swan* events (see Section 4.1.5), are not considered in the analysis. Also gross human errors can usually not reasonably be quantified [Melchers, 1999]. This is why it has been noted that referring to P_f as the *nominal/notional* probability of failure instead would be more appropriate [Melchers, 1999; Kulhawy et al., 1983].

As P_f might easily get misinterpreted as an actual probability of failure, instead of as a notional quantity, the reliability index β is often a more appropriate measure to communicate the anticipated level of safety of a structure.

Chapter 5

Numerical Methods for Reliability Analysis

Reliability Analysis belongs to the category of *forward analysis* (Chapter 4) and is discussed in Section 4.4. This entire chapter is devoted to numerical methods for reliability analysis (known as *reliability methods*), because reliability methods are used as a basis to derive efficient algorithms for Bayesian inference in Chapter 7.

5.1 Introduction

5.1.1 Reliability methods – an overview

The integral in Eq. (4.1) can often not be evaluated directly, because the failure domain $\Gamma_f = \{\boldsymbol{\theta} \in \Gamma | g(\boldsymbol{\theta}) \leq 0\}$ is not known explicitly. Instead, Eq. (4.1) is usually solved numerically. The probabilities that we are dealing with in reliability analysis are typically rather small; i.e., $P_f \ll 10^{-2}$. This renders the numerical treatment of the integral in Eq. (4.1) difficult, because the failure domain Γ_f constitutes only a small part of the total domain Γ.

The class of numerical methods specifically designed to solve Eq. (4.1) are referred to as *reliability methods*. The various reliability methods differ in their treatment of the reliability integral Eq. (4.1). The most straight-forward (and simplest) method to solve Eq. (4.1) is *Monte Carlo simulation* (MCS), which is discussed in Section 5.2. However, for small failure probabilities, MCS requires a considerable number of limit-state function evaluations. Most reliability methods aim at minimizing the number of required limit-state function calls. This is because in structural reliability, the limit-state function is commonly expressed as a function that depends on the outcome of a finite element analysis. Consequently, for every limit-state function evaluation, a finite element analysis must be performed – which renders the

reliability analysis of large finite element systems computationally expensive.

Besides MCS, other well-known reliability methods are the *First Order Reliability Method* (FORM) [Hasofer and Lind, 1974; Rackwitz and Flessler, 1978], the *Second Order Reliability Method* (SORM) [Breitung, 1984], *importance sampling methods* including *line sampling* [Hohenbichler and Rackwitz, 1988; Koutsourelakis et al., 2004; Rackwitz, 2001] and *directional importance sampling* [Bjerager, 1988; Ditlevsen et al., 1990], and *Subset Simulation* (SuS) [Au and Beck, 2001].

5.1.2 Transformation to standard Normal space

For many reliability methods, it is convenient to express Eq. (4.1) in terms of a stochastic vector $\mathbf{u} \in \mathbb{R}^M$ whose coefficients are independent standard Normal variables, instead of a vector $\boldsymbol{\theta}$ of possibly dependent and arbitrary distributed stochastic variables [Ditlevsen and Madsen, 2007; Melchers, 1999]. For FORM and SORM, such a transformation is compulsory. The transformation requires a mapping $\mathrm{T}^{-1} : \mathbf{u} \rightarrow \boldsymbol{\theta}$ (for more details see Section 3.2). Let the limit-state function in terms of the independent standard Normal random variables be defined as $G(\mathbf{u}) = g\left(\mathrm{T}^{-1}(\mathbf{u})\right)$. The reliability problem in standard Normal space can then be expressed as

$$P_f = \int_{G(\mathbf{u})\leq 0} \varphi_M(\mathbf{u})\,\mathrm{d}\mathbf{u} \tag{5.1}$$

with $\varphi_M(\mathbf{u}) = \varphi(u_1) \cdot \varphi(u_2) \cdot \ldots \cdot \varphi(u_M)$, where $\varphi(\cdot)$ denotes the probability density function (PDF) of the standard Normal distribution.

For most problems of practical relevance, the transformation T^{-1} can be readily established. The two most-commonly employed transformation methods are the *Rosenblatt transformation* (Section 3.2.1) and the *Nataf transformation* (Section 3.2.2). Note: The limit-state function $G(\mathbf{u})$ in underlying standard Normal space is not explicitly required (see *Algorithm (5.1)*).

Algorithm 5.1. *Generation of samples* $\boldsymbol{\theta}$ *based on an underlying independent standard Normal sample* $\mathbf{u}$:

1. Generate sample $\mathbf{u}$ from the independent multivariate standard Normal distribution; $\mathbf{u}$ and $\boldsymbol{\theta}$ have the same dimension M.
2. Transform sample $\mathbf{u}$ to $\boldsymbol{\theta}$. Often either the *Rosenblatt transformation* (Section 3.2.1) or the *Nataf transformation* (Section 3.2.2) are employed.
3. Evaluate limit-state function $g(\boldsymbol{\theta})$.

On the one hand, when implementing a reliability method in a software code, the transforma-

tion to independent standard Normal space adds an additional layer of complexity. One the other hand, transforming the problem to the independent standard Normal space normalizes the joint PDF of the the stochastic variables of the problem. This allows us to set-up importance sampling densities or Markov chain proposal distributions that achieve an acceptable performance for a wide range of problems – independent of the variance of the stochastic variables in $\boldsymbol{\theta}$.

The inverse mapping of T^{-1} is denoted as $\mathrm{T} : \boldsymbol{\theta} \rightarrow \mathbf{u}$. However, most reliability methods require only the mapping T^{-1}. An exception is FORM, which in some cases can require both mappings, T^{-1} as well as T.

5.2 Monte Carlo simulation

5.2.1 Interpretation of the reliability problem in Monte Carlo simulation

For Monte Carlo simulation (MCS), a so-called indicator function is introduced in order to rewrite the integral in Eq. (4.1) or Eq. (5.1). The indicator function is by definition *one* if failure occurs and *zero* otherwise. The integral over the failure region can then be expressed as an integral over $\mathbb{R}^M$. For the problem formulated in independent standard Normal space, Eq. (5.1) becomes:

$$P_f = \int_{\mathbb{R}^M} I(\mathbf{u}) \cdot \varphi_M(\mathbf{u}) \, \mathrm{d}\mathbf{u} \tag{5.2}$$

where $I(\mathbf{u})$ denotes the indicator function defined as

$$I(\mathbf{u}) = \begin{cases} 1 & \text{if } G(\mathbf{u}) \leq 0 \\ 0 & \text{otherwise} \end{cases} \tag{5.3}$$

MCS approximates the integral in Eq. (5.2) as a sum over K samples $\mathbf{u}^{(k)}$, $k = 1, \ldots, K$, where $\mathbf{u}^{(k)}$ are samples of probability distribution $\varphi_M(\mathbf{u})$:

$$P_f = \mathrm{E}_{\mathbf{Y}}\left[I(\mathbf{u})\right] \approx p_{f,\mathrm{MCS}} = \frac{1}{K} \sum_{i=1}^{K} I\left(\mathbf{u}^{(k)}\right) \tag{5.4}$$

Essentially, the estimator $p_{f,\mathrm{MCS}}$ is obtained from the number of occurred failures $H = \sum_{i=1}^{K} I(\mathbf{u}^{(k)})$ divided by the total number of samples K used. The estimator $p_{f,\mathrm{MCS}}$ gives an unbiased estimate for the probability of failure.

Proof 5.1. The estimate of the probability of failure obtained with $p_{f,\mathrm{MCS}}$ is unbiased.

$$\mathrm{E}[p_{f,\mathrm{MCS}}] = \mathrm{E}\left[\frac{1}{K} \sum_{i=1}^{K} I\left(\mathbf{u}^{(k)}\right)\right]$$

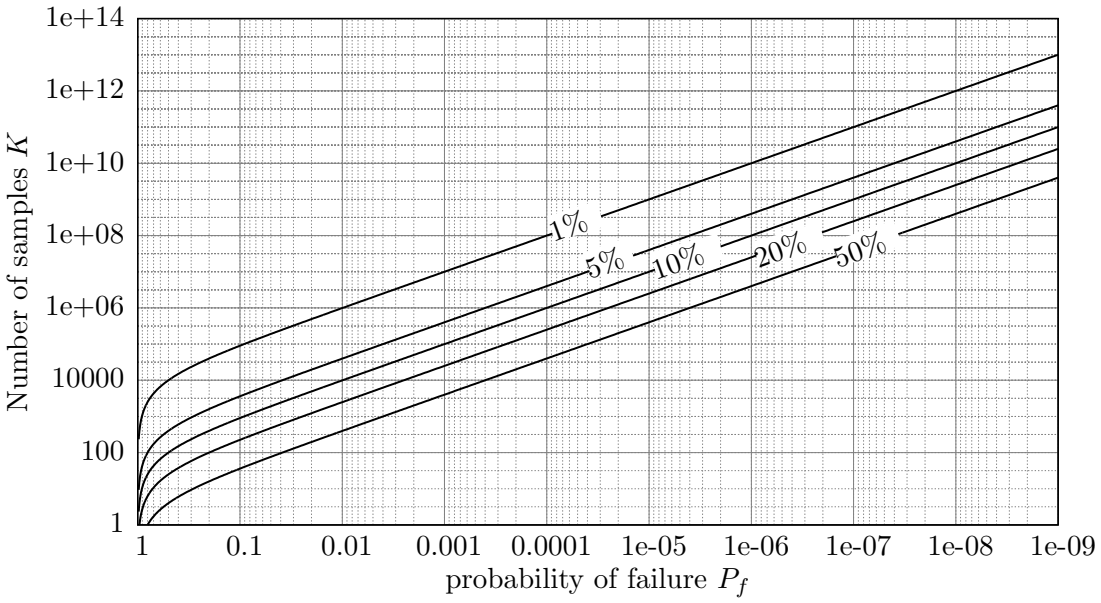


Figure 5.1: Number of samples K required to reach different coefficients of variation δ_{MCS}.

$$
\begin{aligned}
&= \frac{1}{K}\left(\sum_{i=1}^{K} \mathrm{E}\left[I\left(\mathbf{u}^{(k)}\right)\right]\right) \\
&= \frac{1}{K}\left(\sum_{i=1}^{K} P_f\right) = \frac{1}{K}\cdot K \cdot P_f \\
&= P_f
\end{aligned}
$$

□

5.2.2 Variability of the estimated probability of failure

The estimator $p_{f,\mathrm{MCS}}$ itself is a stochastic variable whose coefficient of variation depends on the (unknown) underlying P_f and the total number of samples K. Furthermore, the $I(\mathbf{u}^{(k)})$, $k = 1, \ldots, K$ can be perceived as a Bernoulli process that has a mean equal to P_f. Consequently, for a certain P_f, the number of observed failures H in K trials follows a binomial distribution. Thus, the variance of the estimator $p_{f,\mathrm{MCS}}$ is:

$$\mathrm{Var}\left[p_{f,\mathrm{MCS}}\right] = \frac{P_f(1-P_f)}{K} \tag{5.5}$$

The coefficient of variation of the estimator $p_{f,\mathrm{MCS}}$ for a fixed number of samples K can consequently be written as:

$$\delta_{\mathrm{MCS}} = \sqrt{\frac{1-P_f}{P_f K}} \tag{5.6}$$

The required number K of samples to maintain a specific target coefficient of variation δ_{MCS} for a given P_f is illustrated in Fig. 5.1. The relation in Fig. 5.1 can also be expressed explicitly: We need K to be at least as large as $(1-P_f)/(P_f \cdot \delta^2_{\mathrm{MCS}})$ to maintain a coefficient of variation of δ_{MCS}. Even though δ_{MCS} depends on the (unknown) underlying P_f, Eq. (5.6) highlights the major strength and weakness of Monte Carlo simulation: On the one hand, the weakness is that for small P_f the total number of samples K must be large to achieve a reasonable coefficient of variation of the estimate. On the other hand, the strength of MCS is that δ_{MCS} does not depend on the number of stochastic variables N, i.e., the dimension of $\boldsymbol{\theta}$ and $\mathbf{u}$. Moreover, MCS can be considered a very robust method: It is the only reliability method whose performance solely depends on the targeted probability of failure – and not on the shape of the limit-state function or on the shape of the failure domain.

Note that both Eqs. (5.5) and (5.6) assume that the actual probability of failure P_f of the problem at hand is known. Both estimators are not valid if the estimated probability of failure $p_{f,\mathrm{MCS}}$ is used instead of P_f. To assess the uncertainty in the estimated probability of failure $p_{f,\mathrm{MCS}}$, the use of a Bayesian strategy (as is explained in the next section) is strongly recommended.

5.2.3 Quantification of the uncertainty about the probability of failure

5.2.3.1 Bayesian interpretation

The number of failures H that occur in K trials follows a binomial distribution with parameter p_f (see Section 5.2.2). Thus, having observed a certain number H of failures in K trials, the likelihood of p_f can be expressed as:

$$L(p_f|H,K) = \binom{K}{H}(p_f)^H(1-p_f)^{K-H} \tag{5.7}$$

For the problem at hand, the beta distribution acts as conjugate prior for the problem. Consequently, for a beta distribution as prior, the posterior distribution is also a beta distribution. Using the parameterization H_{p} and $K_{\mathrm{p}} - H_{\mathrm{p}}$, the prior can be written as:

$$p(p_f) = \frac{p_f^{H_{\mathrm{p}}-1} \cdot (1-p_f)^{K_{\mathrm{p}}-H_{\mathrm{p}}-1}}{\mathrm{B}(H_{\mathrm{p}}, K_{\mathrm{p}}-H_{\mathrm{p}})} \tag{5.8}$$

where $\mathrm{B}(\cdot,\cdot)$ denotes the beta function defined as $\mathrm{B}(\alpha,\beta) = \int_0^1 t^{\alpha-1}\,(1-t)^{\beta-1}\,\mathrm{d}t$ with $\alpha,\beta > 0$.

Consequently, having observed a certain number H of failures in K trails, the uncertainty about the true value of p_f can be quantified by means of Bayes' theorem as ([Zuev et al.,

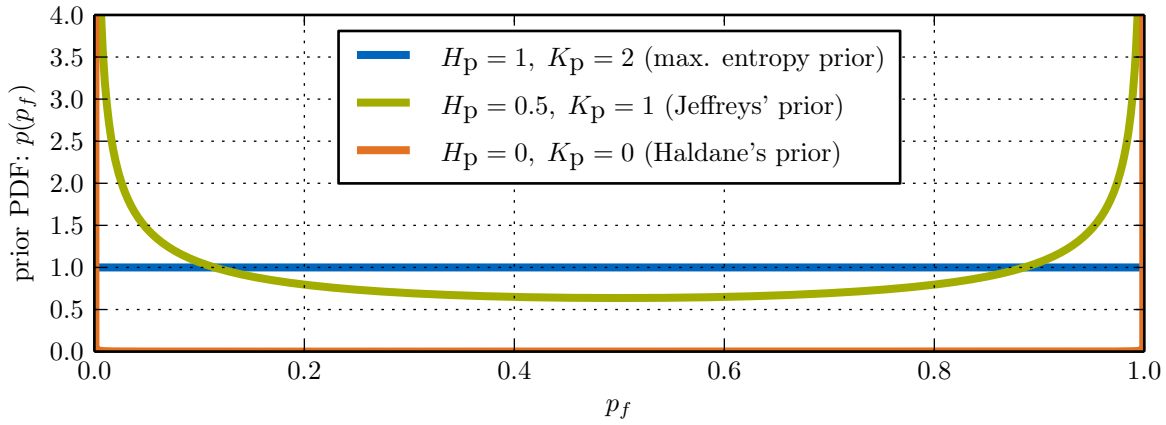


Figure 5.2: Common prior distributions for p_f.

2012; MacKay, 2003; Brown et al., 2001]):

$$p(p_f|H,K) = \frac{{p_f}^{H+H_\mathrm{p}-1} \cdot (1-p_f)^{K-H+K_\mathrm{p}-H_\mathrm{p}-1}}{\mathrm{B}(H+H_\mathrm{p}, K-H+K_\mathrm{p}-H_\mathrm{p})} \tag{5.9}$$

where $p(p_f|H,K)$ denotes the PDF of p_f being the true P_f conditioned on H and K.

5.2.3.2 Discussion of prior distributions for p_f

The parameters H_p and K_p control the shape of the prior distribution. In the following, we discuss three potential prior distributions for p_f that differ in the choice of H_p and K_p:

Maximum entropy prior If nothing is known about p_f in advance except that $p_f \in [0,1]$, then H_p and K_p could be selected as $H_\mathrm{p} = 1$, $K_\mathrm{p} = 2$ [Zuev et al., 2012] in accordance with the Principle of Maximum Information Entropy [Jaynes, 2003, 1957]. For $H_\mathrm{p} = 1$, $K_\mathrm{p} = 2$, the prior distribution is a uniform distribution on the interval $[0,1]$.

Jeffreys' prior Selecting $H_\mathrm{p} = 0.5$ and $K_\mathrm{p} = 1$ gives the so-called Jeffreys' prior [Jeffreys, 1998, 1946] for the problem at hand [Brown et al., 2001]. The Jeffreys prior is an uninformative prior that is invariant under reparameterization.

Haldane's prior Haldane [Haldane, 1932] proposed to select a beta-distribution with $H_p = K_p = 0$ as prior. This choice results in an improper prior, where the prior p_f is either *zero* or *one* with equal probability.

The shape of the prior distributions listed above is illustrated in Fig. 5.2. Fig. 5.3 plots the probability that the real probability of failure P_f is actually contained inside the estimated posterior credible intervals. Two-sided credible intervals as well as credible intervals for the

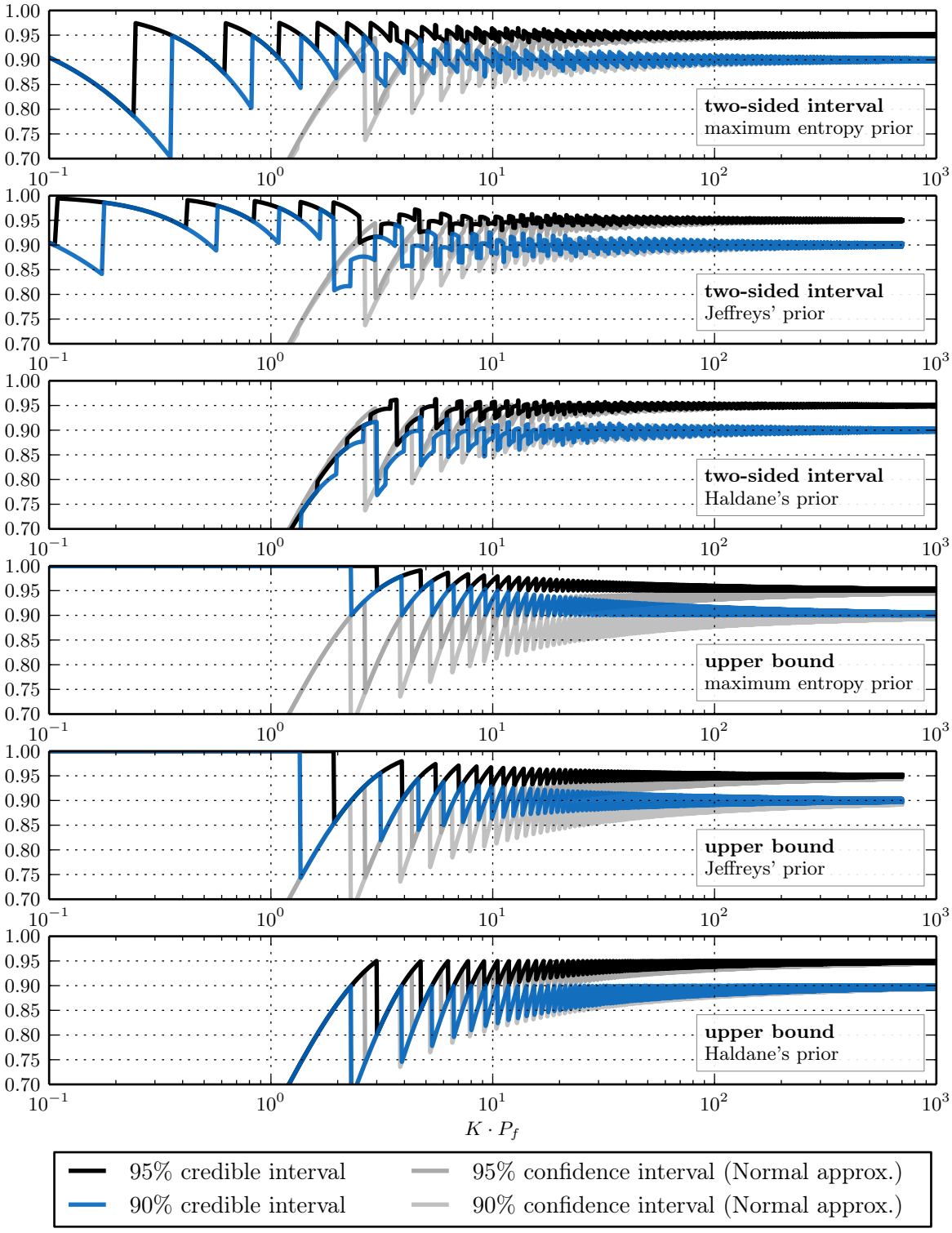


Figure 5.3: The plots show the probability that the real probability of failure P_f is actually contained inside the posterior credible intervals. The study is preformed for different $K \cdot P_f$; the plot is valid for $P_f < 1\%$. The performance of different prior choices is compared to the confidence interval obtained with a primitive standard Normal approximation. The upper three plots investigate the two-sided equal-tailed credible intervals – the probability of p_f being below is as likely as being above the interval. The lower three plots investigate the credible interval $[0, a]$, where p_f being smaller or equal than a is either 95% or 90%.

upper bound are shown. Haldane's prior is clearly not giving conservative credible intervals – especially with respect to the credible interval for the upper bound. A similar performance is obtained using a primitive standard Normal approximation; i.e., the confidence intervals of a Normal distribution centered at $p_{f,\mathrm{MCS}}$ with variance according to Eq. (5.5). Rather good credible intervals that tend to be on the conservative side are obtained with the maximum entropy prior. Note that $\mathrm{E}[p_f|H,K] = (H+H_p)/(K+K_p)$ is biased for virtually all $H_\mathrm{p}, K_\mathrm{p} \neq 0$. However, the Haldane's prior ($H_p = K_p = 0$) is usually not a good choice to express our uncertainty about p_f, as is shown in Fig. 5.3.

5.2.3.3 Maximum entropy prior

In the following, we consider only the prior based on the Principle of Maximum Information Entropy; i.e., $H_\mathrm{p} = 1$, $K_\mathrm{p} = 2$. In this case, Eq. (5.9) reduces to:

$$p(p_f|H,K) = \frac{{p_f}^H \cdot (1-p_f)^{K-H}}{\mathrm{B}(H+1, K-H+1)} \tag{5.10}$$

Thus, the posterior uncertainty about the probability of failure follows a beta distribution with parameters $H+1$ and $K-H+1$. The expectation of Eq. (5.10) is:

$$\mathrm{E}\left[p_f|H,K\right] = \frac{H+1}{K+2} \tag{5.11}$$

The variance of the distribution in Eq. (5.10) can be derived as:

$$\mathrm{Var}\left[p_f|H,K\right] = \frac{(H+1)\cdot(K-H+1)}{(K+2)^2\cdot(K+3)} \tag{5.12}$$

The coefficient of variation of the distribution in Eq. (5.10) can be derived as:

$$\delta_{p_f|H,K} = \sqrt{\frac{K-H+1}{(K+3)\cdot(H+1)}} \tag{5.13}$$

Note that if applied to failure probabilities (i.e., $P_f \ll 1$), then $\mathrm{E}[p_f|H,K] > \mathrm{E}[p_{f,\mathrm{MCS}}] = P_f$, and $\mathrm{Var}[p_f|H,K] > \mathrm{Var}[p_{f,\mathrm{MCS}}]$ on average – which can be considered as conservative. However, the reverse is true for the coefficient of variation, i.e., $\mathrm{C.o.V.}[p_f|H,K] < \mathrm{C.o.V.}[p_{f,\mathrm{MCS}}]$ on average. For this reason, the interpretation of the coefficient of variation of distribution Eq. (5.10) should be handled with care.

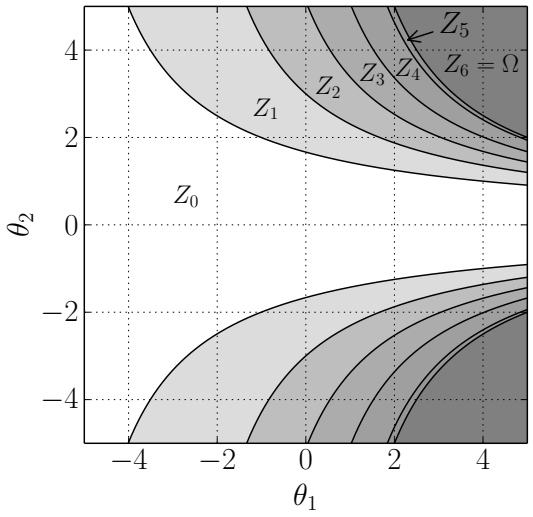


Figure 5.4: The basic idea behind Subset Simulation is illustrated using limit-state function g_4 and $\alpha_4 = 10$. The initial domain Z_0 covers the entire parameter space of θ_1 and θ_2. The intermediate domains Z_1 to Z_5 are chosen such that $Pr(Z_i|Z_{i-1}) = 10\%$, $i = 1, \ldots, 5$; i.e., $h_1 = 6.01$, $h_2 = 3.33$, $h_3 = 1.95$, $h_4 = 0.97$ and $h_5 = 0.16$. Domain Z_6 is equivalent to the failure domain Γ_f, i.e. $h_6 = 0$, and $Pr(Z_6|Z_5) = 0.542$. Consequently, the probability of failure can be evaluated as $P_f = 0.1^5 \cdot 0.542 = 5.42 \cdot 10^{-6}$.

5.3 Subset Simulation

5.3.1 Overview

Subset Simulation (SuS) was proposed by Au and Beck in [Au and Beck, 2001] and is an adaptive Monte Carlo method that is efficient for estimating small probabilities in high dimensional problems. The method represents the probability of failure as a product of larger conditional probabilities. This is done by expressing the failure domain $\Gamma_f = \{\mathbf{u} \in \mathbb{R}^M | G(\mathbf{u}) \leq 0\}$ as the intersection of N intermediate nested domains Z_i, where $Z_0 = \mathbb{R}^M \supset Z_1 \supset \cdots \supset Z_N = \Gamma_f$. The domains Z_i are defined as the sets $\{G(\mathbf{u}) \leq h_i\}$, where h_i are positive coefficients such that $h_0 = \infty > h_1 > h_2 > \cdots > h_N = 0$ holds. The probability of failure can then be written as:

$$P_f = \Pr(\Gamma_f) = \prod_{i=1}^{N} \Pr(Z_i|Z_{i-1}) \tag{5.14}$$

Note that the joint PDF of samples in Z_i is:

$$\varphi_N(\mathbf{u}|Z_i) = \begin{cases} \varphi_N(\mathbf{u})/\Pr(Z_i) & \text{if } \mathbf{u} \in Z_i \\ 0 & \text{otherwise} \end{cases} \tag{5.15}$$

where $\Pr(Z_i) = \prod_{j=1}^{i} \Pr(Z_j|Z_{j-1})$. The idea behind Subset Simulation is illustrated in Fig. 5.4.

The first conditional probability $\Pr(Z_1|Z_0)$ can directly be estimated by means of Monte Carlo simulation, because $\varphi_M(\mathbf{u}|Z_0) = \varphi_M(\mathbf{u})$. Let K_1 be the number of samples used, and H_1 be the number of samples that were observed to be in domain Z_1. An unbiased estimator for the probability $\Pr(Z_1|Z_0)$ is given by the ratio $p_1 = H_1/K_1$ (see Section 5.2).

For all other conditional probabilities $\Pr(Z_i|Z_{i-1}),\ i > 1$, samples from $\varphi_M(\mathbf{u}|Z_i)$ cannot be generated directly. In Subset Simulation, Markov chain Monte Carlo (MCMC) methods [Liu, 2001; Neal, 1993; Robert and Casella, 2004] are applied to generate samples in Z_{i-1}. For each sample that was found to be in Z_{i-1} when estimating $\Pr(Z_{i-1}|Z_{i-2})$, a Markov chain is started using that sample as seed. In order to generate K_i samples of Z_{i-1}, the length of the individual chains needs to be K_i/H_{i-1}. Note that in this case the Markov chains do not suffer from a burn-in, since the seeds of the chains already follow the target distribution [Au and Beck, 2001]. An estimator for the probability $\Pr(Z_i|Z_{i-1})$ is given by the ratio $p_i = H_i/K_i$, where H_i is the number of samples observed to be part of Z_i. An estimator for the probability of failure using Subset Simulation is:

$$P_f \approx p_{f,\mathrm{SuS}} = \prod_{i=1}^{N} p_i \tag{5.16}$$

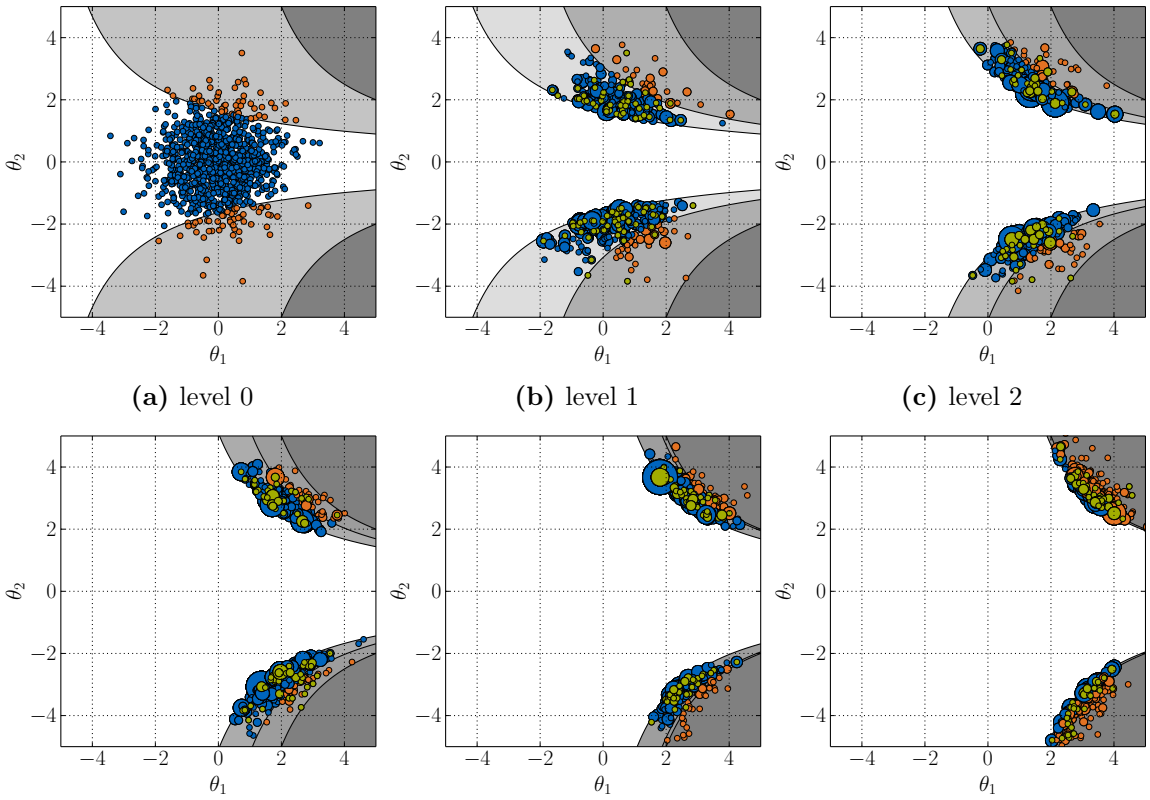
The intermediate threshold values h_i should ideally be selected such that the conditional probabilities are approximately the same. However, for typical limit-state functions neither the probability of failure of the problem nor the shape of the limit-state function are known. Consequently, the h_i cannot efficiently be selected in advance. Instead, the h_i are typically selected adaptively during the simulation such that the conditional probabilities match a predefined probability, denoted p_t [Au and Beck, 2001].

An exemplary run of Subset Simulation is illustrated in Fig. 5.5.

5.3.2 MCMC algorithms for Subset Simulation

For SuS working in independent standard Normal space (see Section 5.1.2), the distribution $\varphi_M(\mathbf{u}|Z_i)$ defined in Eq. (5.15) is sampled by means of MCMC simulation (see Sections 3.4 and 3.5), for $i \geq 1$. *Algorithm (3.3)* provides the general framework for SuS-based MCMC sampling in standard Normal space. Different MCMC algorithms proposed for Subset Simulation are discussed in [Papaioannou et al., 2015]. In this contribution, we will focus on the CS (*conditional sampling in standard Normal space*) algorithm (*Algorithm (3.5.6)*) proposed by [Papaioannou et al., 2015] (see Section 3.5.6). Some alternative MCMC algorithms for Subset Simulation are presented in Section 3.5. The most commonly used MCMC algorithm for SuS is the cwMH algorithm proposed in [Au and Beck, 2001] (see Section 3.5.4).

To improve the efficiency of MCMC in SuS, the spread of the proposal distribution can be



(a) level 0 **(b)** level 1 **(c)** level 2

(d) level 3 **(e)** level 4 **(f)** level 5

Figure 5.5: Subset Simulation is performed for limit-state function g_4 with 10^3 samples per level. The parameter α_4 of limit-state function g_4 is set to 10, which gives a target probability of failure of $5.42 \cdot 10^{-6}$.
Samples at the intermediate levels i (i.e., samples in domain Z_i) of Subset Simulation are shown. The blue samples are located in domain $Z_i \setminus Z_{i+1}$. The orange points indicate samples from Z_i that are also in domain Z_{i+1}. The samples highlighted green are the seeds used to initiate the Markov chains to generate samples in domain Z_i. Note that the seeds (i.e., the green samples) correspond to the orange samples from the previous level of Subset Simulation. The area of the illustrated samples is an indicator for how often the corresponding sample appears in the set of samples at level i; repeated samples are due to the accept/reject step in MCMC.
At the ith level of Subset Simulation, $i \in \{0, \ldots, N\}$, the coefficients h_{i+1} that define domain Z_{i+1} are selected such that 10% of the samples in domain Z_i fall also in domain Z_{i+1}. Note that h_{i+1} depends on the samples at level i, and, thus, the coefficients h_i differ in each run of Subset Simulation. Consequently, the intermediate domains Z_i depicted in sub-figures (a)-(f) do not correspond exactly to the intermediate domains shown in Fig. 5.4.
The iteration is stopped if the number of samples in Z_i that also fall in the actual failure domain Γ_f exceeds 10% of the total number of samples generated at each level of Subset Simulation. In the example at hand, the final level is shown in sub-plot (f), it is level 5. In the final level of Subset Simulation, 636 samples fall into the actual failure domain Γ_f. Thus, an estimate for the probability of failure can be evaluated as: $p_{f,\mathrm{SuS}} = 0.1^5 \cdot \frac{636}{10^3} = 6.4 \cdot 10^{-6}$.

improved adaptively [Papaioannou et al., 2015]. Adaptive MCMC strategies are discussed in Section 3.5.9. In this contribution, *Algorithm (3.9)* is employed to adaptively modify the MCMC proposal spread.

5.3.3 Implementation of Subset Simulation

In the following the algorithm for Subset Simulation with adaptive learning of the spread of the proposal distribution is described:

Algorithm 5.2. *Subset Simulation*:
Define values for p_t, and K_i (e.g.: $p_\mathrm{t} = 10\%$, $K_i = 1000$). The K_i should be selected such that $p_\mathrm{t} \cdot K_i$ is an integer value. Typically, the same value is assigned to all K_i. Moreover, set $i = 1$.

The algorithm evaluates an estimate for the probability of failure P_f of the reliability problem at hand.

1. Perform the initial Monte Carlo sampling:
   (a) Draw K_1 samples $\mathbf{u}_j^{(1)}$, $j = 1, \ldots, K_1$, from distribution $\varphi_M(\mathbf{u})$.
   (b) Transform all $\mathbf{u}_j^{(1)}$ to $\boldsymbol{\theta}_j^{(1)}$: $\boldsymbol{\theta}_j^{(1)} = \mathrm{T}^{-1}(\mathbf{u}_j^{(1)})$
   (c) For each sample $\boldsymbol{\theta}_j^{(1)}$ evaluate $g_j^{(1)} = g\left(\boldsymbol{\theta}_j^{(1)}\right)$.
2. Find the threshold h_i:
   (a) Sort the K_i values of $g_j^{(i)}$ and the corresponding $\mathbf{u}_j^{(i)}$ in an increasing order with respect to the value of $g_j^{(i)}$.
   (b) Set integer $m = p_\mathrm{t} \cdot K_i$.
   (c) If $g_m^{(i)} = g_{m+1}^{(i)}$ then increase m until $g_m^{(i)} \neq g_{m+1}^{(i)}$.
   (d) If $g_m^{(i)} \leq 0$ then increase m until $g_{m+1}^{(i)} > 0$.
   (e) If $g_m^{(i)} > 0$, then set the threshold $h_i = \frac{1}{2}\left(g_m^{(i)} + g_{m+1}^{(i)}\right)$. Otherwise, set $h_i = 0$.
   (f) Compute $p_i = H_i/K_i$, where $H_i = m$.
3. If $h_i = 0$, then go to *(9.)*.
4. For $j = 1, \ldots, H_i$: Shuffle the $\mathbf{u}_j^{(i)}$ and the corresponding $g_j^{(i)}$ randomly. This step is required because an ordered sequence of $\mathbf{u}_j^{(i)}$ and $g_j^{(i)}$ will increase the bias of the SuS estimate if the spread of the proposal distribution is learned adaptively (e.g., by means of *Algorithm (3.9)*).
5. Increase i by one: $i = i + 1$.
6. Determine the length of the individual chains for MCMC sampling:
   (a) Set m as the first integer smaller or equal than K_i/H_{i-1}.
   (b) Set $l_k^{(i)} = m$ with $k = 1, \ldots, H_{i-1}$.

(c) For all $l_k^{(i)}$ with $k \leq K_i - m \cdot H_{i-1}$ do: $l_k^{(i)} = l_k^{(i)} + 1$. This ensures that $\sum_{k=1}^{H_{i-1}} l_k^{(i)} = K_i$.

7. Draw samples from $\varphi_M(\mathbf{u}|Z_{i-1})$ by means of MCMC sampling:
   (a) Set $j = 1$ and $k = 1$.
   (b) If $l_k^{(i)} = 0$ go to *(i)*.
   (c) Set $\mathbf{u}_j^{(i)} = \mathbf{u}_k^{(i-1)}$, $g_j^{(i)} = g_k^{(i-1)}$, and $m = 1$.
   (d) Perform *Algorithm (3.9)* to learn the spread of the proposal distribution adaptively.
   (e) Increase both m and j: $m = m + 1$ and $j = j + 1$.
   (f) If $m > l_k^{(i)}$ go to *(i)*.
   (g) Draw sample $\mathbf{u}_j^{(i)}$ based on sample $\mathbf{u}_{j-1}^{(i)}$ using MCMC by means of *Algorithm (3.3)*.
   (h) Go back to *(e)*.
   (i) Increase k by one: $k = k + 1$.
   (j) If $k \leq H_{i-1}$ go back to *(b)*.
8. Go back to *(2.)*.
9. Evaluate $p_{f,\mathrm{SuS}} = p_1 \cdot \ldots \cdot p_i$.

For large K_i the efficiency of the described algorithm can be improved if samples $\mathbf{u}_j^{(i)}$ are not re-ordered and shuffled directly, but the re-ordering and shuffling is done on an index-vector that points to the corresponding $\mathbf{u}_j^{(i)}$.

5.3.4 Assessing the uncertainty from sampling in SuS

5.3.4.1 Coefficient of variation of the p_i

The samples employed to estimate $\Pr(Z_i|Z_{i-1})$ for $i > 1$ are dependent since they are generated by means of MCMC. Au and Beck [Au and Beck, 2001] proposed to regard the coefficient of variation δ_i of the estimator p_i as equal to the coefficient of variation of a MCS with a reduced number of samples, denoted $K_{i,\mathrm{eff}}$:

$$K_{i,\mathrm{eff}} = \frac{1}{1+\gamma_i} K_i \tag{5.17}$$

where $\gamma_i \geq 0$ is a factor that accounts for the dependency of the samples used to estimate $\Pr(Z_i|Z_{i-1})$. The larger γ_i, the larger is the dependency between the samples. For $\gamma_i = 0$, the samples are independent. Based on the effective number of samples introduced in Eq. (5.17), the coefficient of variation δ_i of the estimator p_i can be approximated as [Au and Beck, 2001]:

$$\delta_i \approx \sqrt{\frac{1-p_i}{p_i K_i} (1+\gamma_i)} \tag{5.18}$$

Note that both Eqs. (5.17) and (5.18) are valid only if the individual chains are considered independent. Thus, only the correlation of the samples within the individual chains (referred to as *chain correlation*) can be taken into account.

In practice, a conservative and reliable estimation of δ_i is difficult, because chain correlation as well as the influence of correlated seeds increase the sampling uncertainty about the value of p_i. An estimate that neglects the influence of correlated seeds is proposed in [Au and Beck, 2001]. Note that seed correlation is not present when estimating $\Pr(Z_2|Z_1)$, because the seeds come from a perfect Monte Carlo simulation. Thus, only chain correlation needs to be considered in δ_2. However, for $i > 2$, the influence of correlated seeds increases the coefficient of variation δ_i.

Zuev et al. [Zuev et al., 2012] proposed an approximation for the PDF of the estimator p_i: They assumed the MCMC samples to be independent and used the distribution specified in Eq. (5.10). In this case, the PDF of p_i can be approximated by a beta distribution. However, this approach under-represents the *true* uncertainty about p_i, as both seed and chain correlation increase the coefficient of variation δ_i of the estimate.

5.3.4.2 Assessing the uncertainty in the estimator $p_{f,\mathrm{SuS}}$

The coefficient of variation of the estimator $p_{f,\mathrm{SuS}}$ from Eq. (5.16) cannot be computed in a straightforward manner, because the $\hat{p}_i$ are dependent. The dependency is due to employing samples from $\varphi_N(\mathbf{u}|Z_{i-1})$ that fall into domain Z_i as seeds to trigger the Markov chains that are used to draw samples from $\varphi_N(\mathbf{u}|Z_i)$. An upper and lower bound for the coefficient of variation of the estimator $p_{f,\mathrm{SuS}}$ was derived in [Au and Beck, 2001]. The lower bound assumes that the p_i are uncorrelated. It can be computed as:

$$\delta_{\mathrm{SuS,low}} = \sqrt{\sum_{i=1}^{M} \delta_i^2} \tag{5.19}$$

The upper bound can be derived by assuming that the p_i are fully correlated:

$$\delta_{\mathrm{SuS,up}} = \sqrt{\sum_{i=1}^{M} \sum_{j=1}^{M} \delta_i \delta_j} \tag{5.20}$$

Note that Eq. (5.20) tends to underestimate the value of the actual upper bound, if the influence of correlated seeds is neglected in the estimate of the δ_i. This effect is illustrated in Fig. 5.6 (see Example 5.1).

Furthermore, as the uncertainty about $p_{f,\mathrm{SuS}}$ cannot be quantified using a standard distribution model (contrary to MCS, the $p_{f,\mathrm{SuS}}$ is not beta-distributed). Moreover, the distribution

of $p_{f,\mathrm{SuS}}$ can be considerably skewed as will be demonstrated by means of numerical examples in the following. Consequently, the coefficient of variation should not be used to derive credible intervals (based on a Normal approximation).

Example 5.1. *Average and estimated coefficient of variation in SuS*:
This example demonstrates that the estimate for the upper bound of the coefficient of variation (C.o.V.) defined in Eq. (5.20) is not conservative if the influence of correlated seeds is neglected in the δ_i.

The average C.o.V. of the SuS estimate and the estimated C.o.V. obtained with SuS by means of Eqs. (5.19) and (5.20) as a function of P_f are shown for different limit-state functions (g_1 to g_5 defined in Section 3.5.3), different number of samples K (10^3 and 10^4) and different target acceptance rates t_{acr} (0.3 and 0.44) in Fig. 5.6. The upper bound of the C.o.V. is approximated, neglecting the influence of correlated seeds. Limit-state functions g_1, g_2, g_3 and g_5 are analyzed for $M = 10$ and limit-state function g_4 is analyzed for $M = 2$. For limit-state function g_2, the parameter κ is set to 10. For limit-state function g_5, the parameter m is set to 4. The results shown in Fig. 5.6 are obtained by at least $5 \cdot 10^5$ ($5 \cdot 10^4$ for $K = 10^4$) repeated runs of SuS.

Contrary to Monte Carlo simulation, for fixed P_f, the C.o.V. of the estimated probability of failure $p_{f,\mathrm{SuS}}$ in SuS depends on the underlying limit-state function. This is due to the MCMC sampling employed in the intermediate levels of SuS: The samples generated with MCMC are dependent; the degree of dependency of the generated samples depends on the underlying limit-state function.

The estimate for the upper bound of the C.o.V. is clearly not conservative for small P_f when the influence of correlated seeds is neglected. For $K = 10^3$ and g_1 to g_4, the C.o.V. for small P_f is smaller for $t_{\mathrm{acr}} = 0.3$ than for $t_{\mathrm{acr}} = 0.44$.

Example 5.2. *Quantiles of the SuS estimate $p_{f,SuS}$*:
This example investigates the bias in the mean and median of the estimated $p_{f,\mathrm{SuS}}$, as well as the 1%, 5%, 95% and 99% quantiles of the estimated $p_{f,\mathrm{SuS}}$. The mentioned quantities are plotted in Fig. 5.7 for different limit-state functions, different number of samples K (10^3 and 10^4) and different target acceptance rates t_{acr} (0.3 and 0.44). Limit-state functions g_1, g_2, g_3 and g_5 (defined in Section 3.5.3) are analyzed for $M = 10$ and limit-state function g_4 is analyzed for $M = 2$. For limit-state function g_2, the parameter κ is set to 10. For limit-state function g_5, the parameter m is set to 4. The results shown in Fig. 5.7 are obtained by at least $5 \cdot 10^5$ ($5 \cdot 10^4$ for $K = 10^4$) repeated runs of SuS.

The bias in the estimated $p_{f,\mathrm{SuS}}$ is negligible compared to the spread of $p_{f,\mathrm{SuS}}$. Looking at the quantiles and the median in relation to the mean, the distribution of $p_{f,\mathrm{SuS}}$ is increasingly right-skewed (*positive* skewness) for decreasing P_f. Especially for limit-state functions g_2 and g_5, the distribution of $p_{f,\mathrm{SuS}}$ is highly skewed. The median indicates that with decreasing P_f it becomes more and more likely to observe a $p_{f,\mathrm{SuS}} < P_f$ than $p_{f,\mathrm{SuS}} > P_f$. The uncertainty about $p_{f,\mathrm{SuS}}$ depends on the formulation of the limit-state function. The shape of the final and

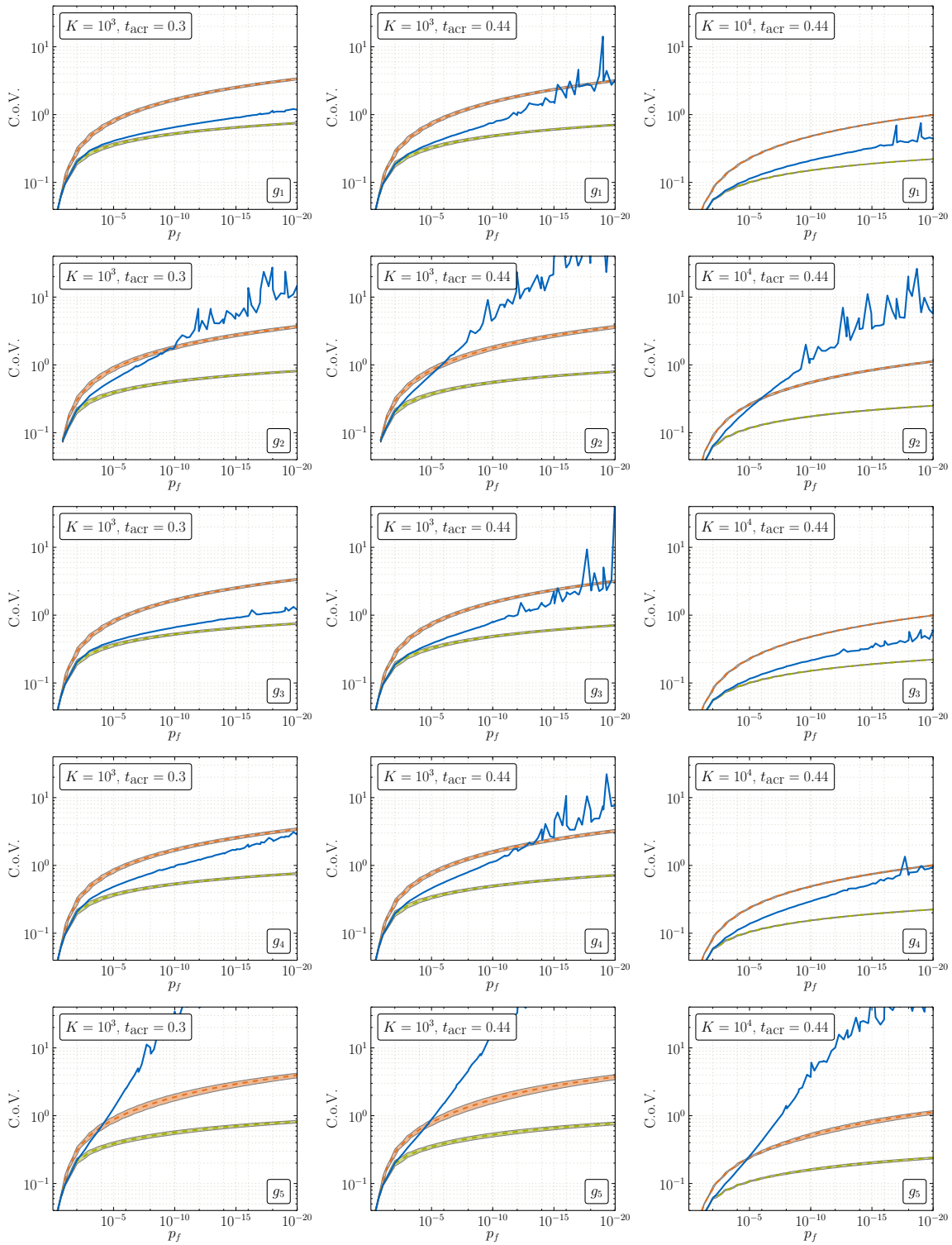


Figure 5.6: Average coefficient of variation (C.o.V.) of SuS (plotted in blue) for limit-state functions g_1 to g_5 and different number of samples K and target acceptance rates t_{acr}. The mean of the SuS estimate for the lower and upper bound of the C.o.V. computed with Eqs. (5.19) and (5.20) is represented by the dashed green and orange line, respectively. The highlighted areas represent the 90% confidence intervals. (Example 5.1)

the intermediate failure domains has a strong influence on the uncertainty about $p_{f,\mathrm{SuS}}$ (this is further investigate in Example 5.5). This is due to the MCMC sampling employed to generate samples conditional on the selected intermediate failure domain.

Example 5.3. *Q-Q plots of the SuS estimate $p_{f,SuS}$*:
This example compares the quantiles of the SuS estimate $p_{f,\mathrm{SuS}}$ with respect to a *log-Normal* and *beta* approximation by means of a Q-Q plot. The log-Normal and beta distribution are selected such that the mean is equal to $\mathrm{E}[p_{f,\mathrm{SuS}}/P_f]$ and the variance is $\mathrm{Var}[p_{f,\mathrm{SuS}}/P_f]$; i.e., the distribution used to approximate the data has the same mean and standard deviation as the data.

Fig. 5.8 shows Q-Q plots of different limit-state functions and different P_f for $K = 10^3$ and $t_{\mathrm{acr}} = 0.33$. Fig. 5.9 shows Q-Q plots of different limit-state functions and different P_f for $K = 10^3$ and $t_{\mathrm{acr}} = 0.44$. Fig. 5.10 shows Q-Q plots of different limit-state functions and different P_f for $K = 10^4$ and $t_{\mathrm{acr}} = 0.44$. Limit-state functions g_1, g_2, g_3 and g_5 (defined in Section 3.5.3) are analyzed for $M = 10$ and limit-state function g_4 is analyzed for $M = 2$. For limit-state function g_2, the parameter κ is set to 10. For limit-state function g_5, the parameter m is set to 4. The results shown in Fig. 5.8 and Fig. 5.9 are obtained by at least $5 \cdot 10^5$ repeated runs of SuS; the results in Fig. 5.10 are obtained by at least $5 \cdot 10^4$ repeated runs of SuS.

Using a log-Normal distribution to approximate the statistics of $p_{f,\mathrm{SuS}}$ is clearly better than employing a beta distribution. This is to be expected, as $p_{f,\mathrm{SuS}}$ is the product of intermediate probabilities p_i. For a large number of SuS levels (and a weak dependence of the p_i), the distribution of $p_{f,\mathrm{SuS}}$ should converge to a log-Normal distribution by means of the *central limit theorem*. However, the log-Normal approximation is not perfect, as the number of levels in SuS is typically not very large and (especially for higher levels) the intermediate p_i are dependent and not identically distributed. The quality of the log-Normal approximation depends on the type of the limit-state function: The log-Normal approximation is good for g_1 and g_3. For g_2 and g_5, the quality of the log-Normal approximation decreases with decreasing P_f. Note that the quality of the log-Normal approximation decreases with an increasing skewness of the underlying distribution in log-scale (compare Example 5.2), as a log-Normal distribution is symmetric in log-scale.

A maximum likelihood based fit might give better results than fitting for the mean and standard deviation. However, this investigation is left for future studies.

Example 5.4. *Sampling strategies in Subset Simulation*:
This example compares two sampling strategies: (1) Subset Simulation is performed successively 10 times with $K = 10^3$; the probability of failure is estimated as the average of the 10 runs. (2) Subset Simulation is performed once with $K = 10^4$. The two sampling strategies have similar computational costs. The target acceptance rate is set to $t_{\mathrm{acr}} = 0.44$.

The bias in the mean and median of the estimated $p_{f,\mathrm{SuS}}$, as well as the 1%, 5%, 95% and 99% quantiles of the estimated $p_{f,\mathrm{SuS}}$ are investigated for the two sampling strategies. The

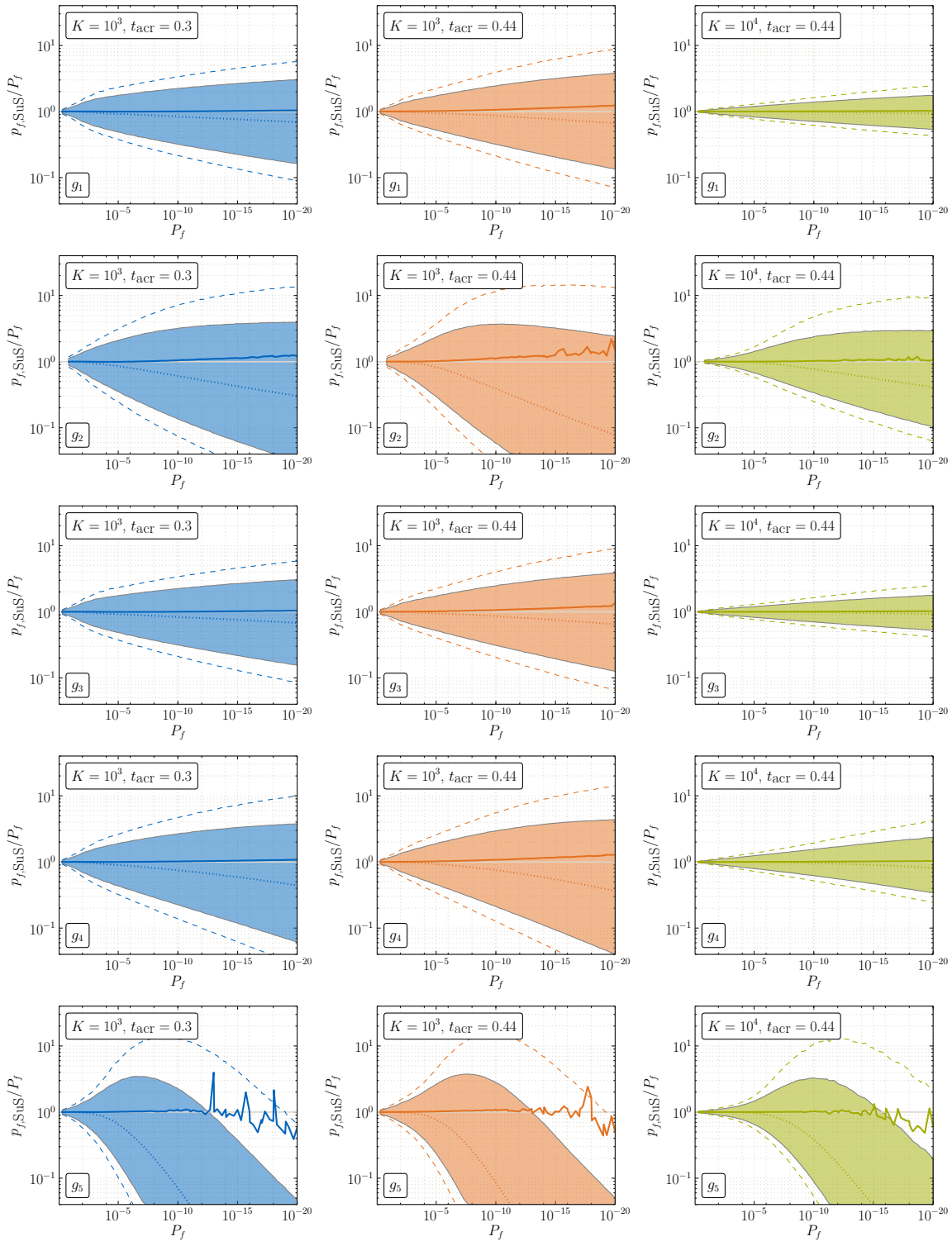


Figure 5.7: The mean of $p_{f,\text{SuS}}/P_f$ is represented by the continuous line. The bias in the SuS estimate $\text{E}[p_{f,\text{SuS}}]$ is *zero*, if $\text{E}[p_{f,\text{SuS}}]/P_f = 1$. The highlighted area shows the 90% confidence interval of $p_{f,\text{SuS}}/P_f$, defined in terms of the 5% and 95% quantiles. The dashed lines represent the 1% and 99% quantiles. The dotted line shows the median of $p_{f,\text{SuS}}/P_f$.
The study is performed for limit-state functions g_1 to g_5 and different number of samples K and target acceptance rates t_{acr}. The results are obtained by at least $5 \cdot 10^5$ ($5 \cdot 10^4$ for $K = 10^4$) repeated runs of SuS. (Example 5.2)

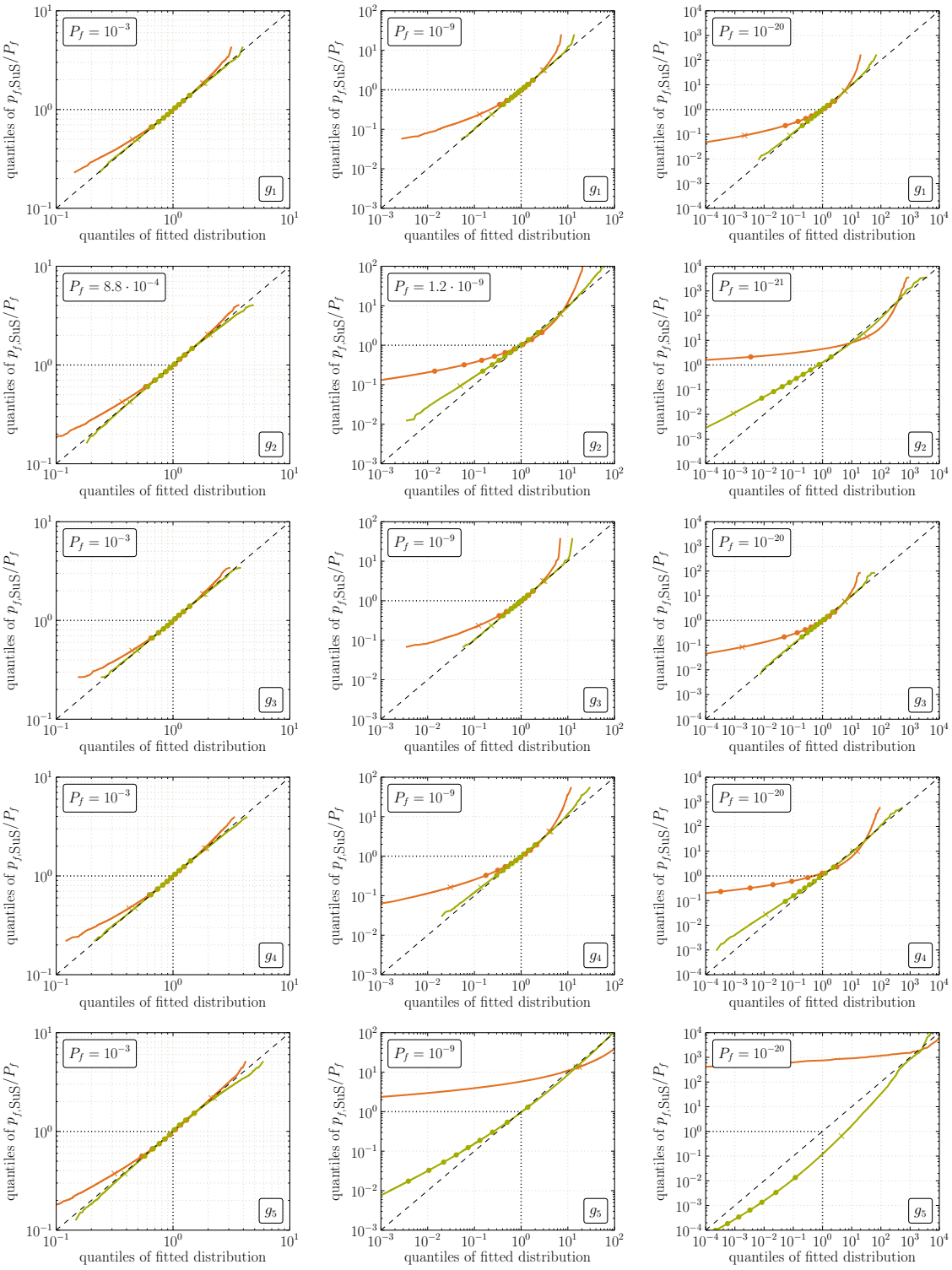


Figure 5.8: Q-Q plots of $p_{f,\mathrm{SuS}}/P_f$ for limit-state functions g_1 to g_5 and different P_f. The number of samples per level and the target acceptance rate of SuS is fixed to $K = 10^3$ and $t_{\mathrm{acr}} = 0.3$, respectively. The orange line compares the fit with respect to a *beta* distribution, and the green line compares the fit with respect to a *log-Normal* distribution. Both distributions are fitted such that the mean and variance is equal to $\mathrm{E}[p_{f,\mathrm{SuS}}/P_f]$ and $\mathrm{Var}[p_{f,\mathrm{SuS}}/P_f]$. The large points in the plot indicate the deciles (i.e., the 10%, 20%, ..., 90% quantiles). The crosses indicates the 1% and 99% quantiles. (Example 5.3)

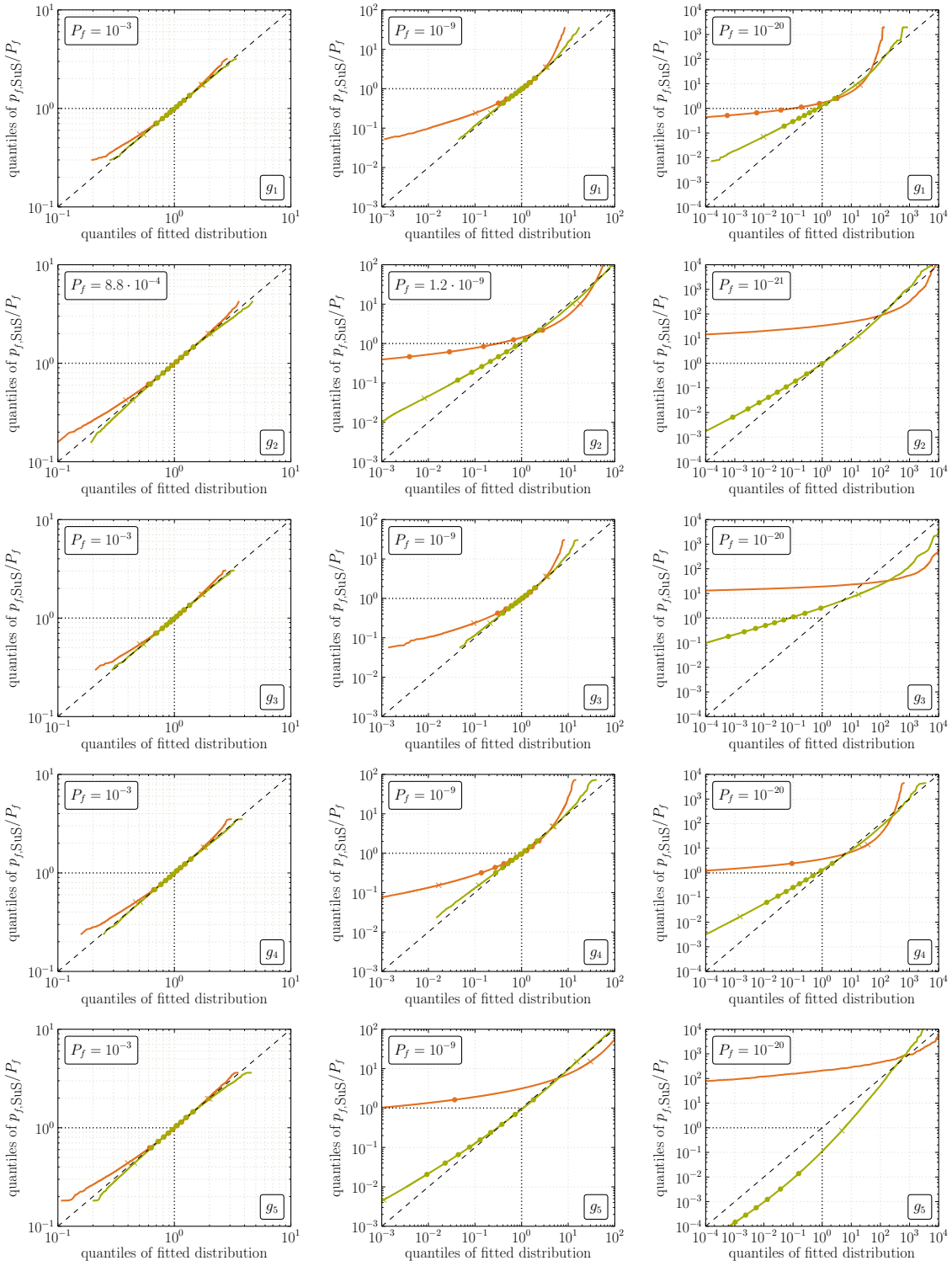


Figure 5.9: Q-Q plots of $p_{f,\mathrm{SuS}}/P_f$ for limit-state functions g_1 to g_5 and different P_f. The number of samples per level and the target acceptance rate of SuS is fixed to $K = 10^3$ and $t_{\mathrm{acr}} = 0.44$, respectively. The orange line compares the fit with respect to a *beta* distribution, and the green line compares the fit with respect to a *log-Normal* distribution. Both distributions are fitted such that the mean and variance is equal to $\mathrm{E}[p_{f,\mathrm{SuS}}/P_f]$ and $\mathrm{Var}[p_{f,\mathrm{SuS}}/P_f]$. The large points in the plot indicate the deciles (i.e., the 10%, 20%, ..., 90% quantiles). The crosses indicates the 1% and 99% quantiles. (Example 5.3)

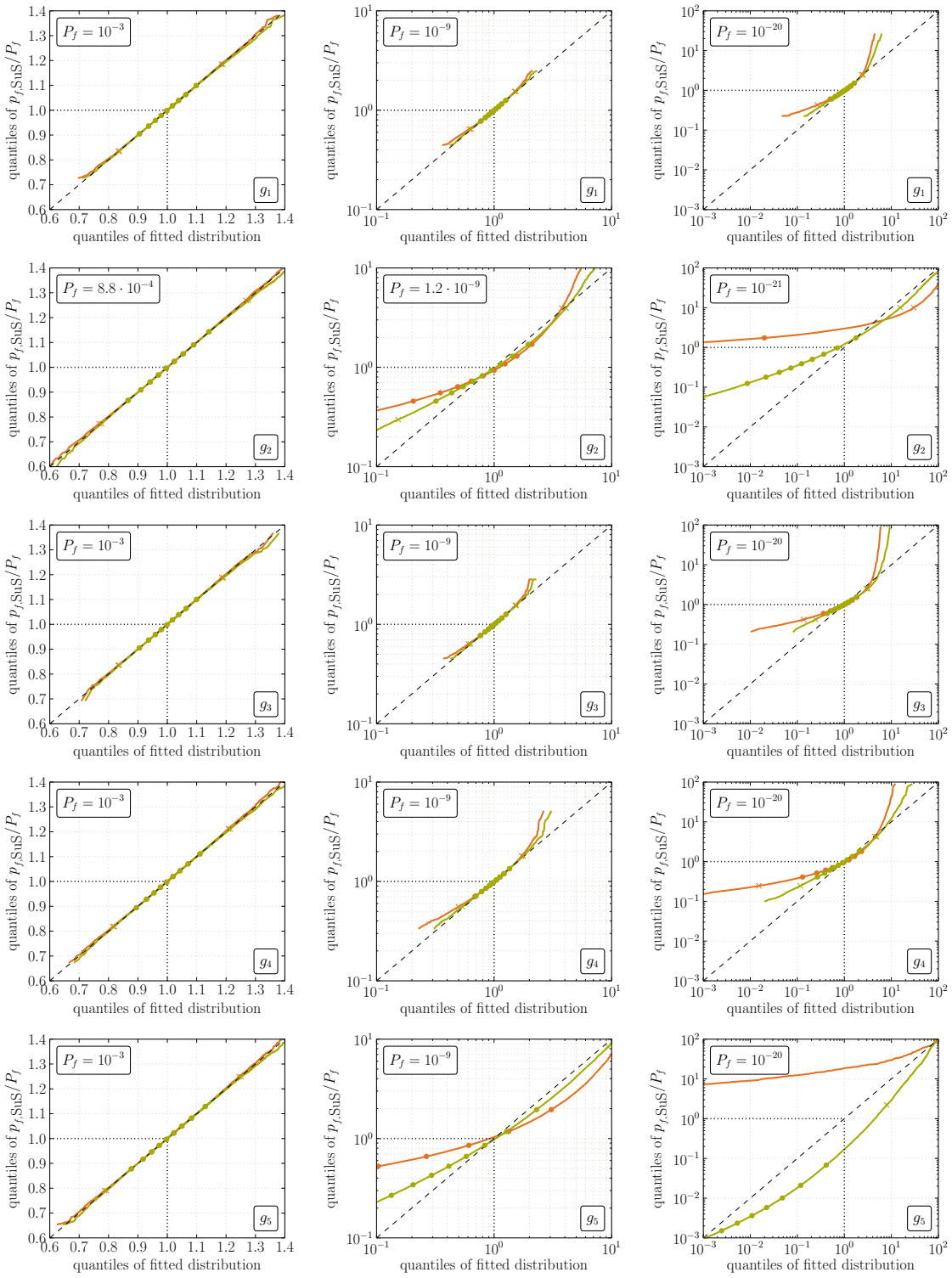


Figure 5.10: Q-Q plots of $p_{f,\mathrm{SuS}}/P_f$ for limit-state functions g_1 to g_5 and different P_f. The number of samples per level and the target acceptance rate of SuS is fixed to $K = 10^4$ and $t_{\mathrm{acr}} = 0.44$, respectively. The orange line compares the fit with respect to a *beta* distribution, and the green line compares the fit with respect to a *log-Normal* distribution. Both distributions are fitted such that the mean and variance is equal to $\mathrm{E}[p_{f,\mathrm{SuS}}/P_f]$ and $\mathrm{Var}[p_{f,\mathrm{SuS}}/P_f]$. The large points in the plot indicate the deciles (i.e., the 10%, 20%, ..., 90% quantiles). The crosses indicates the 1% and 99% quantiles. (Example 5.3)

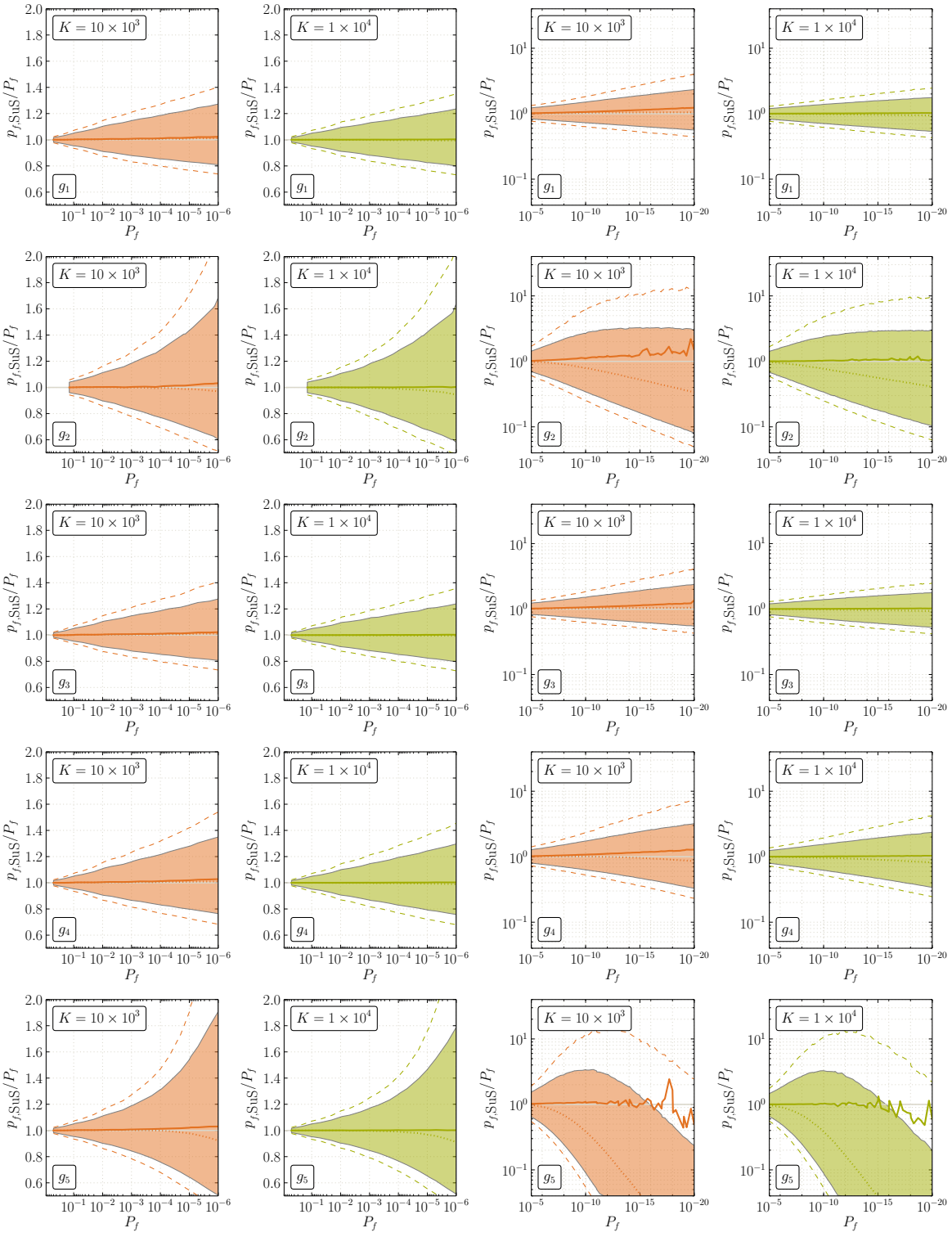


Figure 5.11: The mean of $p_{f,\mathrm{SuS}}/P_f$ is represented by the continuous line. The bias in the SuS estimate $\mathrm{E}[p_{f,\mathrm{SuS}}]$ is *zero*, if $\mathrm{E}[p_{f,\mathrm{SuS}}]/P_f = 1$. The highlighted area shows the 90% confidence interval of $p_{f,\mathrm{SuS}}/P_f$, defined in terms of the 5% and 95% quantiles. The dashed lines represent the 1% and 99% quantiles. The dotted line shows the median of $p_{f,\mathrm{SuS}}/P_f$.
Two sampling strategies are investigated: (1) 10 SuS runs with $K = 10^3$; the estimated $p_{f,\mathrm{SuS}}$ is taken as the average of the 10 runs. (2) A single SuS run with $K = 10^4$ is performed. The study is performed for limit-state functions g_1 to g_5. The target acceptance rate is set to $t_{\mathrm{acr}} = 0.44$. The results are obtained by at least 10^4 repeated runs of the simulation. (Example 5.4)

results are shown in Fig. 5.11 for limit-state functions g_1 to g_5 (as defined in Section 3.5.3). For limit-state function g_2, the parameter κ is set to 10. For limit-state function g_5, the parameter m is set to 4.

The plots in Fig. 5.11 suggest that a single run of Subset Simulation with a large number K of samples is slightly better than multiple smaller runs of SuS with a reduced number of samples. However, multiple smaller runs of SuS are more helpful to quantify the uncertainty in the estimated $p_{f,\mathrm{SuS}}$ than a single large run of SuS.

Example 5.5. *Subset Simulation: influence of the shape of intermediate failure domains*:
This example investigates limit-state function g_5 for different parameter values $m \in \{0, 1, 2, 3, 4\}$. The shape of the actual failure domain does not depend on m. However, the shape of the intermediate failure domains is influenced by m. For $m = 0$, the intermediate failure domains have an optimal shape. For $m = 4$, the intermediate failure domains have an unfavorable shape. The number of uncertain model parameters is set to 10. The target acceptance rate is set to $t_{\mathrm{acr}} = 0.3$. The bias in the mean and median of the estimated $p_{f,\mathrm{SuS}}$, as well as the 1%, 5%, 95% and 99% quantiles of the estimated $p_{f,\mathrm{SuS}}$ are investigated for limit-state function g_5 and different m. The results are shown in Fig. 5.12.

Even though the shape of the actual failure domain does not depend on m, the uncertainty in the estimated $p_{f,\mathrm{SuS}}$ depends considerably on m. The skewness increases strongly with increasing m and decreasing P_f. For $m = 0$, the distribution of $p_{f,\mathrm{SuS}}$ is only slightly skewed. For $m = 4$ and $P_f < 10^{-6}$, the median deviates significantly from the mean. This has to be attributed to the MCMC sampling used to generate samples conditional on the selected intermediate failure domains. If the shape of the final or the intermediate failure domains is not optimal for the selected MCMC sampling strategy, the uncertainty about $p_{f,\mathrm{SuS}}$ increases. This is a disadvantage of Subset Simulation compared to Monte Carlo simulation: In Monte Carlo integration, the estimated probability of failure depends only on P_f, but not the formulation of the limit-state function and not on the shape of the final failure domain.

Next we look at the confidence interval spanned by the 1% and 99% quantiles of Subset Simulation and Monte Carlo simulation. For Monte Carlo simulation, the number of samples is set equal to the average number of limit-state function calls needed in the corresponding Subset Simulation run. The confidence intervals of Monte Carlo simulation can be evaluated explicitly for a given number of samples and known P_f: Then number of samples that fall in the failure domain follows a binomial distribution, and $p_{f,\mathrm{MCS}}$ is obtained by dividing the number of samples in the failure domain by the total number of samples. For up to two subset levels, the Monte Carlo simulation returns a more narrow confidence interval than Subset Simulation. However, generally this is of little practical relevance, as reliabilities in this magnitude can be approximated well with both methods.

Example 5.6. *Q-Q plots of the average $p_{f,SuS}$ from multiple runs of Subset Simulation*:
In this example, Q-Q plots of the average $p_{f,\mathrm{SuS}}$ of $A \in \{10, 100, 10^3, 10^4\}$ independent runs of Subset simulation are generated for limit-state function g_5 with $m = 4$. The studied reference

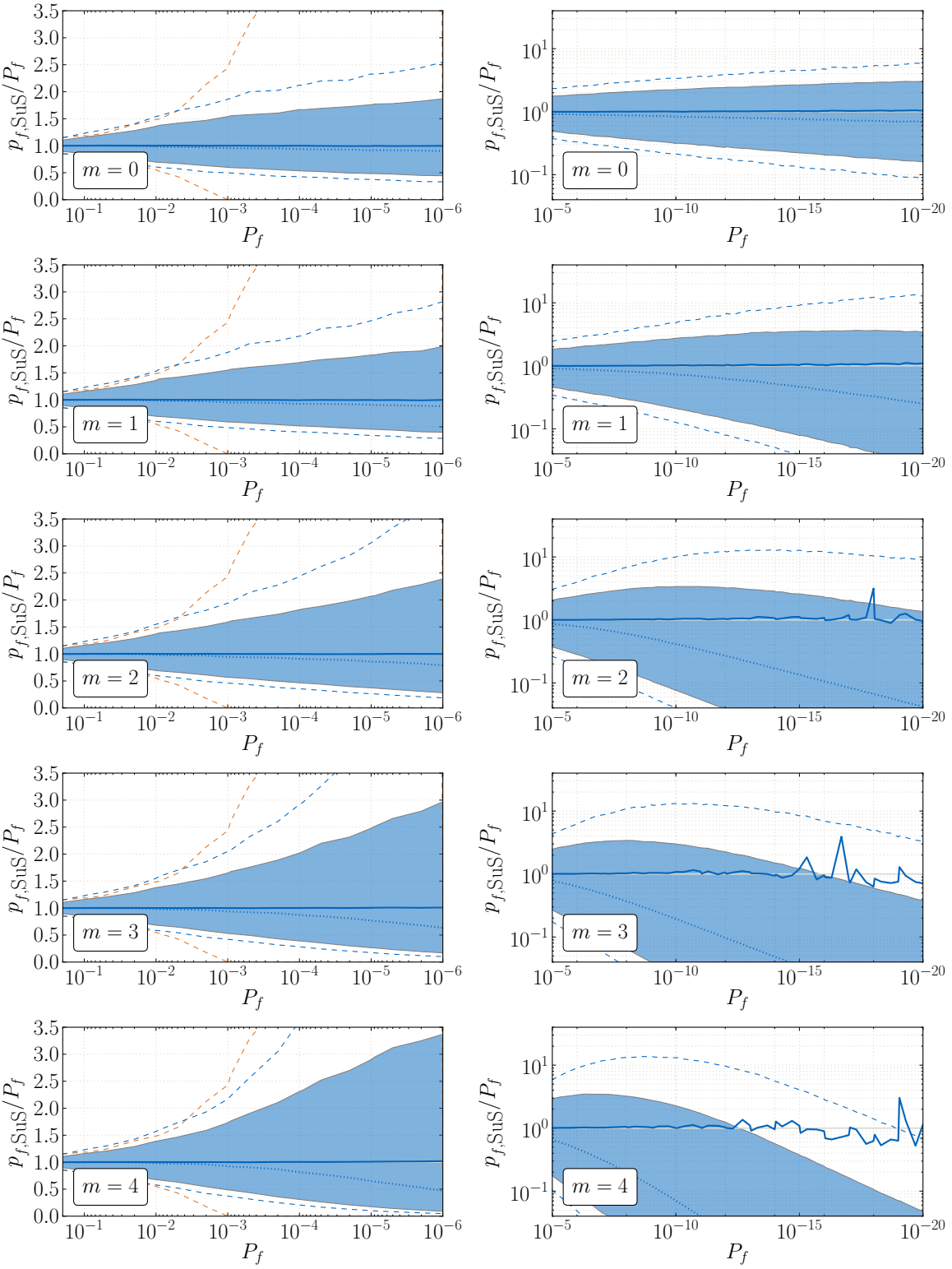


Figure 5.12: The performance of limit-state function g_5 is investigated for different parameter values $m \in \{0, 1, 2, 3, 4\}$. The number of samples per subset level is set to $K = 10^3$, and the target acceptance rate is set to $t_{\mathrm{acr}} = 0.44$. The mean of $p_{f,\mathrm{SuS}}/P_f$ is represented by the continuous line. The bias in the SuS estimate $\mathrm{E}[p_{f,\mathrm{SuS}}]$ is *zero*, if $\mathrm{E}[p_{f,\mathrm{SuS}}]/P_f = 1$. The highlighted area shows the 90% confidence interval of $p_{f,\mathrm{SuS}}/P_f$, defined in terms of the 5% and 95% quantiles. The dashed blue lines represent the 1% and 99% quantiles. The dashed orange lines represent the 1% and 99% quantiles that one would obtain with Monte Carlo simulation and the number of samples set equal to the average number of required limit-state function calls in the corresponding SuS run. The dotted line shows the median of $p_{f,\mathrm{SuS}}/P_f$. (Example 5.5)

probabilities of failure are chosen as $P_f \in \{10^{-3}, 10^{-9}, 10^{-20}\}$. The number of uncertain model parameters is set to 10. The target acceptance rate is set to $t_{\text{acr}} = 0.3$. The Normal and the log-Normal distribution are selected such that the mean is equal to $\text{E}[p_{f,\text{SuS}}/P_f]$, and the variance is equal to $\text{Var}[p_{f,\text{SuS}}/P_f]/\sqrt{A}$, where $\text{Var}[p_{f,\text{SuS}}/P_f]$ refers to the variance obtained from a single run of Subset Simulation. According to the central limit theorem, the distribution of the average $p_{f,\text{SuS}}$ converges for increasing A asymptotically to the employed Normal distribution. The Q-Q plots are shown in Fig. 5.13.

For $P_f = 10^{-3}$, the Normal and log-Normal distribution provide a reasonable fit already for $A = 10$, where the log-Normal distribution exhibits a slightly better fit than the Normal distribution. However, for $P_f = 10^{-9}$ and $P_f = 10^{-20}$, the Normal and the log-Normal distribution no longer describe the distribution of the average $p_{f,\text{SuS}}$ well. Even the average estimated probability of failure from 10^4 independent runs of Subset Simulation is not approximated well by a Normal distribution. Consequently, also the coefficient of variation of the average $p_{f,\text{SuS}}$ is not a good measure to quantify the uncertainty about the average $p_{f,\text{SuS}}$, as the underlying distribution is still asymmetric even for a large number of repeated SuS runs.

For $P_f = 10^{-9}$ and $P_f = 10^{-20}$ the Normal distribution associates too much probability weight with too small $p_{f,\text{SuS}}$. For the investigate limit-state function g_5 with $m = 4$, the 99% quantile of the Normal distribution is comparatively close to the true value. However, for larger quantiles, the Normal distribution puts not enough weight to large $p_{f,\text{SuS}}$. The log-Normal distribution is also not an appropriate probabilistic model, however it provides a slightly better fit than the Normal distribution.

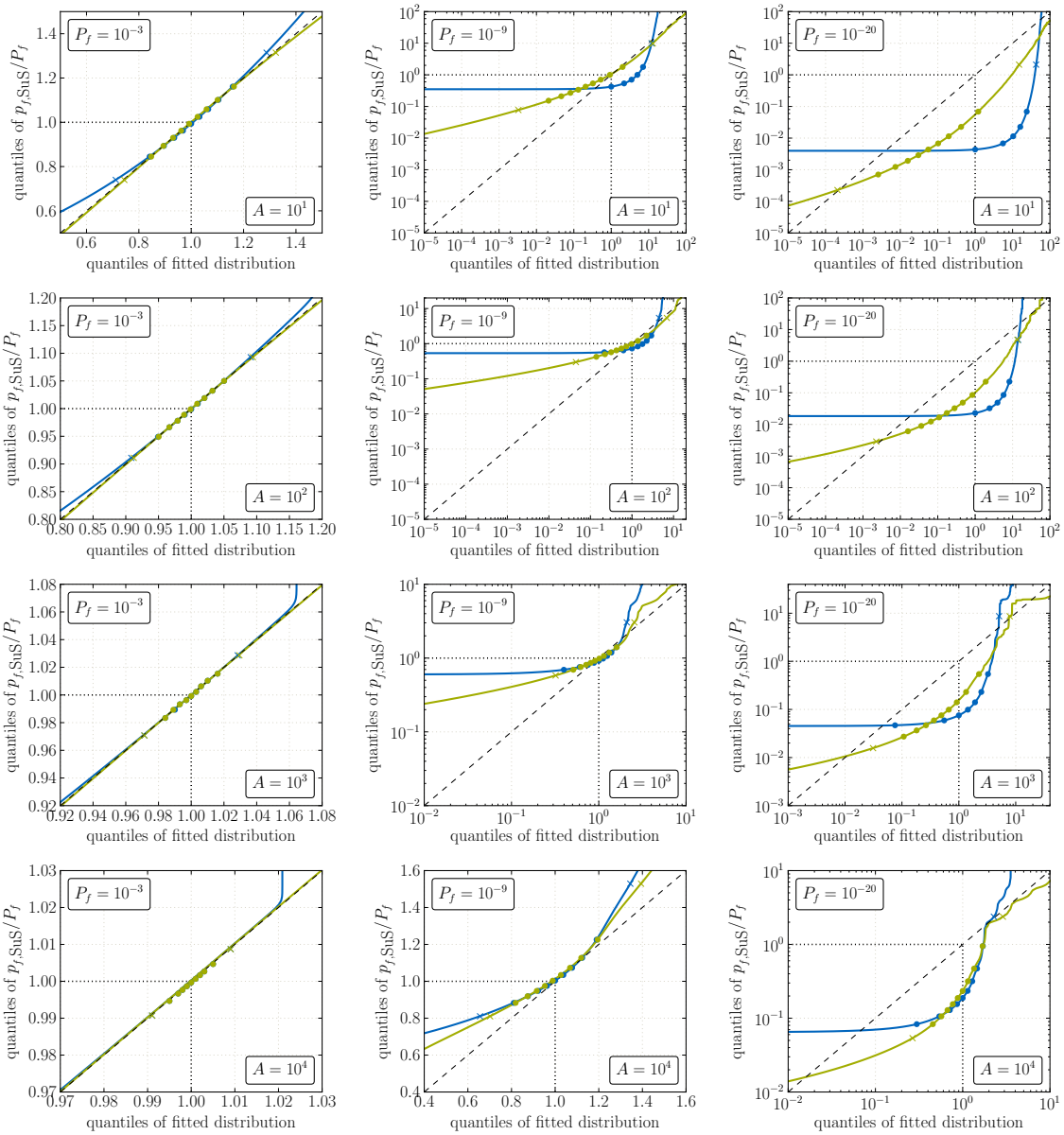


Figure 5.13: Q-Q plots of the average $p_{f,\mathrm{SuS}}/P_f$ of A independent SuS runs. Limit-state function function g_5 with $m = 4$ is investigate for different P_f. The number of samples per level and the target acceptance rate of SuS is fixed to $K = 10^3$ and $t_{\mathrm{acr}} = 0.3$, respectively. The blue line compares the fit with respect to a b*Normal* distribution, and the green line compares the fit with respect to a *log-Normal* distribution. Both distributions are fitted such that the mean corresponds to $\mathrm{E}[p_{f,\mathrm{SuS}}/P_f]$, and the variance is equal to $\mathrm{Var}[p_{f,\mathrm{SuS}}/P_f]/\sqrt{A}$. The large points in the plot indicate the deciles (i.e., the 10%, 20%, ..., 90% quantiles). The crosses indicates the 1% and 99% quantiles. (Example 5.6)

Chapter 6

Bayesian Analysis

6.1 Introduction

Let $\boldsymbol{\theta}$ denote the vector of uncertain model parameters, and let $\mathcal{M}$ denote the associated *stochastic model class* introduced in Section 4.1. In Bayesian inference, two fundamentally different types of information related to $\boldsymbol{\theta}$ are combined:

Prior information: Information about $\boldsymbol{\theta}$ that we already acquired in the past and/or that expresses our *initial belief*. The plausibilities of different realizations of the parameter vector $\boldsymbol{\theta}$ are expressed in terms of a probability density function (PDF), denoted $p(\boldsymbol{\theta}|\mathcal{M})$. The probabilistic model associated with $p(\boldsymbol{\theta}|\mathcal{M})$ is referred to as *prior distribution*. Probabilistic modeling approaches for the prior distribution are discussed in Section 6.4.4.

Data: Information that becomes available to us in form of *measurements or observations*, denoted $\mathcal{D}$. The data $\mathcal{D}$ is embedded in the Bayesian inference through the *likelihood function* $L(\boldsymbol{\theta}|\mathcal{D},\mathcal{M})$. The likelihood expresses the *plausibility* of observing $\mathcal{D}$ given a certain $\boldsymbol{\theta}$, i.e., $L(\boldsymbol{\theta}|\mathcal{D},\mathcal{M}) \propto p(\mathcal{D}|\boldsymbol{\theta},\mathcal{M})$.

The *posterior belief* emanates from combining the *prior information* and the *data* (see Fig. 6.1). The PDF of the posterior is denoted by $p(\boldsymbol{\theta}|\mathcal{D},\mathcal{M})$. The learning process in Bayesian inference is formalized through *Bayes' theorem* as:

$$p(\boldsymbol{\theta}|\mathcal{D},\mathcal{M}) = {c_{\mathrm{E}|\mathcal{M}}}^{-1} \cdot L(\boldsymbol{\theta}|\mathcal{D},\mathcal{M}) \cdot p(\boldsymbol{\theta}|\mathcal{M}) \; . \tag{6.1}$$

The constant $c_{\mathrm{E}|\mathcal{M}}$ in Eq. (6.1) acts as a normalizing scalar and is defined as:

$$c_{\mathrm{E}|\mathcal{M}} = \int_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\mathcal{D},\mathcal{M}) \cdot p(\boldsymbol{\theta}|\mathcal{M}) \, \mathrm{d}\boldsymbol{\theta} \tag{6.2}$$

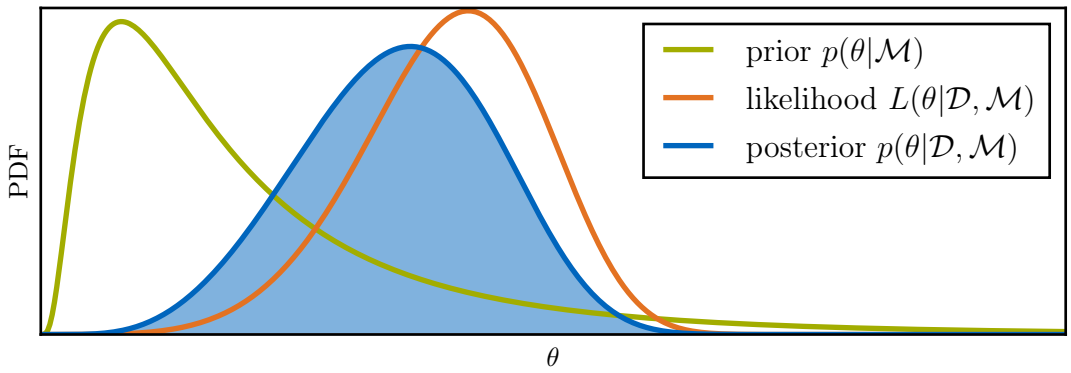


Figure 6.1: Information from the prior $p(\boldsymbol{\theta}|\mathcal{M})$ and the data $L(\boldsymbol{\theta}|\mathcal{D},\mathcal{M})$ is combined to a posterior belief $p(\boldsymbol{\theta}|\mathcal{D},\mathcal{M})$.

Note that the posterior is conditional on $\mathcal{M}$. Consequently, if the assumptions contained in $\mathcal{M}$ are invalid or improper, the corresponding posterior plausibilities can be misleading (see also Sections 4.1.2 and 6.4.2).

Example 6.1. *rate of vehicles on a street* :
We are interested in the number of vehicles passing in front of a gymnasium in the rush hour on weekdays between 4pm and 7pm. It is assumed that the number of vehicles passing in any fixed time interval of length Δt follows a Poisson distribution with a constant mean $\lambda \cdot \Delta t$, where λ denotes the rate of vehicles.

We quantify our prior belief about the value of λ as follows: With a probability of 5%, the value of λ is larger than 2 per minute, and with a probability of 95%, the value of λ is smaller than 20 per minute. Additionally, we choose the Gamma distribution as prior distribution for λ. Based on this assumptions, the mean and standard deviation of the Gamma distribution that represents our uncertainty about λ can be evaluated as 8.96 and 5.74 cars per minute, respectively.

In order to reduce our uncertainty about λ, we hire a student that counts at five different days for an hour the number of cars passing in front of a gymnasium. After performing the measurements, the students submits the following report:

day 1: Counted 358 cars in 60 minutes.

day 2: Counted 277 cars in 50 minutes.

day 3: Sat there for 80 minutes. However, fell asleep at some point – don't know how long. Counted 283 cars in the time not asleep.

day 4: Counted for 60 minutes. The number of cars that passed is either 352 or 353.

day 5: Counted 195 cars in 30 minutes.

Based on the report we can formulate the likelihood functions of the five days as (where λ has unit *one* car per minute):

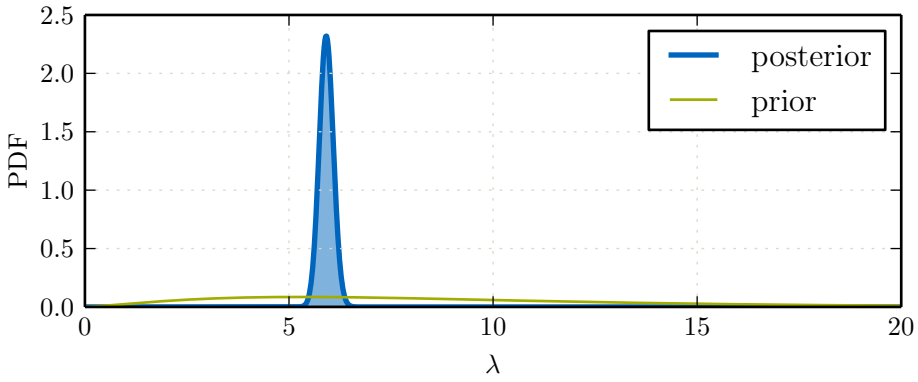


Figure 6.2: Prior and posterior PDF for λ. (Example 6.1)

day 1: $L_1 = (\lambda \cdot 60)^{358} \cdot \exp(-\lambda \cdot 60)/(358!)$

day 2: $L_2 = (\lambda \cdot 50)^{277} \cdot \exp(-\lambda \cdot 50)/(277!)$

day 3: $L_3 = 1 - \exp(-\lambda \cdot 80) \cdot \sum_{i=0}^{282} (\lambda \cdot 80)^i/(i!)$

day 4: $L_4 = \frac{1}{2}\left[(\lambda \cdot 60)^{352} \cdot \exp(-\lambda \cdot 60)/(352!) + (\lambda \cdot 60)^{353} \cdot \exp(-\lambda \cdot 60)/(353!)\right]$

day 5: $L_5 = (\lambda \cdot 30)^{195} \cdot \exp(-\lambda \cdot 30)/(195!)$

Non-overlapping observations of a Poisson process are independent. Thus, the individual likelihood functions can be multiplied to get the combined likelihood function:

$$L = \prod_{j=1}^{5} L_i \tag{6.3}$$

The posterior is defined in Eq. (6.1); i.e., the product of prior density and likelihood function is proportional to the posterior density function. The prior and the posterior distribution are illustrated in Fig. 6.2. The mean and standard deviation of the posterior is 5.92 and 0.17 cars per minute, respectively.

Note that the data the student submitted is incomplete: One day he fell asleep, the other day he was not exactly sure if 352 or 353 cars passed. Nonetheless, such incomplete observations can be considered within a Bayesian inference. The example could even be modified slightly to infer how long the student was asleep on *day 3*.

6.2 Evidence of a stochastic model class

6.2.1 Definition and interpretation

$c_{\mathrm{E}|\mathcal{M}}$ defined in Eq. (6.2) is a measure for the plausibility of the investigated model class $\mathcal{M}$; i.e., $c_{\mathrm{E}|\mathcal{M}} \propto p(\mathcal{D}|\mathcal{M})$. In the Bayesian community $c_{\mathrm{E}|\mathcal{M}}$ is usually referred to as the *evidence* of the assumed model class [Beck and Yuen, 2004], however, $c_{\mathrm{E}|\mathcal{M}}$ is also known as *marginal likelihood* or *integrated likelihood*. The evidence is essentially the expectation of the likelihood

with respect to the prior distribution, i.e.:

$$c_{\mathrm{E}|\mathcal{M}} = \mathrm{E}_{\boldsymbol{\theta}|\mathcal{M}}\left[L(\boldsymbol{\theta}|\mathcal{D},\mathcal{M})\right] \tag{6.4}$$

Except for special cases it is usually a non-trivial task to evaluate the evidence associated with stochastic model class $\mathcal{M}$.

If $\int_{\mathcal{D}} L(\boldsymbol{\theta}|\mathcal{D},\mathcal{M})\,\mathrm{d}\mathcal{D} = 1$ or, equivalently, $L(\boldsymbol{\theta}|\mathcal{D},\mathcal{M}) = p(\mathcal{D}|\boldsymbol{\theta},\mathcal{M})$, then $c_{\mathrm{E}|\mathcal{M}} = p(\mathcal{D}|\mathcal{M})$. If not explicitly mentioned otherwise, in the following it is assumed that $L(\boldsymbol{\theta}|\mathcal{D},\mathcal{M}) = p(\mathcal{D}|\boldsymbol{\theta},\mathcal{M})$. Note that in a general setting, a likelihood function is, unlike a probability density function, not necessarily normalized.

The log-evidence can be expressed by the following difference [Muto and Beck, 2008; Cover and Thomas, 2006]:

$$\begin{aligned}\ln p(\mathcal{D}|\mathcal{M}) = {} & \mathrm{E}_{\boldsymbol{\theta}|\mathcal{D},\mathcal{M}}\left[\ln p(\mathcal{D}|\boldsymbol{\theta},\mathcal{M})\right] \\ & - \mathrm{E}_{\boldsymbol{\theta}|\mathcal{D},\mathcal{M}}\left[\ln \frac{p(\boldsymbol{\theta}|\mathcal{D},\mathcal{M})}{p(\boldsymbol{\theta}|\mathcal{M})}\right]\end{aligned} \tag{6.5}$$

where the expectations are taken with respect to the posterior distribution. The first term in Eq. (6.5) is a measure for the *average data-fit* of the stochastic model class [Beck, 2010] with respect to the employed likelihood function. The second term in Eq. (6.5) represents the *relative entropy* or *Kullback–Leibler divergence*, denoted $D_{\mathrm{KL}}\left(p(\boldsymbol{\theta}|\mathcal{D},\mathcal{M})\|p(\boldsymbol{\theta}|\mathcal{M})\right)$; it measures the *information gain* of the posterior relative to the prior (see Section 2.4.5).

The *relative entropy* serves as a penalty-term in Eq. (6.5), since it is always non-negative: If two model classes have the same data-fit, the stochastic model for which observation $\mathcal{D}$ is less surprising has the larger plausibility. This behavior is called Bayesian Ockham razor [Beck, 2010, 2014]. In other words: A good model should have a good data-fit, and at the same time should have a posterior that is similar to the prior. The plausibility of the stochastic model class decreases if either the data fit decreases, or if the difference between the posterior and the prior (i.e., the relative entropy) increases. Note that the evidence $c_{\mathrm{E}|\mathcal{M}}$ is, therefore, influenced by both the likelihood and the prior.

Example 6.2. *rate of vehicles on a street (cont'd)* :
Continuing Example 6.1, the evidence of the problem can be evaluated as $2.57 \cdot 10^{-9}$. Thus, the log-evidence is -19.78. The associated average data-fit is -16.97, and the relative entropy is 2.81. For this simple example, the mentioned quantities were computed by means of numerical integration.

Example 6.3. *M-dimensional independent Normal prior and likelihood* :

Let the prior distribution be $p(\boldsymbol{\theta}) = \prod_{i=1}^{M} \varphi(\theta_i)$, where M is the dimension of the problem (i.e., number of uncertain parameters). The likelihood is defined as $L(\boldsymbol{\theta}|\mathcal{D}) = \sigma_l^{-M} \cdot \prod_{i=1}^{M} \varphi\left(\frac{\theta_i - \mu_l}{\sigma_l}\right)$, where $\mu_l \in \mathbb{R}$ is the mean of the likelihood function and $\sigma_l \in \mathbb{R}_{>0}$ is the standard deviation of the likelihood function.

The maximum value the likelihood can take is

$$L_{\max}(\sigma_l, M) = \left(2\pi \cdot \sigma_l^2\right)^{-\frac{M}{2}} \tag{6.6}$$

The posterior of each compontent θ_i is an independent Normal distribution with mean and standard deviation:

$$\mu_{\theta_i|\mathcal{D}} = \frac{\mu_l}{\sigma_l^2 + 1} \tag{6.7}$$

$$\sigma_{\theta_i|\mathcal{D}} = \frac{\sigma_l}{\sqrt{\sigma_l^2 + 1}} \tag{6.8}$$

The evidence and log-evidence can be derived as:

$$c_{\mathrm{E}|\mathcal{M}} = \left(\sigma_l^2 + 1\right)^{-\frac{M}{2}} \cdot \left[\varphi\left(\frac{\mu_l}{\sqrt{\sigma_l^2 + 1}}\right)\right]^M \tag{6.9}$$

$$\ln\left(c_{\mathrm{E}|\mathcal{M}}\right) = -\frac{M}{2} \cdot \left[\ln(\sigma_l^2 + 1) + \frac{\mu_l^2}{\sigma_l^2 + 1} + \ln(2\pi)\right] \tag{6.10}$$

Note: The following results are derived, because they are required at a later point.

We are interested in the distribution $L|\mathcal{D}$ of likelihood values for samples $\boldsymbol{\theta}|\mathcal{D}$ that follow the posterior distribution. To simplify the notation, we introduce:

$$d_\mu = \sqrt{\sum_{i=1}^{M} \left(\mu_l - \mu_{\boldsymbol{\theta}|\mathcal{D}}\right)^2} = \sqrt{M} \cdot \mu_l \cdot \left(1 - \frac{1}{\sigma_l^2 + 1}\right), \qquad \text{with } d_\mu \geq 0 \tag{6.11}$$

as well as the distance R_{L} between sample $\boldsymbol{\theta}$ and the center of the likelihood function:

$$R_{\mathrm{L}} = \sqrt{\sum_{i=1}^{M} (\theta_i - \mu_l)^2} \tag{6.12}$$

Note that all samples $\boldsymbol{\theta}$ with the same R_{L} have the same likelihood value $L \in [0, L_{\max}]$. Thus, the likelihood function can be expressed in terms of R_{L} as:

$$L(R_{\mathrm{L}}) = \left(2\pi \cdot \sigma_l^2\right)^{-\frac{M}{2}} \cdot \exp\left(-\frac{R_{\mathrm{L}}^2}{2\sigma_l^2}\right) \tag{6.13}$$

The inverse function $R_{\mathrm{L}}(L)$ of $L(R_{\mathrm{L}})$ is:

$$R_{\mathrm{L}}(L) = \sigma_l \cdot \sqrt{-2 \cdot \ln\left[L \cdot \left(2\pi \cdot \sigma_l^2\right)^{\frac{M}{2}}\right]} \tag{6.14}$$

Furthermore, the distance between sample $\boldsymbol{\theta}$ and the center of the posterior distribution is

denoted as:

$$R = \sqrt{\sum_{i=1}^{M} \left(\theta_i - \mu_{\boldsymbol{\theta}|\mathcal{D}}\right)^2} \tag{6.15}$$

It can be shown that the probability $\Pr\left[L(\boldsymbol{\theta}) \geq L(R_{\mathrm{L}})|R, R_{\mathrm{L}}, \mathcal{D}\right]$ that sample $\boldsymbol{\theta}|R, \mathcal{D}$ with R conditionally fixed has a likelihood value larger than $L(R_{\mathrm{L}})$ is:

- If $R > d_\mu + R_{\mathrm{L}}$ then $\Pr\left[L(\boldsymbol{\theta}) \geq L(R_{\mathrm{L}})|R, R_{\mathrm{L}}, \mathcal{D}\right] = 0$.
- If $R < d_\mu - R_{\mathrm{L}}$ then $\Pr\left[L(\boldsymbol{\theta}) \geq L(R_{\mathrm{L}})|R, R_{\mathrm{L}}, \mathcal{D}\right] = 0$.
- If $R \leq R_{\mathrm{L}} - d_\mu$ then $\Pr\left[L(\boldsymbol{\theta}) \geq L(R_{\mathrm{L}})|R, R_{\mathrm{L}}, \mathcal{D}\right] = 1$.
- Otherwise, $\Pr\left[L(\boldsymbol{\theta}) \geq L(R_{\mathrm{L}})|R, R_{\mathrm{L}}, \mathcal{D}\right]$ is equal to the probability that the angle between a M-dimensional Normal vector and an arbitrarily selected vector is smaller than angle 2ω, where angle ω is defined as follows: Lets assume a triangle with sides of length R, R_{L} and d_μ. The angle between sides R and R_{L} is denoted as ω, with

  $$\cos(\omega) = \frac{R^2 + d_\mu^2 - R_{\mathrm{L}}^2}{2 \cdot R \cdot d_\mu} \tag{6.16}$$

  with $\cos(\omega) = 0$ if $R = 0$. In Section 2.3.4.3.3 it is shown that the quantity $\cos^2(\omega)$ is *beta*-distributed with shape parameters $\alpha = 0.5$ and $\beta = (M-1)/2$. Consequently, $1 - cos^2(\omega)$ is also *beta*-distributed with shape parameters $\alpha = (M-1)/2$ and $\beta = 0.5$. Thus,

  $$\Pr\left[L(\boldsymbol{\theta}) \geq L(R_{\mathrm{L}})|R, R_{\mathrm{L}}, \mathcal{D}\right] = \frac{1}{2} \cdot \left(1 - \operatorname{sign}(\cos\omega) \cdot P_{\mathrm{beta},\,\alpha=0.5,\,\beta=(M-1)/2}(\cos^2(\omega))\right) \tag{6.17}$$

  where $P_{\mathrm{beta},\,\alpha,\,\beta}(\cdot)$ denotes the CDF of the beta distribution.

In order to get the probability that a posterior sample has a likelihood value larger than $L(R_{\mathrm{L}})$, R must be integrated out by means of the total probability theorem. As R follows a *chi* distribution with M degrees of freedom:

$$\Pr\left[L(\boldsymbol{\theta}) \geq L(R_{\mathrm{L}})|R_{\mathrm{L}}, \mathcal{D}\right] = \int_{\max(0, d_\mu - R_{\mathrm{L}})}^{d_\mu + R_{\mathrm{L}}} \Pr\left[L(\boldsymbol{\theta}) \geq L(R_{\mathrm{L}})|R, R_{\mathrm{L}}, \mathcal{D}\right] \cdot \frac{1}{\sigma_l} \cdot p_{\mathrm{chi},M}\left(\frac{R}{\sigma_l}\right) \mathrm{d}R \tag{6.18}$$

where $p_{\mathrm{chi},M}(\cdot)$ is the PDF of a *chi* distribution with M degrees of freedom.

6.2.2 Uniqueness of the evidence of $\mathcal{M}$

The evidence is a unique property of the selected stochastic model class $\mathcal{M}$ and invariant to a transformation (Section 2.3.1.6) of the underlying probabilistic description. Let $\mathrm{T} : \mathbf{Z} \to \boldsymbol{\theta}$ be the transformation of random variable $\mathbf{Z}$ to $\boldsymbol{\theta}$. The evidence can then be written as:

$$c_{\mathrm{E}|\mathcal{M}} = \int_{\mathbf{Z}} L(\mathrm{T}(\mathbf{Z})|\mathcal{D}, \mathcal{M}) \cdot p(\mathbf{Z}|\mathcal{M})\, \mathrm{d}\mathbf{Z} = \mathrm{E}_{\mathbf{Z}|\mathcal{M}}\left[L\left(\mathrm{T}(\mathbf{Z})|\mathcal{D}, \mathcal{M}\right)\right] \tag{6.19}$$

This follows directly from inserting Eq. (6.4) in Eq. (A.7).

6.3 Bayesian model class selection and model averaging

6.3.1 Bayesian model class selection

For a set $\mathbf{M}$ of m competing stochastic model classes $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_m$, the evidence $p(\mathcal{D}|\mathcal{M}_i)$, $i \in \{1, \ldots, m\}$ allows us to evaluate the posterior probabilities $\Pr(\mathcal{M}_i|\mathcal{D}, \mathbf{M})$ of the individual stochastic model classes:

$$\Pr(\mathcal{M}_i|\mathcal{D}, \mathbf{M}) = \frac{p(\mathcal{D}|\mathcal{M}_i) \cdot \Pr(\mathcal{M}_i|\mathbf{M})}{\sum_{j=1}^{m} p(\mathcal{D}|\mathcal{M}_j) \cdot \Pr(\mathcal{M}_j|\mathbf{M})} \tag{6.20}$$

where $\Pr(\mathcal{M}_i|\mathbf{M})$ is the prior probability of the ith model class. The posterior probabilities $\Pr(\mathcal{M}_i|\mathcal{D}, \mathbf{M})$ are required for Bayesian model class selection and Bayesian model averaging [MacKay, 1992a; Wasserman, 2000].

Note that $\Pr(\mathcal{M}_i|\mathcal{D}, \mathbf{M})$ is the posterior plausibility of model class $\mathcal{M}_i$ with respect to all model classes contained in $\mathbf{M}$, and not the posterior probability that model class $\mathcal{M}_i$ is the true model. It is not implicitly assumed that $\mathbf{M}$ contains the *perfect* model (there is no such thing as a perfect model). This distinction is crucial for a correct interpretation of Bayesian model class selection and model averaging: If all models contained in $\mathbf{M}$ are insufficient or inadequate, it will not become apparent that none of the models produces good results by just looking at the posterior plausibilities of the individual model classes. The best models in $\mathbf{M}$ set the standard to which all model classes are compared.

Therefore, it can be considered good practice to include at least one model in $\mathbf{M}$ that is known to be relatively simplistic. If the posterior plausibility of the simple model is not small compared to the model classes with the largest posterior plausibilities, this might be an indicator that the overall performance of the considered model classes could be poor.

Example 6.4. *M-dimensional independent Normal prior and likelihood (cont'd)*:
This example continues Example 6.3. Now, we assume that the parameters μ_l and σ_l are not known. We try to find the values of μ_l and σ_l that maximize the evidence $c_{\mathrm{E}|\mathcal{M}}$ for fixed M.

For fixed $c_{\mathrm{E}|\mathcal{M}}$, M and σ_l, the mean of the likelihood can be derived as:

$$\mu_l = \sqrt{\sigma_l^2 + 1} \cdot \sqrt{-\frac{2}{M} \cdot \ln\left(c_{\mathrm{E}|\mathcal{M}}\right) - \ln(\sigma_l^2 + 1) - \ln(2\pi)} \tag{6.21}$$

The above expression is only defined if σ_l is selected between:

$$0 < \sigma_l \leq \sqrt{\frac{1}{c_{\mathrm{E}|\mathcal{M}}{}^{\frac{2}{M}} \cdot 2\pi} - 1} \tag{6.22}$$

The upper bound in Eq. (6.22) requires that either $M \leq \frac{-2\ln(c_{\mathrm{E}|\mathcal{M}})}{\ln(2\pi)}$ or $c_{\mathrm{E}|\mathcal{M}} \leq \left(\frac{1}{2\pi}\right)^{M/2}$. Thus,

the most plausible likelihood model under the specified constraints must have evidence equal to $c_{\mathrm{E}|\mathcal{M}} = \left(\frac{1}{2\pi}\right)^{M/2}$. For which $\sigma_l = 0$ and $\mu_l = 0$.

Example 6.5. *Tensile test on timber specimens*:
We are given data[1] of a tensile test on 50 timber specimens. The measured tensile strengths are (in $\mathrm{N/mm^2}$):

```
37.1 36.9 40.5 46.8 37.6 30.0 29.0 27.5 30.7 43.9 29.2 28.7 18.3 34.3 35.0 58.0
20.9 26.8 18.4 34.0 28.4 35.2 27.1 38.0 28.0 30.3 54.9 55.4 32.8 39.5 19.3 54.2
23.1 50.3 66.5 32.7 35.7 36.2 26.4 54.8 27.5 25.3 48.3 47.7 52.9 29.9 42.4 24.8
28.9 26.0
```

In this example, we are interested in the distribution underlying the test specimens. The procedure is as follows:

(A) First, we select the investigated *models*. The following distributions are assessed as candidates for the distribution underlying the test specimens:

model 1: log-Normal distribution,

model 2: Gamma distribution,

model 3: truncated Normal distribution with support $[0, \infty)$.

Additionally, we add a model that we expect to perform worse than the other three models (because we expect the mode of the underlying distribution to be larger than *zero*):

model 4: exponential distribution.

If none of the first three models clearly outperforms *model 4*, we should better scrutinize the validity of the investigated models and our *prior* assumptions. The mean μ_i, $i \in \{1, \ldots, 4\}$ and standard deviation σ_i, $i \in \{1, \ldots, 3\}$ of the investigated distributions are considered uncertain; i.e., are treated as uncertain model parameters.

(B) Next, we select the likelihood that quantifies the plausibility of the data conditional on the uncertain model parameters. To keep the example simple, measurement uncertainties are neglected. In this case, the likelihood function $L_i(\boldsymbol{\theta}_i|\mathcal{D}, \mathcal{M}_i)$ can be expressed as:

$$L_i(\boldsymbol{\theta}_i|\mathcal{D}, \mathcal{M}_i) = p(\mathcal{D}|\boldsymbol{\theta}_i, \mathcal{M}_i) = \prod_{j=1}^{50} p_i(d_j|\mu_i, \sigma_i, \mathcal{M}_i) \tag{6.23}$$

where $\mathcal{D}$ represents the 50 measured tensile strengths, d_j denotes the jth measured tensile strength, $\boldsymbol{\theta}_i$ is the vector of uncertain model parameters ($\boldsymbol{\theta}_i = [\mu_i, \sigma_i]$), and $p_i(d_j|\mu_i, \sigma_i, \mathcal{M}_i)$ is the PDF of the distribution associated with the ith model (i.e., log-Normal for *model 1*, Gamma for *model 2*, truncated Normal for *model 3*, and exponential for *model 4*).

(C) Now, we need to select *prior* distributions for the μ_i, $i \in \{1, \ldots, 4\}$ and σ_i, $i \in \{1, \ldots, 3\}$. It is assumed that all μ_i have the same distribution and all σ_i have the same distribution.

[1] The data is taken from the lecture *Risk Analysis* of *Prof. Straub* at *Technische Universität München* and was originally provided by *Lehrstuhl für Holzbau und Baukonstruktion* at *Technische Universität München*.

Table 6.1: Results of the Bayesian inference: data from a tensile test on timber specimens (see Example 6.5).

stoch. model class	$\Pr(\mathcal{M}_i\|\mathbf{M})$	$p(\mathcal{D}\|\mathcal{M}_i)$	$\Pr(\mathcal{M}_i\|\mathcal{D},\mathbf{M})$	data-fit	rel. entropy
$\mathcal{M}_1$	40%	$9.7 \cdot 10^{-85}$	69%	-189.6	3.9
$\mathcal{M}_2$	30%	$5.4 \cdot 10^{-85}$	30%	-190.2	3.9
$\mathcal{M}_3$	25%	$2.6 \cdot 10^{-86}$	1%	-193	4
$\mathcal{M}_4$	5%	$1.0 \cdot 10^{-100}$	0%	-229.3	0.97

The uncertainty in the μ_i is modeled with a log-Normal distribution that has mean 50 and standard deviation 30. The uncertainty about the value of σ_i is described by a log-Normal distribution that has mean $0.2 \cdot \mu_i$ and standard deviation $0.1 \cdot \mu_i$. The assumptions in model, likelihood and prior are comprised in the stochastic model classes $\mathcal{M}_i$, $i \in \{1, \ldots, 4\}$, and $\mathbf{M} = \{\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4\}$ is the set containing all investigated stochastic model classes.

Furthermore, for stochastic model class selection, prior plausibilities $\Pr(\mathcal{M}_i|\mathbf{M})$ for each of the four stochastic model classes $\mathcal{M}_i$ need to be selected. Based on our initial belief, we assign the following probabilities: $\Pr(\mathcal{M}_1|\mathbf{M}) = 40\%$, $\Pr(\mathcal{M}_2|\mathbf{M}) = 30\%$, $\Pr(\mathcal{M}_3|\mathbf{M}) = 25\%$, $\Pr(\mathcal{M}_5|\mathbf{M}) = 5\%$.

(D) Based on the prior and the likelihood, the evidence $p(\mathcal{D}|\mathcal{M}_i)$ can be evaluated for each stochastic model class $\mathcal{M}_i$ in $\mathbf{M}$. In this simple example, the evidence is evaluated by means of numerical integration of Eq. (6.2). Based on the evidence of all stochastic model classes, the posterior probability $\Pr(\mathcal{M}_i|\mathcal{D},\mathbf{M})$ of each stochastic model class can be calculated by means of Eq. (6.20). The results of the analysis are listed in Table 6.1. The most plausible stochastic model class is $\mathcal{M}_1$, followed by $\mathcal{M}_2$. The posterior plausibility of stochastic model class $\mathcal{M}_3$ is small compared to $\mathcal{M}_1$ and $\mathcal{M}_2$, and the posterior plausibility of $\mathcal{M}_4$ is negligible. $\mathcal{M}_4$ has a considerably smaller data-fit than the other three model classes. The relative entropy of $\mathcal{M}_4$ is comparatively small, as $\mathcal{M}_4$ is modeled by a single uncertain parameter (μ_4). However, the small relative entropy does not significantly alleviate the degradation due to the data-fit term.

This example is continued as Example 6.6 in Section 6.3.2.

6.3.2 Bayesian model averaging

The posterior probabilities $\Pr(\mathcal{M}_i|\mathcal{D},\mathbf{M})$ of the individual stochastic model classes can be employed in the total probability theorem to formulate the posterior distribution of $\boldsymbol{\theta}$ conditional on $\mathbf{M}$; i.e., $p(\boldsymbol{\theta}|\mathcal{D},\mathbf{M})$:

$$p(\boldsymbol{\theta}|\mathcal{D},\mathbf{M}) = \sum_{i=1}^{m} p(\boldsymbol{\theta}|\mathcal{D},\mathcal{M}_i) \cdot \Pr(\mathcal{M}_i|\mathcal{D},\mathbf{M}) \tag{6.24}$$

Note that $\boldsymbol{\theta}$ comprises the joint set of uncertain parameters used in at least one of the m stochastic model classes.

Similarly, the prior distribution of the entire set $\mathbf{M}$ of stochastic model classes can be ex-

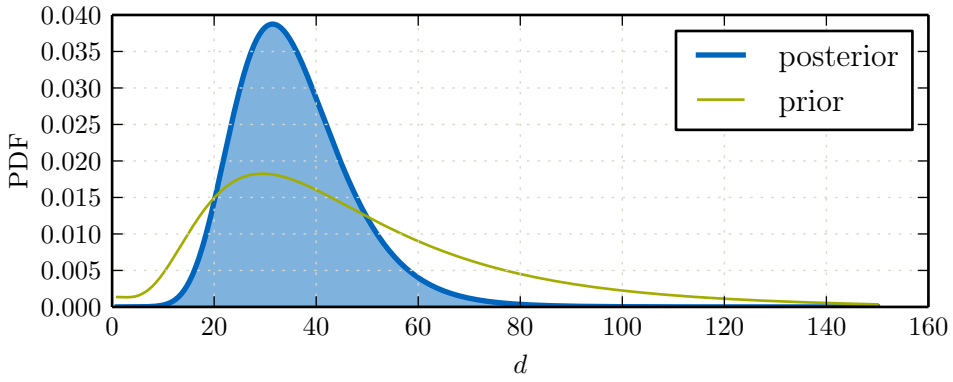


Figure 6.3: Prior and posterior PDF of the set of stochastic model classes $\mathbf{M}$ obtained with Bayesian model averaging. (Example 6.5)

pressed:

$$p(\boldsymbol{\theta}|\mathbf{M}) = \sum_{i=1}^{m} p(\boldsymbol{\theta}|\mathcal{M}_i) \cdot \Pr(\mathcal{M}_i|\mathbf{M}) \tag{6.25}$$

Example 6.6. *Tensile test on timber specimens (cont'd)*:
This example continues Example 6.5; the distribution underlying the test specimens is derived conditional on the set of stochastic model classes $\mathbf{M}$ and *Bayesian model averaging*.

Our prior belief about the distribution of the tensile strength d of the test specimen can be expressed as:

$$p(d|\mathbf{M}) = \sum_{i=1}^{4} \int_0^\infty \int_0^\infty p_i(d|\mu_i, \sigma_i, \mathcal{M}_i) \cdot p(\mu_i, \sigma_i|\mathcal{M}_i)\, \mathrm{d}\sigma_i\, \mathrm{d}\mu_i \cdot \Pr(\mathcal{M}_i|\mathbf{M}) \tag{6.26}$$

where $p(\mu_i, \sigma_i|\mathcal{M}_i)$ is the prior distribution of μ_i and σ_i.

The posterior belief about the distribution of the tensile strength d of the test specimen can be expressed as:

$$p(d|\mathcal{D}, \mathbf{M}) = \sum_{i=1}^{4} \int_0^\infty \int_0^\infty p_i(d|\mu_i, \sigma_i, \mathcal{M}_i) \cdot p(\mu_i, \sigma_i|\mathcal{D}, \mathcal{M}_i)\, \mathrm{d}\sigma_i\, \mathrm{d}\mu_i \cdot \Pr(\mathcal{M}_i|\mathcal{D}, \mathbf{M}) \tag{6.27}$$

where $p(\mu_i, \sigma_i|\mathcal{D}, \mathcal{M}_i)$ is the posterior distribution of μ_i and σ_i (see Example 6.5).

The prior and posterior PDF of the set of stochastic model classes $\mathbf{M}$ evaluated with Eqs. (6.26) and (6.27) are shown in Fig. 6.3.

6.4 The Bayesian modeling framework

6.4.1 Formal representation of data $\mathcal{D}$ in Bayesian inference

Within a Bayesian framework, a major objective is to learn the uncertain parameters $\boldsymbol{\theta}$ of a model that approximates the behavior of a real system: Our uncertainty about the state of the parameters $\boldsymbol{\theta}$ is reduced based on the information $\mathcal{D}$. Typically, the real system produces some output and is driven by its input conditions. In Bayesian inference, the data $\mathcal{D}$ contains observations/measurements of the system output. However, $\mathcal{D}$ can also contain information about the system input. Thus, the available information $\mathcal{D}$ for the Bayesian learning process is composed of the observed output states $\mathbf{z}$ and the available information $\mathbf{s}$ about the system input, i.e., $\mathcal{D} = [\mathbf{z}^\intercal, \mathbf{s}^\intercal]^\intercal$.

To increase clarity, it is often helpful to use a notation that represents the observed output $\mathbf{z}$ and input $\mathbf{s}$ separately. Directly employing this notation in *Bayes' theorem* as stated in Eq. (6.1) gives:

$$p(\boldsymbol{\theta}|\mathbf{z},\mathbf{s},\mathcal{M}) = \frac{\int_{\Gamma_\mathbf{f}} p(\mathbf{z}|\boldsymbol{\theta},\mathbf{f},\mathcal{M}) \cdot p(\mathbf{s}|\mathbf{f},\boldsymbol{\theta}_\mathbf{f},\mathcal{M}) \cdot p(\mathbf{f}|\boldsymbol{\theta}_\mathbf{f},\mathcal{M}) \cdot p(\boldsymbol{\theta}|\mathcal{M})\,\mathrm{d}\mathbf{f}}{p(\mathbf{z},\mathbf{s}|\mathcal{M})} \,. \tag{6.28}$$

where $p(\mathbf{f}|\boldsymbol{\theta},\mathcal{M})$ is the probabilistic description of the model input, and $p(\mathbf{s}|\mathbf{f},\boldsymbol{\theta})$ represents the observation-error of the model input. The quantities $\mathbf{z}$ (observed system output), $\mathbf{s}$ (observed model input) and $\mathbf{f}$ (actual model input) are defined according to Section 4.1. However, the formulation of *Bayes' theorem* as stated in Eq. (6.28) is often not convenient. Usually, it is more appropriate to represent *Bayes' theorem* as:

$$p(\boldsymbol{\theta}|\mathbf{z},\mathbf{s},\mathcal{M}) = \frac{\int_{\Gamma_\mathbf{f}} p(\mathbf{z}|\boldsymbol{\theta},\mathbf{f},\mathcal{M}) \cdot p(\mathbf{f}|\mathbf{s},\boldsymbol{\theta}_\mathbf{f},\mathcal{M}) \cdot p(\boldsymbol{\theta}|\mathbf{s},\mathcal{M})\,\mathrm{d}\mathbf{f}}{p(\mathbf{z}|\mathbf{s},\mathcal{M})} \,. \tag{6.29}$$

Contrary to Eq. (6.28), all elements in Eq. (6.29) are conditional on $\mathbf{s}$. This holds also for the evidence: The evidence $p(\mathbf{z}|\mathbf{s},\mathcal{M})$ is expressed as the plausibility of observing output $\mathbf{z}$ conditional on having observed input $\mathbf{s}$.

Note that Eq. (6.28) can be transformed to Eq. (6.29) as follows:

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{z},\mathbf{s},\mathcal{M}) &= \frac{\int_{\Gamma_\mathbf{f}} p(\mathbf{z}|\boldsymbol{\theta},\mathbf{f},\mathcal{M}) \cdot p(\mathbf{s}|\mathbf{f},\boldsymbol{\theta}_\mathbf{f},\mathcal{M}) \cdot p(\mathbf{f}|\boldsymbol{\theta}_\mathbf{f},\mathcal{M}) \cdot p(\boldsymbol{\theta}|\mathcal{M})\,\mathrm{d}\mathbf{f}}{p(\mathbf{z},\mathbf{s}|\mathcal{M})} \\ &= \frac{\int_{\Gamma_\mathbf{f}} p(\mathbf{z}|\boldsymbol{\theta},\mathbf{f},\mathcal{M}) \cdot p(\mathbf{s}|\mathbf{f},\boldsymbol{\theta}_\mathbf{f},\mathcal{M}) \cdot p(\mathbf{f}|\boldsymbol{\theta}_\mathbf{f},\mathcal{M}) \cdot p(\boldsymbol{\theta}|\mathcal{M})\,\mathrm{d}\mathbf{f}}{p(\mathbf{z}|\mathbf{s},\mathcal{M}) \cdot p(\mathbf{s}|\mathcal{M})} \\ &= \frac{\int_{\Gamma_\mathbf{f}} p(\mathbf{z}|\boldsymbol{\theta},\mathbf{f},\mathcal{M}) \cdot p(\mathbf{f}|\mathbf{s},\boldsymbol{\theta}_\mathbf{f},\mathcal{M}) \cdot p(\boldsymbol{\theta}|\mathbf{s},\mathcal{M})\,\mathrm{d}\mathbf{f}}{p(\mathbf{z}|\mathbf{s},\mathcal{M})} \end{aligned}$$

if it is loosely assumed that $p(\boldsymbol{\theta}|\mathcal{M})$ can be replaced with $p(\boldsymbol{\theta}|\mathbf{s},\mathcal{M})$. In this case, Eq. (6.28)

can be viewed as consisting of two concatenated inference problems: Assessing (i) the probabilistic model of the model input, and (ii) the probabilistic model of the system response. In contrast, Eq. (6.29) assesses only the probabilistic model of the system response. Note that the inference problem that assesses the probabilistic model of the model input, i.e.,

$$p(\mathbf{f}|\mathbf{s}, \boldsymbol{\theta}_\mathbf{f}, \mathcal{M}) = \frac{p(\mathbf{s}|\mathbf{f}, \boldsymbol{\theta}_\mathbf{f}, \mathcal{M}) \cdot p(\mathbf{f}|\boldsymbol{\theta}_\mathbf{f}, \mathcal{M})}{p(\mathbf{s}|\mathcal{M})} \tag{6.30}$$

can be solved separately prior to the inference problem specified in Eq. (6.29). When assessing the performance of a stochastic model class by means of Bayesian model class selection and model averaging, we are typically interested only in the performance of the probabilistic model of the system response; i.e., in the inference problem specified in Eq. (6.29).

In addition to that, consider for example the following scenarios that highlight the difference between Eqs. (6.28) and (6.29):

- A bridge is loaded with heavy trucks to induce a test load. The response of the bridge under the test load is measured. Before the trucks enter the bridge, they are weighed. However, measurement uncertainty about the actual weight applied remains.
  In this case, the number and weight of the trucks is pre-specified; the analysis is conditional on the applied test load. Thus, the uncertainty about the actual load applied is most conveniently directly modeled by means of density $p(\mathbf{f}|\mathbf{s}, \boldsymbol{\theta}_\mathbf{f}, \mathcal{M})$.
  Moreover, for Bayesian model class selection and model averaging, we are typically interested in the evidence $p(\mathbf{z}|\mathbf{s}, \mathcal{M})$ instead of $p(\mathbf{z}, \mathbf{s}|\mathcal{M})$: The plausibility of observing the response $\mathbf{z}$ conditional on the applied test load $\mathbf{s}$.

- In a hydrological catchment, precipitation and discharge are observed for a certain time period. The data is used to reduce the uncertainty about the parameters of a model that approximates the hydrological system.
  In Eq. (6.28), the probabilistic model $p(\mathbf{s}|\mathbf{f}, \boldsymbol{\theta}_\mathbf{f}, \mathcal{M}) \cdot p(\mathbf{f}|\boldsymbol{\theta}, \mathcal{M})$ needs to be formulated explicitly, whereas Eq. (6.29) directly requires $p(\mathbf{f}|\mathbf{s}, \boldsymbol{\theta}_\mathbf{f}, \mathcal{M})$. In this case it is much simpler and more straight-forward to directly assume $p(\mathbf{f}|\mathbf{s}, \boldsymbol{\theta}_\mathbf{f}, \mathcal{M})$ known, instead of representing it through Eq. (6.30).

6.4.2 Stochastic model class

In Section 4.1, the concept of a *stochastic model class* was introduced for *forward analysis* (Chapter 4). The *stochastic model class* is also the central element that bundles all our modeling assumptions in *Bayesian analysis*. The definition in Section 4.1 needs, however, the following extension:

Additional to the

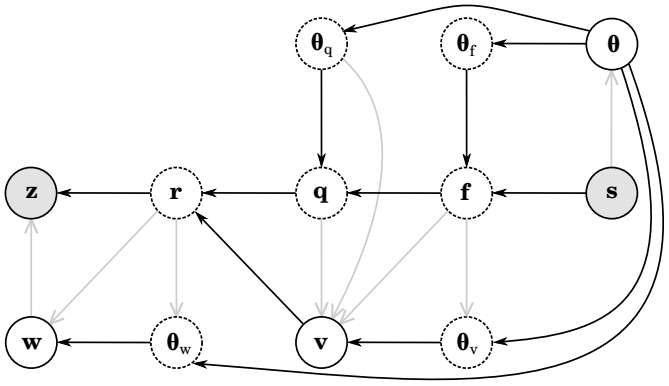


Figure 6.4: Assumed dependency structure in a stochastic model class (as discussed in Sections 4.1 and 6.4.2) represented as a Bayesian network. The figure is an extended variation of Fig. 4.1. The dependencies are represented as arrows. The gray arrows denote dependencies that exist in the real world but that are usually not considered explicitly in modeling. The nodes with a continuous border denote random variables. The nodes with a dashed border can be computed deterministically if all input quantities are conditionally fixed. The gray nodes denote the observed system input and output. The parameter vector $\boldsymbol{\theta}$ is composed of $\boldsymbol{\theta} = \left[\boldsymbol{\theta}_q{}^\intercal, \boldsymbol{\theta}_f{}^\intercal, \boldsymbol{\theta}_v{}^\intercal, \boldsymbol{\theta}_w{}^\intercal\right]^\intercal$, where $\boldsymbol{\theta}_q$ is the vector of uncertain parameters of the deterministic model, $\boldsymbol{\theta}_f$ is the parameter vector of the *input model*, $\boldsymbol{\theta}_v$ is the parameter vector of the *prediction-error model*, and $\boldsymbol{\theta}_w$ is the parameter vector of the *observation-error model*.

(1) *prior probability model* $p(\boldsymbol{\theta}|\mathcal{M})$, the

(2) *stochastic forward model* $p(\mathbf{r}|\mathbf{f}, \boldsymbol{\theta}, \mathcal{M})$, and the

(3) *stochastic input model* $p(\mathbf{f}|\mathbf{s}, \boldsymbol{\theta}_f, \mathcal{M})$,

the (4) *stochastic observation-error model* $p(\mathbf{z}|\mathbf{r}, \boldsymbol{\theta}_w, \mathcal{M})$ is needed, if the observed output states $\mathbf{z}$ of the system are contaminated with measurement errors or relate to imperfect observations. The *stochastic observation-error model* $p(\mathbf{z}|\mathbf{r}, \boldsymbol{\theta}_w, \mathcal{M})$ quantifies our belief in the measurement uncertainty, where $\boldsymbol{\theta}_w$ is the parameter vector of the observation-error model. The stochastic observation-error model $p(\mathbf{z}|\mathbf{r}, \boldsymbol{\theta}_w, \mathcal{M})$ is usually expressed explicitly as $\mathbf{z} = \mathbf{r} + \mathbf{w}$, where $\mathbf{w}$ is the uncertain *observation-error*. Instead of the additive error structure, also a multiplicative error structure could be selected, i.e., $\mathbf{z} = \mathbf{r} \cdot \mathbf{w}$.

A particular example where large observation errors can occur is the observation of runoff from a hydrological catchment: Typically, only water levels are observed directly and so-called rating curves are used to transform water levels to discharges.

Note that strictly we should denote the actually observed output as $\hat{\mathbf{z}}$, that is a particular realization of stochastic variable $\mathbf{z}$. However, if an explicit distinction is not required, we will use $\mathbf{z}$ to denote both the actual observation and the stochastic variable.

Note: In applications of Bayesian inference, the *output prediction-error* is usually the predominant link between the observed system response and the model output. *Observation errors*

(e.g., measurement errors) contribute as well, however, they are often small compared to the *output prediction-error*. Therefore, the probabilistic description of the *output prediction-error* has a considerable influence on the learning process in Bayesian inference. This is a delicate issue, as the *output prediction-error* cannot be quantified exactly (see Section 4.1.4).

The assumed dependency structure of a stochastic modeling class for Bayesian analysis is depicted in Fig. 6.4.

6.4.3 Objectivity of the likelihood function

Contrary to the prior, the likelihood is often considered the objective element of the Bayesian analysis, because it contains the data $\mathcal{D}$. This point of view can, however, be misleading: It is true that the data can be usually regarded as objective – even though $\mathcal{D}$ can also contain subjective assessment. What is often not considered when the likelihood is called *objective* is that modeling and observation errors – if present – constitute a fundamental part of the likelihood. However, the probabilistic description of errors is usually far from objective (consider also the discussion in Section 2.2.5). Thus, it is important to acknowledge that the likelihood is – as is the prior – conditional on $\mathcal{M}$.

6.4.4 Probabilistic modeling approaches for the prior

6.4.4.1 Overview

Typically, the probability models of $\mathbf{v}$, $\mathbf{w}$, $\mathbf{f}$, and $\boldsymbol{\theta}$ are not known explicitly. Sometimes we can express our prior knowledge only in terms of constraints (e.g., unbiased, given variance structure, non-negative) that come either from reasonable assumptions or are imposed by the physics of the problem. In this case, our uncertainty is not only about the specific value of e.g. the first and second moment, but the type of the underlying distribution itself is uncertain. For example, consider the Young's modulus, a typical parameter of structural engineering models: This quantity can take only positive values. If we additionally assume that its mean and standard deviation are – at least conditionally – given, there exist many distributions that can be picked to represent our uncertainty associated with this parameter. This includes the often used log-Normal distribution, but also the Gamma, Weibull or truncated Normal distribution.

In this section we discuss probabilistic modeling approaches for prior distributions. Recommendations for modeling of prior distributions in engineering problems are provided in the summary at the end of the section. For specific guidelines on how to select adequate probability models of error structures see Section 6.4.6.

6.4.4.2 Weakly-informative priors

If no prior knowledge is available or if the available prior information is sparse, *weakly-informative* priors can be used. The underlying assumption is that the posterior is dominated by the data (i.e., by the likelihood function) and the influence of the prior model is negligible. Therefore, the prior distribution is chosen such that it contains nearly no information.

6.4.4.2.1 Uninformative or vague priors

Uninformative or vague priors are often modeled as constant, spanning the entire range of a parameter. If the parameter space is unbounded, the use of constant prior densities spanning the entire parameter space leads to improper prior distributions [Berger et al., 2006]. The use of improper priors needs careful handling and is often better avoided in engineering models because of the following reasons: (i) The resulting posterior distribution is often a proper distribution, but is not guaranteed to be proper. Especially if the problem is solved numerically, identification of improper posterior distributions is difficult [Berger et al., 2006]. (ii) Engineering models are typically driven by many parameters. Even if all uncertain parameters have a proper posterior distribution, the posterior distribution of some parameters can still be quite vague despite a large number of observations. For these parameters, it is not proper to assume an uninformative prior distribution as the prior has a considerable influence on the posterior [Kass and Wasserman, 1996]. Typically, in engineering models the parameters have a physical basis and even if the available prior information is sparse, we typically know that some parameter values are more plausible than others. (iii) Models based on improper priors are typically not fit for Bayesian model class selection and model averaging, because the evidence of the associated stochastic model class is *zero* – and, consequently, the evidence looses its significance. Moreover, priors that are proper but overly diffuse should also be avoided in Bayesian model class assessment, because they enforce a strong preference toward simpler models [Cheung and Beck, 2010].

6.4.4.2.2 Uninformative prior on truncated parameter space

To avoid the problem of unbounded parameter ranges and to simplify the numerical solution of the Bayesian inference problem, the parameter space is sometimes truncated. The bounds are selected such that they contain the assumed relevant posterior range. For physical models, this approach is unsatisfactory because the imposed bounds and in particular the probability jumps at the bounds are unphysical. Additionally, as for *uninformative* or *vague* priors, this approach is not fit for Bayesian model class selection and model averaging: The evidence of the associated stochastic model class depends on the chosen bounds. Moreover, this approach should not be used to compensate for improper posterior distributions, because in this case the resulting posterior distribution will also depend on the chosen bounds.

6.4.4.2.3 Other weakly-informative approaches

The *Jeffreys prior* [Jeffreys, 1946, 1998] is a special weakly-informative prior distribution that is invariant under re-parameterization of the parameter vector. In the so-called *reference prior* approach [Bernardo, 1979; Berger and Bernardo, 1992; Berger et al., 2009] the prior is chosen such that the model-averaged information gain between the prior and the posterior is maximized in an asymptotic sense. For multiparameter problems this approach is often advantageous compared to the *Jeffreys prior*. A disadvantage of the *reference prior* approach is that for complex engineering models it is not straightforward to obtain the prior distribution. A literature review of weakly-informative approaches is provided by [Kass and Wasserman, 1996].

6.4.4.3 Principle of Maximum Information Entropy

If prior knowledge is available, it is a subjective choice to neglect the available prior information. Therefore, the argument that weakly informative priors are "objective" is misguided. One way to incorporate the available prior information is to choose the probability model amongst all feasible models that includes the least amount of information given the imposed constraints [Beck, 2010]; i.e., the *Principle of Maximum Information Entropy* [Jaynes, 1983, 2003] is applied to infer the probability model. This approach offers an objective technique to pick a probability model, given a set of (possibly subjective) constraints.

For example, consider an uncertain parameter that can take only positive values, and whose mean and standard deviation are (assumed) given (e.g., the Young's modulus or hydraulic conductivity). In this case, the *maximum entropy probability (MEP)* model is the truncated Normal distribution. However, the MEP model is only one out of many feasible probability models. For the stated example, also e.g. the log-Normal, Gamma or Weibull distribution could be picked instead of the MEP model.

6.4.4.4 Informative priors

Another way to select the respective probability model is to choose the probability model that appears most adequate given the imposed constraints. Selected probability distributions are listed in Table 6.2, ordered according to their range of support. If more than one probabilistic model is deemed appropriate, different stochastic model classes can be considered: Together with our belief about the prior plausibilities of the considered model classes, the evidence of the stochastic model classes provides a measure for their posterior plausibilities.

This approach is sometimes accused of being subjective, because no formal rule is used to select the prior distribution. However, if multiple alternative probabilistic models are investigated, the objectivity of the approach is at least partially recovered, because the most

Table 6.2: List of selected continuous probability distributions ordered according to their support. N_{para} denotes the number of parameters required for the definition of the distribution. Parameters defining the support of the distribution are not considered in N_{para}.
Note that by truncation, shifting and other transformations, the support of the listed distributions can be adopted.

support	distribution	N_{para}	support	distribution	N_{para}
$(-\infty, \infty)$			$(0, \infty)$		
	Normal	2		log-Normal	2
	Gumbel	2		Gamma	2
	Cauchy	2		Exponential	1
	Student's t (standard)	1		Weibull	2
	Student's t (generalized)	3		Fréchet	2
	Laplace	2		Rayleigh	1
	Logistic	2		Chi-squared	1
				Chi	1
$[a, b]$				Fisher's F	2
	uniform	0		inverse Gaussian	2
	Beta	2		log-Logistic	2
	truncated Normal	2			
	trapezoidal	2			

Table 6.3: Results of the simple toy-example discussed in Example 6.7. The *flat prior* does not maintain the specified constraints for the prior distribution, it is added for illustratory purposes only. The value in each row that is considered advantageous (conservative) compared to the other entries is highlighted in blue. The value in each row that is considered disadvantageous compared to the other entries is highlighted in orange. (Example 6.7)

	MEP prior	log-Normal prior	Gamma prior	flat prior
entropy of the prior	-0.884	-0.891	-0.887	∞
entropy of the posterior	-1.23	-1.07	-1.12	-0.884
evidence	$5.4 \cdot 10^{-3}$	$9.6 \cdot 10^{-3}$	$7.9 \cdot 10^{-3}$	0
relative entropy	3.22	3.54	3.44	∞
data fit	-1.99	-1.11	-1.40	0.884
mean of posterior	1.25	1.29	1.28	1.5
std. dev. of posterior	$7.1 \cdot 10^{-2}$	$8.3 \cdot 10^{-2}$	$7.9 \cdot 10^{-2}$	$10 \cdot 10^{-2}$

plausible model in the set will be favored. In Example 6.7, the MEP prior is compared to alternative prior distributions that also maintain the imposed constraints.

Example 6.7. *Informative prior distributions*:
We consider the following simple toy-example that incorporates prior information: The knowledge about parameter θ is to be updated. All that is a-priori known about θ is the mean $m = 1.0$, the standard deviation $s = 0.1$ and the support $\theta \in [0, \infty)$. For simplicity, the likelihood function is given explicitly as a Normal distribution with mean 1.5 and standard deviation

0.1. The MEP prior for the example at hand is the truncated Normal distribution - which is in this case essentially equivalent to a Normal distribution, since the coefficient of variation of θ is only 10%. We compare the MEP prior with a log-Normal prior and a Gamma prior that also maintain the specified constraints

The results are shown in Table 6.3. For illustratory reasons, also the results for a flat prior spanning the entire positive real line are given. As expected, the entropy of the MEP prior is larger than the entropy of the log-Normal or Gamma prior (the flat prior does not maintain the specified constraints for the prior distribution). However, the model utilizing a log-Normal prior has a larger posterior entropy, a larger evidence and a larger posterior standard deviation compared to the model with the MEP prior. Consequently, the log-Normal prior is preferred over the MEP prior for the example at hand. The only "drawback" of the model with the log-Normal prior is that it is the one surprising us more. However, this disadvantage is more than outweighed by a larger data-fit compared to the model with the MEP prior.

Summarizing the results of Example 6.7: The MEP model can be regarded as an objective choice that incorporates the least amount of prior information. However, the MEP model should not be considered a conservative choice or as a choice that retains the largest uncertainty in the posterior distribution.

6.4.4.5 Summary

This section attempts to give a short overview for modeling of prior distributions in engineering problems:

(1) Typically, parameters in engineering problems are related to physical quantities. For this type of parameters, certain subsets of parameter values can often be identified to be more plausible than parameter values in other subsets. For example, stiffness values are usually neither very close to *zero* nor are they unreasonably large. Moreover, it is also not a good idea to simply prescribe bounds for reasonable parameter values and to consider all values inside the bounds as equally plausible, because the bounds impose incomprehensible jumps in the plausibilities. Instead, rational reasoning suggests that the plausibilities should only vary gradually. As a consequence, *weakly-informative* priors are often not an appropriate choice. Thus, MEP priors that maintain the imposed constraints and other informative prior distributions are often a better choice.

(2) If the choice of a distribution for which only constraints are given is expected to have a considerable influence on the posterior results, we recommend to investigate multiple alternative probability models – and not only the MEP prior. In this case a robust posterior distribution is obtained through Bayesian model class selection or model averaging – provided that the involved computational costs remain manageable. If the choice of the distribution is not considered to have a large influence on the posterior or if the computational costs of a more detailed analysis are too large, the MEP prior is a simple and objective choice.

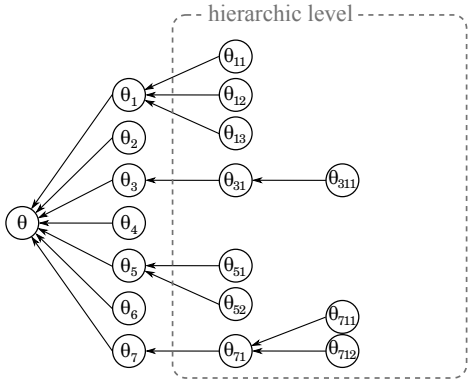


Figure 6.5: Example of a parameter vector $\boldsymbol{\theta}$ represented by a hierarchic structure. By means of a hierarchic probabilistic model, the uncertainties can be represented in an intuitive fashion.

6.4.5 Hierarchical stochastic models

Sometimes it can be appropriate to further parameterize the components of the uncertain parameter vector $\boldsymbol{\theta}$. In this case the stochastic model class is said to be *hierarchical*. An exemplary illustration of a hierarchic representation for parameter vector $\boldsymbol{\theta}$ is given in Fig. 6.5. A hierarchic model structure offers an intuitive and visual way of representing existing uncertainties, and can simplify the probabilistic modeling.

A scenario where it is appropriate to further parameterize the probability model is: Assume we have the following information about an uncertain parameter: (i) The parameter can only take positive values. (ii) The prior mean is known. (iii) For the prior standard deviation we have some incomplete knowledge; i.e., we cannot directly specify a fixed value for the standard deviation. In this case we can first find a reasonable distribution conditional on a fixed standard deviation (e.g., the MEP model) and, in a second step, express the standard deviation itself as an uncertain parameter. The distribution of the uncertain standard deviation is either specified directly using the available information, or is itself based on e.g. the Principle of Maximum Information Entropy if the available information only allows us to impose constraints for the standard deviation.

Example 6.8. *Hierarchic prior distributions*:
We continue and extend the simple toy-example presented in Example 6.7, in which the knowledge about parameter θ is updated. Again, we assume that a-priori the (conditional) mean m of θ is known. However, contrary to Example 6.7, for the standard deviation s of θ we presume that only incomplete knowledge is available: The standard deviation s is treated as a hyperparameter with a coefficient of variation of $\delta_s = 30\%$. Consequently, the prior can be expressed as $p(\theta, s) = p(\theta|s) \cdot p(s)$. The prior distribution of $p(\theta|s)$ is considered fixed: It is a log-Normal distribution that has conditional mean m and standard deviation s. For the distribution $p(s)$

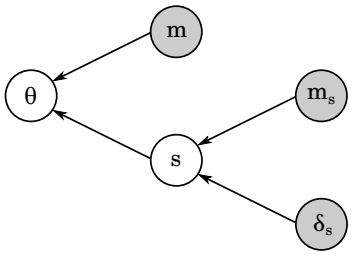


Figure 6.6: Hierarchic prior probabilistic model used in Example 6.8 – represented as a Bayesian network. The white nodes denote the parameters that are considered uncertain, the gray nodes are the parameters whose value is fixed. (Example 6.8)

Table 6.4: Results of the simple toy-example discussed as Example 6.8 in Section 6.4.5: Comparison of three hyper-prior models for the standard deviation s of parameter θ. The value in each row that is considered advantageous (conservative) compared to the other entries is highlighted in blue. The value in each row that is considered disadvantageous compared to the other entries is highlighted in orange. The last three columns that are printed in grey list the results of the Bayesian analysis for fixed values of s; the underlying models do not maintain the a-priori prescribed constraints for s and are added for illustratory purposes only. The column $s = 0.10$ repeats the results of Example 6.7 listed in Table 6.3 for the log-Normal prior. A value of $s = 0.40$ maximizes the evidence amongst all feasible models with fixed $s \in (0, \infty)$; and $s = 0.29$ maximizes the evidence weighted with the PDF of the log-Normal hyper-prior. (Example 6.8)

	MEP hyper-prior	log-Normal hyper-prior	Gamma hyper-prior	fixed s $s = 0.10$	fixed s $s = 0.40$	fixed s $s = 0.29$
mean of marginal prior θ	1.000	1.000	1.000	1.000	1.000	1.000
std. dev. of marginal prior θ	0.100	0.100	0.100	0.1	0.4	0.29
entropy of marginal prior θ	0.182	0.180	0.181	−0.891	0.391	0.120
entropy of prior	−0.88	−0.93	−0.90	−0.891	0.391	0.120
entropy of posterior	−2.03	−1.99	−1.99	−1.07	−0.890	−0.899
entropy of marginal posterior θ	−0.884	−0.886	−0.885	−1.07	−0.890	−0.899
evidence	0.268	0.270	0.269	0.010	0.329	0.295
relative entropy	2.12	2.11	2.11	3.53	1.96	2.03
data fit	0.802	0.798	0.799	−1.11	0.852	0.810
mean of posterior θ	1.46	1.46	1.46	1.29	1.47	1.46
std. dev. of posterior θ	0.100	0.100	0.100	0.083	0.099	0.099
mean of posterior s	0.326	0.320	0.322	0.1	0.4	0.29
std. dev. of posterior s	0.078	0.087	0.084	–	–	–

that has mean m_s and coefficient of variation δ_s, different alternative hyper-prior models are compared: the MEP hyper-prior (a truncated Normal distribution), a log-Normal hyper-prior and a Gamma hyper-prior. The probabilistic model of the prior is illustrated in Fig. 6.6.

The likelihood function is adopted without modification from Example 6.7: It is a Normal distribution with mean 1.5 and standard deviation 0.1. To maintain comparability to Example 6.7, the conditional mean m of θ as well as the mean m_s of s are selected such that the mean and standard deviation of the marginal prior of θ are 1.0 and 0.1 – in accordance with Example 6.7. This results in $m = 1.0001$ and $m_s = 0.3031$.

The results are listed in Table 6.4. The model employing the MEP hyper-prior has the prior

and marginal prior of θ with the largest entropy. Also the entropy of the posterior is maximized for the MEP hyper-prior compared to the other investigated hyper-prior models. However, the same cannot be said about the entropy of the marginal posterior of θ. The largest evidence is produced by the model class employing the log-Normal hyper-prior. However, there is no considerable difference in the evidence of all three investigated hyper-prior models. The model class with the MEP hyper-prior has the best data-fit, but also a larger relative entropy compared to the other hyper-prior models. All three hyper-prior models have the same posterior mean and standard deviation for parameter θ.

Comparing the models investigated in this example (Example 6.8) with the models investigated in Example 6.7, the models in this example clearly have a larger evidence. Note that the constraints on θ are the same in Example 6.7 and Example 6.8. Therefore, the models in Example 6.8 can be considered more appropriate than the models in Example 6.7. The larger evidence is a consequence of (i) a larger data-fit of the models in Example 6.8, but also of a (ii) smaller *relative entropy* of the models in Example 6.8. Remember that *relative entropy* can be interpreted as a measure for model complexity – a larger relative entropy acts as a penalty term for more complex models. Therefore, at a first glance, it is surprising that the models in Example 6.8 have a smaller relative entropy than the models in Example 6.7, because the models in Example 6.8 employ a larger number of parameters (i.e., *two* parameters) than the models in Example 6.7 (which have only *one* parameter). However, in this context, model complexity cannot be interpreted as a simple function of the number of parameters in a stochastic model class: The prior uncertainty that is associated with the respective model parameters has a considerable influence.

The last three columns in Table 6.4 investigate a partial hierarchic modeling approach. Instead of describing the uncertainty about parameter s explicitly, the value of s is fixed at a particular $\hat{s}$. This columns are added for illustratory purposes only, because they do not maintain the a-priori specified constraints for s. The column $\hat{s} = 0.10$ repeats the results listed in Table 6.3 for the log-Normal prior (Example 6.7). Remember that the three hyper-prior models can be directly compared to the models listed in Table 6.3, because all models maintain the prior constraints imposed on θ. The value $\hat{s} = 0.40$ is selected such that the evidence of the underlying non-hierarchic stochastic model class is maximized (compared to all other feasible values for s). In this case, no prior information about s is taken into account – and the results deviate clearly from the results obtained employing a full hierarchic modeling approach. Results that are closer to the hierarchic approach can be obtained with $\hat{s} = 0.29$, which maximizes the evidence weighted with the PDF of the log-Normal hyper-prior at location s. This approach takes prior information about s into account, but selects only the model class belonging to a particular s that appears most plausible.

As Example 6.7, Example 6.8 demonstrates that the MEP prior is a feasible choice, but should be considered as one viable prior model out of many. Wherever feasible, a Bayesian analysis should take different alternative prior modeling strategies into account. The evidence of the associated stochastic model classes can be used to obtain posterior plausibilities of the investigated model classes.

6.4.6 Probability models for error structures

This section is manly written with respect to *output prediction-errors*, as they are usually the largest errors in the analysis. However, the following discussion can be directly transfered to other error structures; e.g., *observation errors* or *input errors*.

6.4.6.1 Quantities that influence the output prediction-error

The probability model of the *output prediction-error* $\mathbf{v}$ should – in theory – depend on $\mathbf{q}$, $\boldsymbol{\theta}_\mathbf{q}$ and $\mathbf{f}$. This dependency can be illustrated by the example of a frame structure that is modeled using linear-elastic beam elements: In this case $\mathbf{f}$ represents the loading of the structure, $\boldsymbol{\theta}_\mathbf{q}$ represents its stiffness, and $\mathbf{q}$ are the computed displacements; the displacements of the real structure are $\mathbf{r}$. If the loading $\mathbf{f}$ of the structure is large compared to the stiffness $\boldsymbol{\theta}_\mathbf{q}$ of the structure, the material behaves no longer elastic and the response $\mathbf{r}$ of the real structure becomes non-linear. Similarly, if the displacements $\mathbf{q}$ of the model become large, it is to be expected that geometrical non-linearities have to be considered to represent the response $\mathbf{r}$ of the real structure. As the model at hand is a linear-elastic one and is based on the assumption of small displacements, the modeling errors clearly depend on $\mathbf{q}$, $\boldsymbol{\theta}_\mathbf{q}$ and $\mathbf{f}$.

In practice, this dependency is typically not modeled explicitly (see Fig. 6.4). Sometimes, the dependency is implicitly assumed by choosing either an additive or a multiplicative error structure for $\mathbf{v}$. If the investigated system is always exposed to similar input conditions, the stochastic model class can very likely cope with the inadequate assumptions about the error structure. However, if the investigated system is exposed to different (sometimes possibly extreme) input conditions, the inadequate[1] error structure will tend to underestimate the uncertainties in the model[2]. Such an undesired behavior clearly confines the prediction capabilities of the selected stochastic model class.

As one of the purposes of Bayesian inference is often to predict the response of a system under extreme conditions, neglecting the dependency of the *output prediction-error* $\mathbf{v}$ on $\mathbf{q}$, $\boldsymbol{\theta}_\mathbf{q}$ and $\mathbf{f}$ can be a considerable source for errors. For many models a reasonable (and still feasible) strategy can be to formulate the standard deviation of $\mathbf{v}$ as a function of the model response $\mathbf{q}$. For example, in case of the previously mentioned linear model, the standard deviation of $\mathbf{v}$ can be gradually increased if $\mathbf{q}$ exceeds a specified threshold. In this example, a bias in the model output for large $\mathbf{q}$ can be implemented as well, by setting the mean of $\mathbf{v}$ to a value different from *zero*, in order to approximately represents non-linear effects for large displacements.

[1] By "inadequate", an error structure is meant that underestimates the errors on average.

[2] The uncertainties tend to be underestimated, because inadequate assumptions mean that the available data is less likely to have been observed; i.e., is more surprising. This means that the *information gain* of the posterior relative to the prior is large. Thus, the posterior tends to be over-confident.

6.4.6.2 Representation of the error mean

Apart from representing *prediction-errors* for "extreme" scenarios that the model was not designed for (see previous section), it is typically appropriate to assume that the *prediction-error* and *observation-error* are unbiased – even if in reality there will probably be a bias. This assumptions needs to be made, because modeling a biased error structure is equivalent to increasing the complexity of the underlying model. If information is available that indicates a (possibly *a priori* uncertain) biased error structure, we could, therefore, parametrize the expected bias and consider it as part of the engineering model. For an additive error structure, assuming unbiasedness corresponds to setting the mean to *zero*. For a multiplicative error structure, the mean has to be set to *one* to get an unbiased response (if the error is considered independent of the response).

6.4.6.3 Representation of the error variance

It is often also safe to assume that the variance of the error structure is - at least conditionally within a hierarchic framework - known (see discussion in Section 6.4.6.1). Assuming the variance conditionally known has the advantage that (i) the uncertainty in the variance can be quantified and modeled separately (e.g., conditional on the model response) of the actual error structure, and that (ii) the complexity of the probabilistic error model decreases by assuming the variance conditionally fixed (see Section 6.4.5). If there is reason to believe the variance structure to be inhomogeneous, one could attempt to parametrize (and learn) the variance in the context of the hierarchic model structure.

Another issue to consider is whether the size of errors and the shape of the error distribution are bounded by physical constraints. This is for example the case if observed or predicted quantities cannot become negative. In such a case it may be more convenient to work with a multiplicative instead of an additive error structure. However, working with a multiplicative error structure, it is important to ensure that for model responses close to *zero*, the expected errors of the model are not underestimated (as the standard deviation is assumed proportional to the response).

6.4.6.4 Dependence structure of the errors

If the mean and variance of the error structure are conditionally asserted and if the size of errors is not bound by physical constraints, the maximum entropy PDF is Normal in the absence of any other limiting constraints. The handling of uncertainties in the correlation structure of (e.g.) a Normal process is more complex than handling uncertainties in the variance or in the mean. Typically, no specific information about the particular correlation structure is available. Therefore, the errors are usually considered to be uncorrelated, in

accordance with the *Principle of Maximum Information Entropy*.

However, based on the physics of the problem, we often know that especially *prediction-errors* tend to exhibit a positive dependence structure in time and/or space. Neglecting this knowledge and modeling the errors as uncorrelated does not necessarily lead to more conservative results [Simoen et al., 2013]. Instead of working with uncorrelated errors, it is therefore recommended to explicitly model the correlation structure as uncertain [Simoen et al., 2013].

This holds in particular for model responses that exhibit a temporal or spatial structure. The errors of such a type of model response are typically dependent, where the dependency decreases with increasing time/distance between two elements. Usually, in engineering, the time-step size is not large enough to consider the *output prediction-errors* of two adjacent time-steps (or grid elements) as uncorrelated. The modeling issue is often the unknown dependence structure of the errors. However, especially for long observation histories, the dependence structure can have a considerable influence on the posterior results. Therefore, it is advisable to work with different dependence structures specified in separate modeling classes, and to consider the dependence parameter (e.g., the correlation length) as uncertain – provided that such an approach is computationally feasible. Dependence structures are commonly expressed in terms of the exponential or exponential squared auto-correlation coefficient function, or as a auto-regressive or moving-average model. Note that the process modeled with an exponential auto-correlation coefficient function is equivalent to a first-order auto-regressive model. This type of dependence structure is often preferred in practice, because it renders the process Markovian.

Example 6.9. *Correlated error structure*:
Let the observed data $\mathcal{D}$ be generated by a standard Normal stochastic process with exponential correlation structure; i.e. the auto-correlation coefficient function of the stochastic process is $\rho_{\mathcal{D}}(d_i, d_j) = \exp\left(\frac{|i-j|}{l_{\text{data}}}\right)$, where $l_{\text{data}} > 0$ is the correlation length, and $i, j \in \{1, \ldots, N_{\text{data}}\}$, with N_{data} the number of observed data points. In the following, it is assumed that the properties of the process generating the data $\mathcal{D}$ are not fully known.

The observed data $\mathcal{D}$ is approximated by a Normal stochastic process that has *zero* mean, standard deviation σ, and exponential correlation structure with correlation length l. Our knowledge about σ and l is considered uncertain. Consequently, the likelihood function $L(\sigma, l|\mathcal{D})$ can be expressed as the PDF of a N-dimensional multivariate Normal distribution that has *zero* mean and covariance matrix $\mathbf{\Sigma}$, where the coefficients $(\mathbf{\Sigma})_{i,j}$ of $\mathbf{\Sigma}$ are defined as $(\mathbf{\Sigma})_{i,j} = \sigma^2 \cdot \exp\left(\frac{|i-j|}{l}\right)$.

In this example, the shape of the *average likelihood function* is assessed. The *average likelihood function* is computed as follows: 10^3 different realizations of the data set $\mathcal{D}$ are generated, the likelihood function of each data set is evaluated, and the average over all such likelihood functions is labeled the *average likelihood function*. The *average likelihood function* is depicted

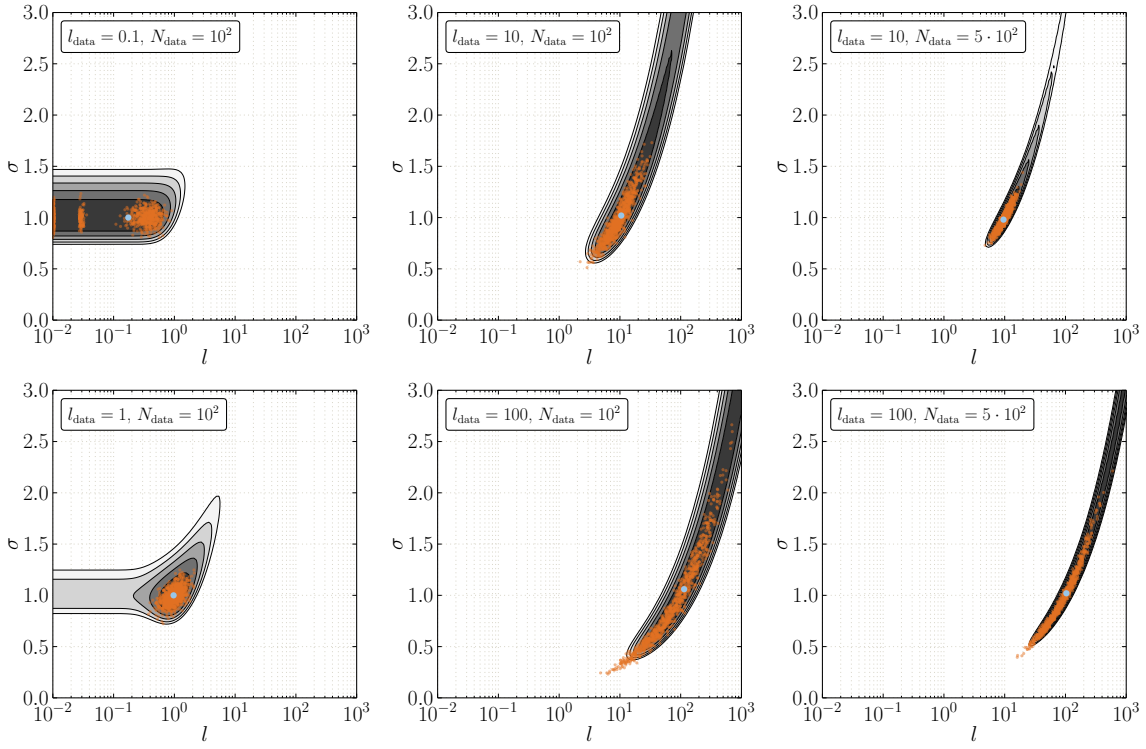


Figure 6.7: The contour lines depict the shape of the *average likelihood function*; i.e., the average of the likelihood functions of 10^3 randomly generated data sets. Each contour line indicates a decrease of the likelihood by a factor of 10 compared to the preceding contour line (or to the absolute maximum for the first contour line). The blue marker indicates the maximum of the *average likelihood function*. The orange markers indicate the maxima of the 10^3 likelihood functions that were used to obtain the *average likelihood function*. The plots differ in the properties of the stochastic process generating the data. Different correlation length l_{data} and size N_{data} of the data sets are investigated. (Example 6.9)

in Fig. 6.7 for different l_{data} and N_{data}.

The shapes of the depicted likelihood functions indicate that for large correlation length l_{data} used to generate the data sets, the assumed uncertain standard deviation σ and the correlation length l cannot be reliably estimated from the data – even if N_{data} is relatively large. Thus, the prior distribution of σ and l will have a considerable influence on the posterior distribution, and the posterior distribution of σ and l is not guaranteed to be centered in the neighborhood of the postulated *true* values σ_{data} and l_{data}. Moreover, assuming the prior standard deviation of the errors conservatively (i.e., too large), will limit the prediction capabilities of the model, as both the posterior standard deviation and the posterior correlation length of the errors are over-estimated.

6.5 Formulation of the likelihood function

The evidence of the stochastic model class is defined as the plausibility of observing output $\mathbf{z}$ conditional on having observed input $\mathbf{s}$ and on the stochastic model class $\mathcal{M}$, i.e., $c_{\mathbf{E}|\mathcal{M}} =$

$p(\mathbf{z}|\mathbf{s},\mathcal{M})$. The evidence can be evaluated as:

$$c_{\mathtt{E}|\mathcal{M}} = \int_{\boldsymbol{\theta}} \int_{\mathbf{f}} \int_{\mathbf{r}} p(\mathbf{z}|\mathbf{r},\boldsymbol{\theta}_{\mathtt{w}},\mathcal{M}) \cdot p(\mathbf{r}|\mathbf{q}(\mathbf{f},\boldsymbol{\theta}_{\mathtt{q}}),\boldsymbol{\theta}_{\mathtt{v}},\mathcal{M}) \cdot p(\mathbf{f}|\mathbf{s},\boldsymbol{\theta}_{\mathtt{f}},\mathcal{M}) \cdot p(\boldsymbol{\theta}|\mathcal{M}) \,\mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{f}\,\mathrm{d}\boldsymbol{\theta} \;, \quad (6.31)$$

where $\boldsymbol{\theta}_{\mathtt{q}}$, $\boldsymbol{\theta}_{\mathtt{f}}$, $\boldsymbol{\theta}_{\mathtt{v}}$ and $\boldsymbol{\theta}_{\mathtt{w}}$ are part of $\boldsymbol{\theta}$.

Contrary to the evidence that is a unique property of the selected stochastic model class $\mathcal{M}$ and the observed data $\mathcal{D}$, the formulation of the Bayesian inference process itself is ambiguous. Some variants to formulate the Bayesian inference process are stated in the following.

Variant (1)
The first variant merges the *prediction-error* and the *observation-error*. The likelihood is expressed as the plausibility of observing $\mathbf{z}$ conditional on the model output $\mathbf{q}(\mathbf{f},\boldsymbol{\theta}_{\mathtt{q}})$:

$$p(\mathbf{z}|\mathbf{q}(\mathbf{f},\boldsymbol{\theta}_{\mathtt{q}}),\boldsymbol{\theta},\mathcal{M}) = \int_{\mathbf{r}} p(\mathbf{z}|\mathbf{r},\boldsymbol{\theta}_{\mathtt{w}},\mathcal{M}) \cdot p(\mathbf{r}|\mathbf{q}(\mathbf{f},\boldsymbol{\theta}_{\mathtt{q}}),\boldsymbol{\theta}_{\mathtt{v}},\mathcal{M}) \,\mathrm{d}\mathbf{r} \quad (6.32)$$

In case of an additive error structure we have $\mathbf{z} = \mathbf{q} + \mathbf{v} + \mathbf{w}$, and, thus, we can write the likelihood as:

$$p(\mathbf{z} = \hat{\mathbf{z}}|\mathbf{q}(\mathbf{f},\boldsymbol{\theta}_{\mathtt{q}}),\boldsymbol{\theta},\mathcal{M}) = p(\mathbf{v}+\mathbf{w} = \hat{\mathbf{z}} - \mathbf{q}(\mathbf{f},\boldsymbol{\theta}_{\mathtt{q}})|\boldsymbol{\theta}_{\mathtt{v}},\boldsymbol{\theta}_{\mathtt{w}},\mathcal{M}) \quad (6.33)$$

The convolution integral appearing in Eq. (6.32), and implicitly also in Eq. (6.33), is an inconvenience. If both $\mathbf{v}$ and $\mathbf{w}$ are Normal, the integral can be solved analytically, because $\mathbf{v}+\mathbf{w}$ is also Normal. With likelihood $p(\mathbf{z}|\mathbf{q}(\mathbf{f},\boldsymbol{\theta}_{\mathtt{q}}),\boldsymbol{\theta},\mathcal{M})$ and prior $p(\mathbf{f}|\mathbf{s},\boldsymbol{\theta}_{\mathtt{f}},\mathcal{M}) \cdot p(\boldsymbol{\theta}|\mathcal{M})$, the Bayesian inference process can be written as:

$$p(\mathbf{f},\boldsymbol{\theta}|\mathbf{z},\mathbf{s},\mathcal{M}) = {c_{\mathtt{E}|\mathcal{M}}}^{-1} \cdot p(\mathbf{z}|\mathbf{q}(\mathbf{f},\boldsymbol{\theta}_{\mathtt{q}}),\boldsymbol{\theta}_{\mathtt{v}},\mathcal{M}) \cdot p(\mathbf{f}|\mathbf{s},\boldsymbol{\theta}_{\mathtt{f}},\mathcal{M}) \cdot p(\boldsymbol{\theta}|\mathcal{M}) \quad (6.34)$$

Consequently, we directly learn the posterior parameter vector $\boldsymbol{\theta}$ and the posterior model input $\mathbf{f}$. In order to obtain the joint posterior PDF that additionally includes the system output $\mathbf{r} = \mathbf{q}(\mathbf{f},\boldsymbol{\theta}_{\mathtt{q}}) + \mathbf{v}$, a post-processing step is required:

$$p(\mathbf{v},\mathbf{f},\boldsymbol{\theta}|\mathbf{z},\mathbf{s},\mathcal{M}) = p(\mathbf{v}|\mathbf{v}+\mathbf{w} = \mathbf{z} - \mathbf{q}(\mathbf{f},\boldsymbol{\theta}_{\mathtt{q}}),\mathbf{z},\mathbf{q}(\mathbf{f},\boldsymbol{\theta}_{\mathtt{q}}),\boldsymbol{\theta},\mathcal{M}) \cdot p(\mathbf{f},\boldsymbol{\theta}|\mathbf{z},\mathbf{s},\mathcal{M}) \quad (6.35)$$

where $p(\mathbf{v}|\mathbf{v}+\mathbf{w} = \mathbf{z} - \mathbf{q}(\mathbf{f},\boldsymbol{\theta}_{\mathtt{q}}),\mathbf{z},\mathbf{q}(\mathbf{f},\boldsymbol{\theta}_{\mathtt{q}}),\boldsymbol{\theta},\mathcal{M})$ can be readily obtained as a conditional Normal distribution, if both $\mathbf{v}$ and $\mathbf{w}$ are Normal.

Variant (2a)

This variant expresses the likelihood directly in terms of the *observation-error*: $p(\mathbf{z}|\mathbf{r}, \boldsymbol{\theta}_{\mathtt{w}}, \mathcal{M})$, and the stochastic forward model $\mathbf{r} = \mathbf{q}(\mathbf{f}, \boldsymbol{\theta}_{\mathtt{q}}) + \mathbf{v}$ is represented explicitly as part of the model. For an additive error structure $\mathbf{z} = \mathbf{r} + \mathbf{w}$, the likelihood is $p(\mathbf{z} = \hat{\mathbf{z}}|\mathbf{r}, \boldsymbol{\theta}_{\mathtt{w}}, \mathcal{M}) = p(\mathbf{w} = \hat{\mathbf{z}} - \mathbf{r}|\boldsymbol{\theta}_{\mathtt{w}}, \mathcal{M})$. The Bayesian inference process is formulated as:

$$p(\mathbf{r}, \mathbf{f}, \boldsymbol{\theta}|\mathbf{z}, \mathbf{s}, \mathcal{M}) = {c_{\mathtt{E}|\mathcal{M}}}^{-1} \cdot p(\mathbf{z}|\mathbf{r}, \boldsymbol{\theta}_{\mathtt{w}}, \mathcal{M}) \cdot p(\mathbf{r}|\mathbf{q}(\mathbf{f}, \boldsymbol{\theta}_{\mathtt{q}}), \boldsymbol{\theta}_{\mathtt{v}}, \mathcal{M}) \cdot p(\mathbf{f}|\mathbf{s}, \boldsymbol{\theta}_{\mathtt{f}}, \mathcal{M}) \cdot p(\boldsymbol{\theta}|\mathcal{M}) \quad (6.36)$$

Contrary to *Variant (1)*, we directly learn the joint posterior PDF of $\mathbf{r}$, $\mathbf{f}$, and $\boldsymbol{\theta}$. Consequently, posterior samples of the marginal PDF $p(\mathbf{r}|\mathbf{z}, \mathbf{s}, \mathcal{M})$ are readily available, corresponding to samples from the robust posterior predictive PDF for the system output history conditioned on the observed input. Therefore, this variant directly provides a filter not only for parameter estimation, but also for the system output. However, this advantage has a price: The peakedness of the likelihood function is more pronounced than in *Variant (1)*, because, the uncertainty in the *observation-errors* is typically much smaller than the uncertainty in the *prediction-errors*.

Variant (2b)

Instead of expressing the likelihood in terms of the *observation-error*, the likelihood can often also be stated exclusively in terms of the *prediction-error*. In this case, the observation-error $\mathbf{w}$ is represented explicitly, and we introduce the stochastic variable $\hat{\mathbf{r}}$ that is linked to the observed output $\hat{\mathbf{z}}$ through $\mathbf{w}$ as $\hat{\mathbf{r}} = \hat{\mathbf{z}} - \mathbf{w}$. The likelihood is then given as the plausibility of observing $\mathbf{r} = \hat{\mathbf{r}}$ given $\mathbf{w}$, $\mathbf{q}(\mathbf{f}, \boldsymbol{\theta}_{\mathtt{q}})$ and $\mathcal{M}$ i.e., $p(\mathbf{r} = \hat{\mathbf{r}}|\hat{\mathbf{z}}, \mathbf{w}, \mathbf{q}(\mathbf{f}, \boldsymbol{\theta}_{\mathtt{q}}), \boldsymbol{\theta}_{\mathtt{v}}, \mathcal{M}) = p(\mathbf{v} = \hat{\mathbf{z}} - \mathbf{w} - \mathbf{q}(\mathbf{f}, \boldsymbol{\theta}_{\mathtt{q}})|\boldsymbol{\theta}_{\mathtt{v}}, \mathcal{M})$. This leads to the following Bayesian inference process:

$$p(\mathbf{w}, \mathbf{f}, \boldsymbol{\theta}|\hat{\mathbf{z}}, \mathbf{s}, \mathcal{M}) = {c_{\mathtt{E}|\mathcal{M}}}^{-1} \cdot p(\mathbf{r} = \hat{\mathbf{r}}|\hat{\mathbf{z}}, \mathbf{w}, \mathbf{q}(\mathbf{f}, \boldsymbol{\theta}_{\mathtt{q}}), \boldsymbol{\theta}_{\mathtt{v}}, \mathcal{M}) \cdot p(\mathbf{w}|, \boldsymbol{\theta}_{\mathtt{w}}, \mathcal{M}) \cdot p(\mathbf{f}|\mathbf{s}, \boldsymbol{\theta}_{\mathtt{f}}, \mathcal{M}) \cdot p(\boldsymbol{\theta}|\mathcal{M}) \quad (6.37)$$

In this variant we explicitly learn the structure of the posterior *observation-error*. We can obtain posterior realizations of the system output $\mathbf{r}$ by using the relation $\mathbf{r} = \hat{\mathbf{z}} - \mathbf{w}$ and inserting posterior samples of the observation-error $\mathbf{w}$.

Variant (3)

In the third variant all errors are modeled explicitly and the likelihood is given as $p(\mathbf{z} = \hat{\mathbf{z}}|\mathbf{w}, \mathbf{v}, \mathbf{q}(\mathbf{f}, \boldsymbol{\theta}_{\mathtt{q}}), \mathcal{M}) = \delta_{\mathbf{z}}(\hat{\mathbf{z}})$, where $\delta_{\mathbf{z}}(\cdot)$ denotes the Dirac mass at $\mathbf{z} = \mathbf{q}(\mathbf{f}, \boldsymbol{\theta}_{\mathtt{q}}) + \mathbf{v} + \mathbf{w}$.

The Bayesian inference process then writes:

$$p(\mathbf{w},\mathbf{v},\mathbf{f},\boldsymbol{\theta}|\hat{\mathbf{z}},\mathbf{s},\mathcal{M}) = {c_{\mathbf{E}|\mathcal{M}}}^{-1} \cdot p(\mathbf{q}(\mathbf{f},\boldsymbol{\theta}_\mathtt{q}) + \mathbf{v} + \mathbf{w} = \hat{\mathbf{z}}|\mathbf{w},\mathbf{v},\mathbf{q}(\mathbf{f},\boldsymbol{\theta}_\mathtt{q}),\mathcal{M}) \cdot p(\mathbf{w}|,\boldsymbol{\theta}_\mathtt{w},\mathcal{M}) \cdot p(\mathbf{v}|,\boldsymbol{\theta}_\mathtt{v},\mathcal{M}) \cdot p(\mathbf{f}|\mathbf{s},\boldsymbol{\theta}_\mathtt{f},\mathcal{M}) \cdot p(\boldsymbol{\theta}|\mathcal{M}) \quad (6.38)$$

Eq. (6.38) can usually not be solved analytically; and a direct numerical treatment of Eq. (6.38) is infeasible, if the space of $\mathbf{z}$ is not finite. The problem in Eq. (6.38) can be tackled approximately by means of *Approximate Bayesian Computation* [Tavaré et al., 1997; Beaumont et al., 2009].

The posterior distributions as well as the value of the evidence are invariant to how the likelihood in the Bayesian inference process is formulated. Nevertheless, if the Bayesian inference is performed numerically, it does matter which of the above variants is used to set-up the problem. Loosely speaking, for most Markov Chain Mote Carlo (MCMC) [Gilks et al., 1996; Gelman et al., 2004a] based algorithms, the computational complexity of generating posterior samples increases as the peakedness of the likelihood function increases. This holds in particular for the BUS approach [Straub and Papaioannou, 2015] and the TMCMC method [Ching and Chen, 2007]. Ordering the above variants with respect to an increasing peakedness we get the sequence: *Variant (1)*, *Variant (2b)*, *Variant (2a)*, *Variant (3)*, where *Variant (1)* has the flattest likelihood. *Variant (2b)* is listed before *Variant (2a)*, because typically the uncertainty in the *prediction-error* is larger than the uncertainty in the *observation-error*. The formulation given in *Variant (1)* is to be preferred if the Bayesian inference process is performed numerically, since it has the flattest likelihood. However, if the convolution integral in Eq. (6.32) cannot be solved analytically, it is recommended to use *Variant (2b)*. *Variant (2a)* is only recommended if the uncertainty in the *observation-error* is larger than the uncertainty in the *prediction-error*. *Variant (3)* constitutes an extreme case, that can only be tackled numerically through *Approximate Bayesian Computation*, since it includes the Dirac function.

Chapter 7

Numerical Methods for Bayesian Analysis

7.1 Introduction

In a Bayesian analysis, typically two main problems can be identified:

1. *generate posterior samples*:
   Except for very simple models, the posterior distribution can usually not be derived analytically. Even if the shape of the posterior distribution is known explicitly, it does often not follow any conventional type. Consequently, samples from the posterior can in most cases only be generated numerically.
   If the underlying model requires low computational costs, Markov chain Monte Carlo (MCMC) methods (see Section 3.4) constitute a popular class of methods to sample from the posterior distribution [Gilks et al., 1996; Gelman et al., 2004a]. However, one problem of MCMC methods is that after an initial burn-in phase the samples may not yet have reached the stationary distribution of the Markov chain [Plummer et al., 2006]. That is, finding an appropriate burn-in period in MCMC is often a non-trivial problem. Another issue is that standard MCMC algorithms usually cannot be applied efficiently for problems with many uncertain parameters. Some specialized MCMC algorithms [Haario et al., 2005; Robert and Tweedie, 1996; Neal, 2011; Cheung and Beck, 2009] can cope with such high dimensional problems, they require however additional evaluations of the likelihood function or its gradient for each generated sample.

2. *evaluate the evidence* $c_{\mathrm{E}|\mathcal{M}}$:
   It is typically challenging to compute the evidence $c_{\mathrm{E}|\mathcal{M}}$, because of the multi-dimensional integral in Eq. (6.2). If the system is globally identifiable [Beck and Katafygiotis, 1991, 1998], asymptotic approximations [Beck and Yuen, 2004; MacKay, 1992b] can be ap-

plied to estimate the evidence. Otherwise, the evidence is usually evaluated numerically. Some methods to compute the evidence $c_{\mathbf{E}|\mathcal{M}}$ are discussed in [Cheung and Beck, 2010].

In Bayesian inference with engineering problems, the likelihood function is typically coupled to an engineering model that approximates the response of the actual system of interest. For each evaluation of the likelihood function, the engineering model needs to be computed for the inquired parameter vector $\boldsymbol{\theta}$. As many engineering models are computationally demanding, the efficiency of Bayesian inference methods is often associated with the required number of likelihood function calls.

This chapter focuses on numerical methods that generate samples from the posterior distribution and provide simultaneously an estimate for the evidence. The BUS appraoch (Section 7.3), an adpative variant of the BUS approach (Section 7.4), nested sampling (Section 7.5) and the TMCMC method (Section 7.6) are presented.

7.2 Investigated example problems for numerical Bayesian inference

The example problems listed in the following are investigated in this chapter. This section uses quantity c that is a positive constant that typically needs to be chosen such that $c \cdot L(\boldsymbol{\theta}|\mathcal{D}) \leq 1$ is maintained for all $\boldsymbol{\theta}$. The constant c is properly introduced in Section 7.3.

- *Example problem 1a:* A one dimensional problem with a standard Normal prior. The uncertain parameter is denoted by θ. The likelihood is a Normal distribution that has mean $\mu_l = 3$ and standard deviation $\sigma_l = 0.3$. This problem has an analytical solution: The posterior distribution is Normal with posterior mean and standard deviation of $\mu_l / \left(\sigma_l^2 + 1\right) = 2.75$ and $1/\sqrt{1 + \sigma_l^{-2}} = 0.287$, respectively. The maximum of the likelihood is $L_{\max} = 1/\left(\sigma_l \sqrt{2\pi}\right) = 1.33$. The evidence of the example problem is $c_{\mathbf{E},\text{ref}} = \varphi\left(\mu_l / \sqrt{1 + \sigma_l^2}\right) / \sqrt{1 + \sigma_l^2} = 6.16 \cdot 10^{-3}$, where $\varphi(\cdot)$ is the PDF of the standard Normal distribution. Consequently, $p_{\Omega,\text{ref}}$ of the rejection sampling algorithm is $4.63 \cdot 10^{-3}$, if $c = 1/L_{\max}$.

- *Example problem 1b:* The formulation of this problem is equivalent to *Example problem 1a*, with the only difference being that the likelihood function has mean $\mu_l = 5$ and standard deviation $\sigma_l = 0.2$. The posterior mean and standard deviation is 4.81 and 0.196, respectively. The evidence for this problem is $c_{\mathbf{E},\text{ref}} = 2.36 \cdot 10^{-6}$. Consequently, $p_{\Omega,\text{ref}}$ of the rejection sampling algorithm is $1.18 \cdot 10^{-6}$, if $c = 1/L_{\max}$ and $L_{\max} = 1.99$.

- *Example problem 2:* A 12-dimensional problem with prior $\prod_{i=1}^{12} \varphi(\theta_i)$, where $\varphi(\cdot)$ denotes the standard Normal PDF and θ_i is the ith component of the 12-dimensional

parameter vector $\boldsymbol{\theta}$. The likelihood function of the problem is $\prod_{i=1}^{12} \varphi\left(\frac{\theta_i-\mu_l}{\sigma_l}\right)/\sigma_l$, with $\sigma_l = 0.6$. The value μ_l is chosen such that the evidence $c_{\mathrm{E,ref}}$ becomes 10^{-6}; i.e., $\mu_l = 0.462$. The posterior mean and standard deviation of each component of $\boldsymbol{\theta}$ is 0.34 and 0.51, respectively. The theoretical maximum that the likelihood function can take is $L_{\max} = \left(0.6 \cdot \sqrt{2\pi}\right)^{-12} = 7.47 \cdot 10^{-3}$. Thus, $p_{\Omega,\mathrm{ref}}$ of the rejection sampling algorithm is $1.34 \cdot 10^{-4}$, if $c = 1/L_{\max}$.

– *Example problem 3:* A two-story frame structure represented as a two-degree-of-freedom shear building model is investigated. This example problem was originally discussed in [Beck and Au, 2002]. BUS is applied in [Straub and Papaioannou, 2015; DiazDelaO et al., 2017] to solve this problem. The two stiffness coefficients k_1 (first story) and k_2 (second story) of the model are considered uncertain. The uncertainty in k_1 and k_2 is expressed as $k_1 = \theta_1 \cdot k_{\mathrm{n}}$ and $k_2 = \theta_2 \cdot k_{\mathrm{n}}$, where θ_1 and θ_2 are uncertain parameters and $k_{\mathrm{n}} = 29.7 \cdot 10^6 \mathrm{N/m}$. The prior distributions of θ_1 and θ_2 are modeled as independent log-Normal distributions with modes 1.3 and 0.8 and standard deviation 1.0. The lumped story masses m_1 (first story) and m_2 (second story) are considered deterministic and have masses $m_1 = 16.5 \cdot 10^3 \mathrm{kg}$ and $m_2 = 16.1 \cdot 10^3 \mathrm{kg}$. The influence of damping is neglected. Bayesian updating is performed based on the measured first two eigen-frequencies of the system: $\tilde{f}_1 = 3.13\mathrm{Hz}$ and $\tilde{f}_2 = 9.83\mathrm{Hz}$. The likelihood of the problem is expressed as $L(\boldsymbol{\theta}) = \exp\left(-0.5 \cdot J(\boldsymbol{\theta})/\sigma_\varepsilon^2\right)$, where $\sigma_\varepsilon = 1/16$ and $J(\boldsymbol{\theta}) = \sum_{j=1}^{2} \lambda_j^2 \left(\frac{f_j^2(\boldsymbol{\theta})}{\tilde{f}_j^2} - 1\right)^2$ with $\lambda_1 = \lambda_2 = 1$ and $f_j(\boldsymbol{\theta})$ as the jth eigen-frequency predicted by the model. The posterior distribution of this problem is bimodal [Beck and Au, 2002; Straub and Papaioannou, 2015]. The reference solution is: $c_{\mathrm{E,ref}} = p_{\Omega,\mathrm{ref}} = 1.52 \cdot 10^{-3}$ (since $L_{\max} = 1$), $\mathrm{E}[k_1|\mathcal{D}] = 1.12$ and $\sigma[k_1|\mathcal{D}] = 0.66$.

The following quantities are introduced for the discussion of the example problems:

– b acts as a normalized version of c^{-1}: Let $b \in (0, 1]$ be defined as $b = L_{\max}/c^{-1}$; i.e., $b = 1 \Leftrightarrow c^{-1} = L_{\max}$ and $b = 0 \Leftrightarrow c^{-1} = 0$. The performance of the investigated algorithm is assessed for different values of b.

– $b_{10^3,\max}$ represents the largest observed likelihood multiplied with c in a set of 10^3 independent posterior samples. Note that $b_{10^3,\max}$ is a stochastic quantity.

– $c_{\mathrm{E,ref}}$ denotes the actual value of the evidence of the example problem. The quantity $p_{\Omega,\mathrm{ref}}$ is defined as $p_{\Omega,\mathrm{ref}} = c_{\mathrm{E,ref}}/L_{\max}$.

– $\hat{c}_{\mathrm{E},K}$ is the evidence estimated by the investigated algorithm based on K posterior samples.

– a_K and s_K denote the estimated mean and standard deviation of the first component of the parameter vector in a set of K posterior samples. Note that a_K and s_K are

Table 7.1: Reference solution of the investigated example problems.

	Example problem			
	1a	**1b**	**2**	**3**
$c_{\mathrm{E,ref}}$	$6.16 \cdot 10^{-3}$	$2.36 \cdot 10^{-6}$	$1.00 \cdot 10^{-6}$	$1.52 \cdot 10^{-3}$
$L_{\max}$	1.33	1.99	$7.47 \cdot 10^{-3}$	1.00
$p_{\Omega,\mathrm{ref}}$	$4.63 \cdot 10^{-3}$	$1.18 \cdot 10^{-6}$	$1.34 \cdot 10^{-4}$	$1.52 \cdot 10^{-3}$
$\mathrm{E}[\theta_1\|\mathcal{D}]$	2.75	4.81	0.34	1.12
$\sigma[\theta_1\|\mathcal{D}]$	0.287	0.196	0.51	0.66
$\mathrm{E}[b_{10^3,\max}]$	1	1	0.46	0.999
$\Pr\left[b_{10^3,\max} > 0.8\right]$	1	1	$1 \cdot 10^{-4}$	1
$\Pr\left[b_{10^3,\max} < 0.99\right]$	10^{-37}	10^{-32}	1	$1 \cdot 10^{-4}$
$\Pr\left[b_{10^3,\max} < 0.999\right]$	$5 \cdot 10^{-12}$	$1 \cdot 10^{-10}$	1	0.38

random variables for finite K. If the investigated algorithm produces posterior samples, we have $\mathrm{E}[a_K] = \mathrm{E}[\theta_1|\mathcal{D}]$ and $\mathrm{E}[s_K] = \sigma[\theta_1|\mathcal{D}]$. If the generated posterior samples are independent, then $\sigma[a_K] = \frac{1}{\sqrt{K}} \cdot \sigma[\theta_1|\mathcal{D}]$. For dependent samples, $\sigma[a_K]$ can be expressed as

$$\sigma[a_K] = \sqrt{\frac{1+\gamma}{K}} \cdot \sigma[\theta_1|\mathcal{D}] \tag{7.1}$$

where $\gamma \geq 0$ quantifies the dependency of the generated samples.

- N_{eff}: the number of effectively independent samples in the generated set of K posterior samples (of the first component θ_1). This quantity specifies how many truly independent posterior samples of θ_1 would give the same variance in the sample mean as $\mathrm{Var}[a_K]$ obtained by aBUS.

$$N_{\mathrm{eff}} = \left(\frac{\mathrm{E}\left[s_K\right]}{\sigma\left[a_K\right]}\right)^2 = \frac{K}{1+\gamma} \tag{7.2}$$

  Note that N_{eff} can be interpreted as a measure for the dependency of the generated posterior samples; the smaller N_{eff} the stronger the dependency.

– n_K is the total number of prior samples needed to generate K posterior samples in *Algorithm (7.3)*.

– $\hat{\theta}_1$ represents a posterior sample obtained with the investigated algorithm.

The reference solutions of the presented example problems are summarized in Table 7.1. Additional to the quantities $c_{\mathrm{E,ref}}$, $L_{\max}$, $p_{\Omega,\mathrm{ref}}$, $\mathrm{E}[\theta_1|\mathcal{D}]$ and $\sigma[\theta_1|\mathcal{D}]$, the statistics of quantity $b_{10^3,\max}$ are listed in the last four rows. It is obvious that *Example problem 2* differs from the other problems with respect to the statistics of $b_{10^3,\max}$: For *Example problems 1a*, *1b* and *3*, the expectation of $b_{10^3,\max}$ is very close to *one*: $\mathrm{E}[b_{10^3,\max}] = 1$ and $\Pr\left[b_{10^3,\max} < 0.999\right] = 5 \cdot 10^{-12}$ for *Example problem 1a*, $\mathrm{E}[b_{10^3,\max}] = 1$ and $\Pr\left[b_{10^3,\max} < 0.999\right] = 1 \cdot 10^{-10}$ for *Example problem 1b*, and $\mathrm{E}[b_{10^3,\max}] = 1$ and $\Pr\left[b_{10^3,\max} < 0.99\right] = 1 \cdot 10^{-4}$ for *Example problem 3*. However, for *Example problem 2*, $\mathrm{E}[b_{10^3,\max}] = 0.46$ and $\Pr\left[b_{10^3,\max} > 0.8\right] =$

$1 \cdot 10^{-4}$. Consequently, it is extremely unlikely that a $b_{10^3,\max}$ close to *one* will be observed in a set of 10^3 posterior samples.

7.3 Bayesian updating with structural reliability methods (BUS)

This section contains material originally published in [Betz et al., 2017]. Some passages and figures are directly taken from the mentioned reference.

7.3.1 Introduction

BUS [Straub and Papaioannou, 2015] is a recently introduced accept/reject sampling method for Bayesian updating that converts sampling from the posterior into sampling from the failure domain of a structural reliability problem. In structural reliability, probabilities of rare events are estimated [Ditlevsen and Madsen, 2007; Melchers, 1999; Straub, 2014] (see Section 4.4). By interpreting the Bayesian updating problem as a structural reliability problem, existing structural reliability methods can be used to perform the Bayesian analysis. Moreover, an estimate for the evidence $c_{\mathrm{E}|\mathcal{M}}$ is obtained as a by-product of BUS.

An often employed reliability method to tackle the BUS problem is Subset Simulation (SuS) introduced in Section 5.3. The use of SuS in BUS is referred to as *BUS-SuS* in the following and explained in detail in Section 7.3.8.

7.3.2 The idea behind BUS

Straub and Papaioannou show in [Straub and Papaioannou, 2015] that a Bayesian updating problem can be interpreted as a structural reliability problem. The principal idea behind BUS (Bayesian Updating with Structural reliability methods) is to add an additional uniformly distributed random variable π with support $[0, 1]$ to the space of random variables spanned by $\boldsymbol{\theta}$. The updating problem is then expressed as a structural reliability problem in the so-obtained augmented random variable space spanned by the compound vector $[\boldsymbol{\theta}, \pi]$. The "failure" domain Ω of this reliability problem is defined as:

$$\Omega = \{\pi \leq c \cdot L(\boldsymbol{\theta}|\mathcal{D})\} \tag{7.3}$$

where c is a positive constant chosen such that $c \cdot L(\boldsymbol{\theta}|\mathcal{D}) \leq 1$ is maintained for all $\boldsymbol{\theta}$. The domain Ω is exemplified in Fig. 7.1. Note that Ω is used to denote both the failure domain and the corresponding event. The link between the domain Ω and the actual Bayesian updating problem is: Samples from the prior distribution of $\boldsymbol{\theta}$ that are contained in Ω follow the

posterior distribution [Straub and Papaioannou, 2015]. The limit-state function of the BUS problem is defined such that it is $g(\boldsymbol{\theta}, \pi) \leq 0$ if $[\boldsymbol{\theta}, \pi] \in \Omega$; and $g(\boldsymbol{\theta}, \pi) > 0$ if $[\boldsymbol{\theta}, \pi]$ is outside of Ω (see Fig. 7.1). The limit-state function $g(\boldsymbol{\theta}, \pi)$ that describes the "failure" domain Ω defined in Eq. (7.3) can be expressed as:

$$g(\boldsymbol{\theta}, \pi) = \pi - c \cdot L(\boldsymbol{\theta}|\mathcal{D}) \tag{7.4}$$

Optimally, the constant c should be chosen as the reciprocal of the maximum of the likelihood function, denoted $L_{\max}$ [Straub and Papaioannou, 2015]. However, $L_{\max}$ is not always known in advance. In such cases, it is difficult to select c appropriately. The implications of choosing c too large or too small are discussed in Section 7.3.9. An efficient strategy based on *BUS-SuS* that renders a prior selection of c unnecessary is developed in Section 7.4.

7.3.3 Structural reliability methods in BUS

BUS employs structural reliability methods to perform Bayesian updating. The most straightforward (and simplest) application of BUS is rejection sampling – which corresponds to crude Monte Carlo simulation in the context of structural reliability. Rejection sampling within BUS is explained in detail in Section 7.3.7. However, as was pointed out in [Straub and Papaioannou, 2015], other structural reliability methods can be used instead of the simple rejection sampling algorithm (i.e., instead of a Monte Carlo simulation). Typically, BUS is combined with Subset Simulation (see for example [Straub and Papaioannou, 2015; DiazDelaO et al., 2017; Betz et al., 2014b; Papaioannou et al., 2013; Betz et al., 2014a]), because it is efficient for very small failure probabilities and its performance does not depend on the dimension M of the vector of uncertain model parameters $\boldsymbol{\theta}$. The combination of BUS and SuS (*BUS-SuS*) is explained in detail in Section 7.3.8.1. Apart from rejection sampling and SuS, the BUS approach has already been combined with the first order reliability method (FORM) and line sampling in [Straub et al., 2016]. FORM solves the reliability problem only approximately by linearizing the limit-state function at the most probable point of failure [Hasofer and Lind, 1974; Rackwitz and Flessler, 1978]. The line sampling method computes a correction factor for the linearized solution by performing a specified number of line searches perpendicular to a linear approximation of the failure surface [Hohenbichler and Rackwitz, 1988; Koutsourelakis et al., 2004; Rackwitz, 2001].

To generate realizations of the posterior, an additional post-processing step is required: Besides computing the probability of failure, samples located in Ω have to be returned. In Monte Carlo simulation and Subset Simulation, samples located in Ω are directly generated during the reliability analysis – and simply have to be stored. Importance sampling based reliability methods require an additional re-sampling step based on the importance weights associated with the "failed" samples to produce equal weighted samples. In case of FORM or line sam-

pling, samples of the approximated "failure" domain can be easily generated. Samples from the posterior distribution can be further used for posterior prediction of quantities of interest. A special application is the use of BUS for updating the probability of rare events based on observed system response: As the target quantity of interest is the posterior probability of failure, no posterior samples have to be generated and the updating problem can be directly solved by structural reliability methods [Straub, 2011; Straub et al., 2016].

7.3.4 Estimating the evidence in BUS

An estimate for the evidence $c_{\mathrm{E}|\mathcal{M}}$ is obtained as a by-product of BUS. Let p_Ω be the probability that samples $[\boldsymbol{\theta}, \pi]$ from the prior distribution fall into Ω, i.e.:

$$p_\Omega = \Pr[\Omega] = \Pr[g(\boldsymbol{\theta}, \pi) \leq 0] \tag{7.5}$$

p_Ω is the target quantity of interest in a reliability analysis and referred to as the *probability of failure*. In BUS, p_Ω is directly linked to the evidence $c_{\mathrm{E}|\mathcal{M}}$ through c [Straub and Papaioannou, 2015]:

$$c_{\mathrm{E}|\mathcal{M}} = \frac{p_\Omega}{c} \tag{7.6}$$

Note that some reliability methods allow us to evaluate uncertainty bounds for the estimate of p_Ω. In this case, the statistical uncertainty in the estimated evidence $c_{\mathrm{E}|\mathcal{M}}$ can be quantified directly, as the evidence is directly proportional to p_Ω.

7.3.5 Outline of a simple proof of BUS

A simple proof that demonstrates the validity of BUS is [Straub and Papaioannou, 2015]: The product $cL(\boldsymbol{\theta}|\mathcal{D})$ can be expressed as:

$$cL(\boldsymbol{\theta}|\mathcal{D}) = \int_{\pi \leq cL(\boldsymbol{\theta}|\mathcal{D})} \mathrm{d}\pi \tag{7.7}$$

Consequently, $L(\boldsymbol{\theta}|\mathcal{D}) \cdot p(\boldsymbol{\theta})$ can be stated as:

$$L(\boldsymbol{\theta}|\mathcal{D}) \cdot p(\boldsymbol{\theta}) = \frac{1}{c} \int_{\pi \leq cL(\boldsymbol{\theta}|\mathcal{D})} p(\boldsymbol{\theta})\, \mathrm{d}\pi \tag{7.8}$$

By inserting Eq. (7.8) into Eq. (6.2) we can easily prove the validity of Eq. (7.6).

7.3.6 BUS in standard Normal space

For some reliability methods it is convenient to transform the reliability problem to the so-called underlying *standard Normal space* (see Section 5.1.2). In the BUS approach, the random variable space of $\boldsymbol{\theta}$ is augmented by the uniform random variable π. Thus, the transformation must be performed in the augmented random variable space. Let $\mathbf{u}^{(\boldsymbol{\theta})}$ be a M-dimensional vector whose M components are independent standard Normal random variables. The transformation of $\mathbf{u}^{(\boldsymbol{\theta})}$ to $\boldsymbol{\theta}$ is denoted as: $\mathrm{T}_{\boldsymbol{\theta}}^{-1} : \mathbf{u}^{(\boldsymbol{\theta})} \rightarrow \boldsymbol{\theta}$. Furthermore, let $\mathbf{u}$ be a $(M+1)$-dimensional vector that extends $\mathbf{u}^{(\boldsymbol{\theta})}$ by one dimension. The last component of $\mathbf{u}$, denoted u_{M+1}, is transformed as: $\pi = \Phi(u_{M+1})$, where $\Phi(\cdot)$ is the CDF of the standard Normal distribution; i.e. u_{M+1} also follows a standard Normal distribution. Thus, the limit-state function $g(\boldsymbol{\theta}, \pi)$ defined in Eq. (7.4) can then be equivalently expressed as

$$g(\boldsymbol{\theta}, \pi) = G(\mathbf{u}) = \Phi(u_{M+1}) - c \cdot L\left(\mathrm{T}_{\boldsymbol{\theta}}^{-1}(\mathbf{u}^{(\boldsymbol{\theta})}) | \mathcal{D}\right) \tag{7.9}$$

Note that the prior distribution of $\mathbf{u}$ is described by $p(\mathbf{u}) = \prod_{i=1}^{M+1} \varphi(u_i)$, where $\varphi(\cdot)$ is the PDF of the standard Normal distribution.

7.3.7 BUS with rejection sampling

The most trivial application of the BUS idea is the *rejection sampling* algorithm[1] [Smith and Gelfand, 1992; Straub and Papaioannou, 2015]: This algorithm repeatedly proposes a sample $[\tilde{\boldsymbol{\theta}}, \tilde{\pi}]$ from the prior distribution and accepts the sample if it is located in the "failure" domain; i.e., if $[\tilde{\boldsymbol{\theta}}, \tilde{\pi}] \in \Omega$. The accepted sample $\tilde{\boldsymbol{\theta}}$ is a sample from the posterior distribution. The algorithm is repeated until K posterior samples are generated. The posterior samples resulting from the rejection sampling algorithm are statistically independent.

The quantity p_Ω is the probability that a proposed sample is accepted; p_Ω is also referred to as the *acceptance probability*. An unbiased estimate $\hat{p}_\Omega$ of p_Ω is [Haldane, 1945]:

$$p_\Omega \approx \hat{p}_\Omega = \frac{K-1}{n-1} \tag{7.10}$$

where n is the total number of prior samples that were proposed to generate K posterior samples. Note that contrary to $\hat{p}_\Omega$, the estimator K/n produces a biased estimate for p_Ω [Haldane, 1945]. An unbiased estimate of the variance of $\hat{p}_\Omega$ is [Finney, 1949]:

$$\mathrm{Var}[\hat{p}_\Omega] \approx \frac{(1-\hat{p}_\Omega)\,\hat{p}_\Omega}{n-2} \tag{7.11}$$

The estimates given in Eqs. (7.10) and (7.11) are frequentist estimates. To appropriately

[1] Standard rejection sampling is explained in Section 3.3.

quantify the uncertainty about p_Ω based on the outcome of a particular run of rejection sampling, a Bayesian approach is recommended. The number n of prior samples needed to generate K posterior samples in the rejection sampling algorithm follows a negative binomial distribution. Thus, having observed a certain n for a given K, the likelihood of p_Ω is:

$$L(p_\Omega|n,K) = \binom{n-1}{K-1}(p_\Omega)^K(1-p_\Omega)^{n-K} \tag{7.12}$$

where $\binom{n-1}{K-1}$ denotes the binomial coefficient. The beta distribution acts as conjugate prior for the problem at hand. If the beta distribution is selected as prior distribution of p_Ω, and the shape parameters of the distribution are selected based on the Principle of Maximum Information Entropy [Jaynes, 1983, 2003], the resulting prior distribution[1] is the uniform distribution on $[0,1]$. In this case, the posterior is also a beta distribution and can be expressed as:

$$p(p_\Omega|K,n) = \frac{{p_\Omega}^K \cdot (1-p_\Omega)^{n-K}}{\mathrm{B}(K+1, n-K+1)} \tag{7.13}$$

where B denotes the *beta function*. The expectation of Eq. (7.13) is:

$$\mathrm{E}\left[p_\Omega|K,n\right] = \frac{K+1}{n+2} \tag{7.14}$$

The variance of the distribution in Eq. (7.13) can be derived as:

$$\mathrm{Var}\left[p_\Omega|K,n\right] = \frac{(K+1)\cdot(n-K+1)}{(n+2)^2\cdot(n+3)} \tag{7.15}$$

For increasing K and n, Eqs. (7.10) and (7.14) as well as Eqs. (7.11) and (7.15) converge to the same value.

Algorithm 7.1. *Rejection sampling algorithm for BUS*:
As input the algorithm requires:

- K, the total number of samples to draw from the posterior distribution.
- c, selected such that $c^{-1} \geq L_{\max}$.

The algorithm evaluates the evidence $c_{\mathrm{E}|\mathcal{M}}$ and returns K unweighted and statistically independent posterior samples $\boldsymbol{\theta}_{(k)}$ with $k = 1,\ldots,K$.

1. Initialize counters $k = 0$ and $n = 0$.
2. **while** $(k < K)$ **do**:
   (a) Propose sample $[\tilde{\boldsymbol{\theta}}, \tilde{\pi}]$:

[1] For a related discussion of a Bayesian interpretation of the probability of failure obtained with Monte Carlo simulation, see Section 5.2.3.1

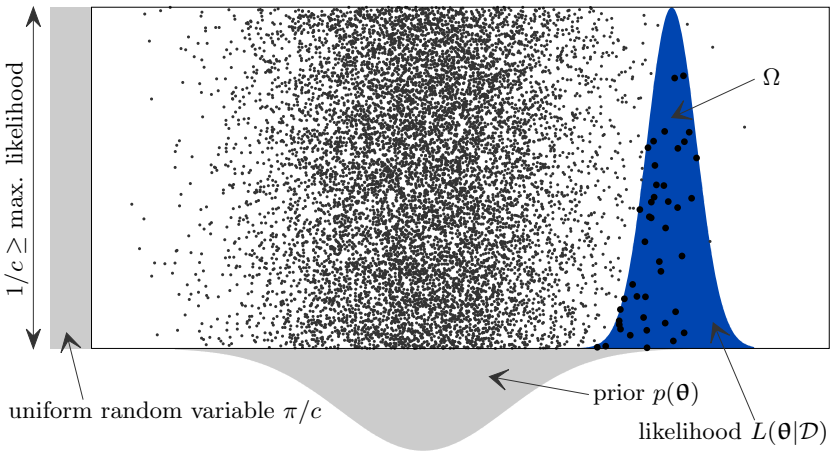


Figure 7.1: Illustration of the principle of the rejection sampling algorithm – *Algorithm (7.1)*. The highlighted region is the domain Ω defined in Eq. (7.3). The limit-state function $g(\boldsymbol{\theta}, \pi)$ introduced in Eq. (7.4) is smaller or equal than *zero* within Ω (it is *zero* at the boundary of Ω), and larger than *zero* outside of Ω. Samples "below" the likelihood (i.e., the samples contained in Ω) are independent samples from the posterior distribution. In this example, 43 out of 10^4 samples are accepted.

   i. Draw $\tilde{\boldsymbol{\theta}}$ from the prior distribution $p(\boldsymbol{\theta})$.

   ii. Draw $\tilde{\pi}$ from the uniform distribution that has support $[0, 1]$.

(b) **if** $\left(g(\tilde{\boldsymbol{\theta}}, \tilde{\pi}) \leq 0\right)$ **then**:

   i. Increase the counter $k = k + 1$.

   ii. Accept the proposed sample $\tilde{\boldsymbol{\theta}}$ as a posterior sample, i.e.: set $\boldsymbol{\theta}_{(k)} = \tilde{\boldsymbol{\theta}}$.

(c) Increase the counter $n = n + 1$.

3. Estimate p_Ω by means of Eq. (7.10) or Eq. (7.13).

4. Evaluate the evidence $c_{\mathrm{E}|\mathcal{M}} = p_\Omega / c$

Algorithm (7.1) is an extended variant of *Algorithm (3.1)*, specifically designed for BUS. On average, the algorithm requires K/p_Ω samples from the prior distribution to generate K samples from the posterior distribution. The principle of the *rejection sampling* algorithm is illustrated in Fig. 7.1. Note that *Algorithm (7.1)* is similar to a Monte Carlo simulation for solving the structural reliability problem that has limit-state function $g(\boldsymbol{\theta}, \pi)$ and random variables $[\boldsymbol{\theta}, \pi]$. The difference is that in a Monte Carlo simulation typically the total number of samples n is fixed whereas in *Algorithm (7.1)* the number K of samples to be generated in the domain Ω is specified.

The main advantage of rejection sampling is that it produces independent samples from the posterior distribution. However, if the posterior distribution does not match the prior distribution well, p_Ω becomes small which renders the rejection sampling algorithm inefficient.

As a consequence, rejection sampling is often inefficient in the BUS framework and typically more advanced reliability methods are employed – see Section 7.3.3.

7.3.8 BUS-SuS: BUS with Subset Simulation

The Subset Simulation (SuS) algorithm [Au and Beck, 2001] (see Section 5.3) is a structural reliability method that is particularly well suited for BUS: (i) SuS can efficiently handle problems with many uncertain parameters; (ii) SuS can efficiently estimate very small probabilities that arise within BUS when the scalar $c_{\mathrm{E}|\mathcal{M}}$ and the constant c are small. Hence, the combination of BUS and SuS (referred to as *BUS-SuS*) is well suited for problems where it is computationally demanding to evaluate the likelihood function.

7.3.8.1 Formulation of the limit-state function

The standard limit-state function of the BUS problem is given in Eqs. (7.4) and (7.9) for the original parameter space $[\boldsymbol{\theta}, \pi]$ and the standard Normal space $\mathbf{u}$, respectively. However, the particular format of the limit-state function of the BUS problem is not uniquely defined: Any limit-state function that has the same probability of failure p_Ω and the same limit-state surface (the surface were the limit-state function equals *zero*) as $g(\boldsymbol{\theta}, \pi)$ for a given c is a valid limit-state function for the respective BUS problem.

For rejection sampling, the performance of the method does not depend on the particular choice of the limit-state function, because the method checks only if a sample is inside or outside of the failure domain. However, for BUS with Subset Simulation, the formulation of the limit-state function has an impact. This is related to the fact that Subset Simulation introduces intermediate failure events. This intermediate failure events are defined as $g(\boldsymbol{\theta}, \pi) \leq h_i$, where h_i is a positive constant (see Section 5.3.1). The particular shapes of the intermediate failure levels depend on the selected limit-state function. Loosely speaking, a smooth transition of the intermediate failure levels has a positive influence on the performance of Subset Simulation.

From a numerical point of view, the limit-state function defined in Eq. (7.4) and Eq. (7.9) is not optimal, because samples with small values of π are preferred over samples with large values of π in the initial levels of Subset Simulation (especially if prior realizations of the likelihood are small compared to c^{-1}). An alternative representation of the limit-state function that has a more appropriate shape is:

$$g_{\mathrm{l}}(\boldsymbol{\theta}, \pi) = \ln(\pi) - \ln\left(c \cdot L\left(\boldsymbol{\theta}|\mathcal{D}\right)\right) \tag{7.16}$$

where $\ln(\cdot)$ denotes the natural logarithm. $g_{\mathrm{l}}(\boldsymbol{\theta}, \pi)$ is obtained by applying the natural logarithm to each of the terms in Eq. (7.4). The transformation of the BUS limit-state function

given in Eq. (7.16) was first used in [DiazDelaO et al., 2017] to propose a variant of the BUS approach. By comparing Eq. (7.16) with Eq. (7.4) it is obvious that both functions have the same failure domain. For enhanced numerical stability, it is usually of advantage to work with the log-transform of the likelihood, denoted $\ln L(\boldsymbol{\theta}|\mathcal{D}) = \ln\left(L(\boldsymbol{\theta}|\mathcal{D})\right)$, instead of using the likelihood directly. Eq. (7.16) can then be expressed as:

$$g_{\mathrm{l}}(\boldsymbol{\theta}, \pi) = \ln(\pi) + \ell - \ln L\left(\boldsymbol{\theta}|\mathcal{D}\right) \tag{7.17}$$

where $\ell = -\ln(c)$. Based on Eq. (7.16), the intermediate failure domains can be stated as:

$$\begin{aligned} Z_i &= \left\{\boldsymbol{\theta} \in \mathbb{R}^M \middle| \ln(\pi) - \ln\left(c \cdot L\left(\boldsymbol{\theta}|\mathcal{D}\right)\right) \leq h_i\right\} \\ &= \left\{\boldsymbol{\theta} \in \mathbb{R}^M \middle| \ln(\pi) \leq \ln\left(c \cdot L\left(\boldsymbol{\theta}|\mathcal{D}\right)\right) + h_i\right\} \\ &= \left\{\boldsymbol{\theta} \in \mathbb{R}^M \middle| \pi \leq c \cdot L\left(\boldsymbol{\theta}|\mathcal{D}\right) \cdot \exp(h_i)\right\} \end{aligned} \tag{7.18}$$

The transition of the intermediate failure levels is illustrated in Fig. 7.2 and Fig. 7.3 for limit-state function $g(\boldsymbol{\theta}, \pi)$ and $g_{\mathrm{l}}(\boldsymbol{\theta}, \pi)$, respectively. Limit-state function $g_{\mathrm{l}}(\boldsymbol{\theta}, \pi)$ (shown in Fig. 7.3) clearly is more appropriate than $g(\boldsymbol{\theta}, \pi)$ (shown in Fig. 7.2), because the intermediate failure domains obtained with $g_{\mathrm{l}}(\boldsymbol{\theta}, \pi)$ converge smoothly to the final failure domain Ω. Another viable representation of the limit-state function that ensures a smooth transition of the intermediate failure domains is [Straub and Papaioannou, 2015]:

$$g_{\mathrm{n}}(\boldsymbol{\theta}, \pi) = \Phi^{-1}(\pi) - \Phi^{-1}\left(c \cdot L\left(\boldsymbol{\theta}|\mathcal{D}\right)\right) \tag{7.19}$$

where $\Phi^{-1}(\cdot)$ is the inverse of the CDF of the standard normal distribution.

In this contribution, we exclusively use limit-state function $g_{\mathrm{l}}(\boldsymbol{\theta}, \pi)$ as defined in Eq. (7.17), because it has particular advantages if the scaling parameter c of *BUS-SuS* is learned adaptively (see Section 7.4).

7.3.8.2 BUS-SuS algorithm

The standard Subset Simulation algorithm is explained in detail in Section 5.3. In the BUS problem, the intermediate failure domains Z_i are defined according to Eq. (7.18); where Z_i is used interchangeably to denote both the domain and the event. Samples conditional on Z_i are denoted $[\boldsymbol{\theta}_{(i,k)}, \pi_{(i,k)}]$, for $k \in \{1, \ldots, K\}$.

The BUS-SuS algorithm is a slightly extended version of *Algorithm (5.2)*:

Algorithm 7.2. *BUS-SuS algorithm (Subset Simulation algorithm for BUS)*:
As input the algorithm requires:

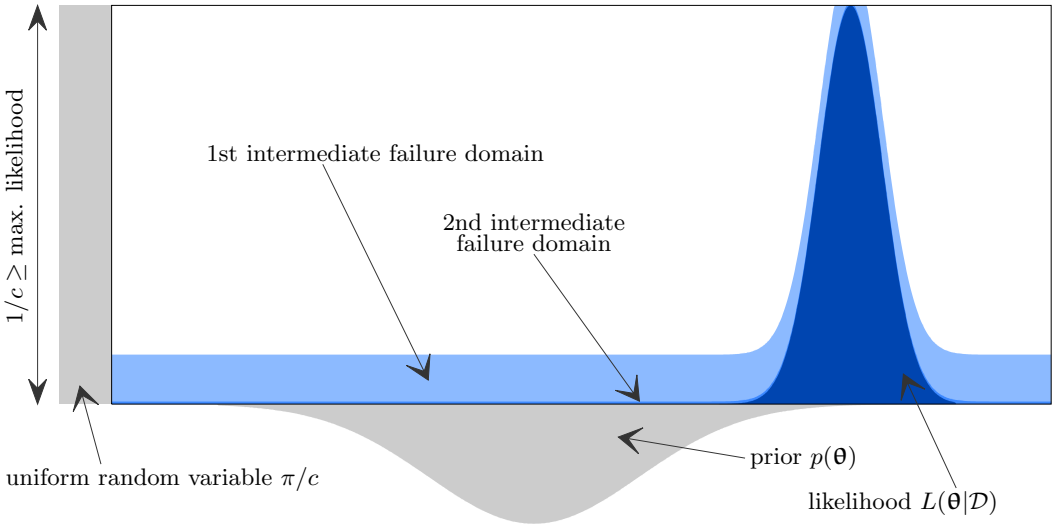


Figure 7.2: Shape of the intermediate failure domains if the original BUS-limit-state function defined in Eq. (7.4) is employed in *BUS-SuS*. The intermediate failure domains obtained with this limit-state function do not exhibit a smooth transition to the final failure domain Ω. Therefore, this particular formulation of the limit-state function should not be used in combination with *BUS-SuS*. Instead, we recommend to use limit-state function $g_1(\boldsymbol{\theta}, \pi)$ defined in Eqs. (7.16) and (7.17) (see Fig. 7.3).

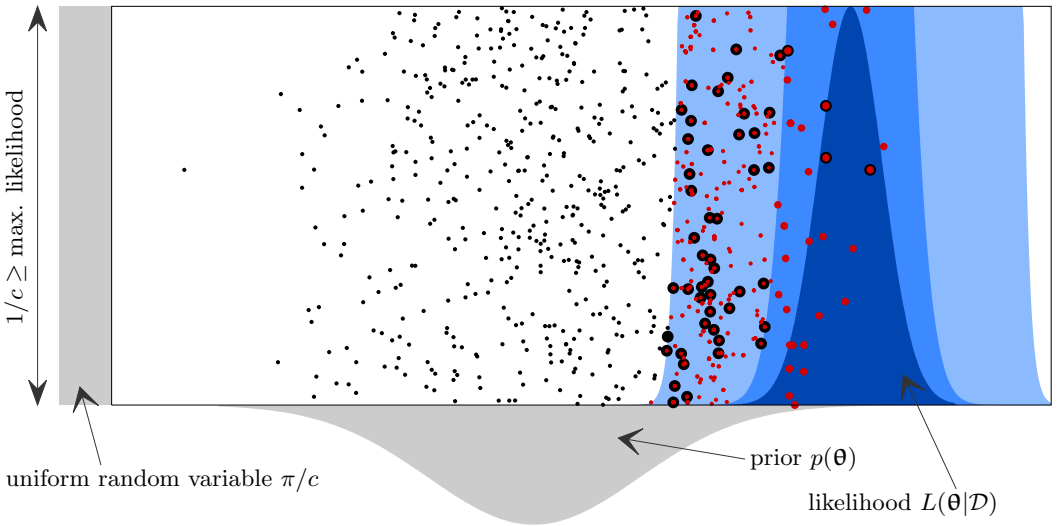


Figure 7.3: Illustration of the principle of the *BUS-SuS* algorithm – *Algorithm (7.2)*. The intermediate failure domains obtained with limit-state function $g_1(\boldsymbol{\theta}, \pi)$ defined in Eq. (7.17) are highlighted. The innermost region is the domain Ω defined in Eq. (7.3); samples within this region follow the posterior distribution. The black samples are the initial samples from the prior distribution ($K = 500$ samples were used per Subset level). The large black dots indicate the $10\% \cdot K$ samples that are located in the first intermediate failure domain Z_1. These samples are used as seed values to generate samples in Z_1 by means of MCMC. The generated samples in Z_1 are highlighted in red. Note that only the *black* samples are independent; the *red* samples are dependent, because they are obtained by means of MCMC. The large red dots indicate the $10\% \cdot K$ samples that are located in the second intermediate failure domain Z_2. The number of Subset levels in this example is $N = 3$.

- K, the total number of samples to draw from the posterior distribution.
- p_t, the probability of the intermediate subsets. p_t needs to be selected such that $p_\mathrm{t} \cdot K$ is an integer number.
- c, selected such that $c^{-1} \geq L_\mathrm{max}$.

The algorithm evaluates the evidence $c_{\mathrm{E}|\mathcal{M}}$ and returns K unweighted but dependent posterior samples $\boldsymbol{\theta}_{(k)}$ with $k = 1, \ldots, K$.

1. Draw K samples $[\boldsymbol{\theta}_{(0,k)}, \pi_{(0,k)}]$, with $k = 1, \ldots, K$, from the prior distribution.
2. Initialize $i = 0$ and $h_0 = \infty$.
3. **while** $(h_i > 0)$ **do**:
   (a) Increase counter i by *one*: $i = i + 1$.
   (b) Select the threshold level h_i:
      i. Sort the K samples $\left\{[\boldsymbol{\theta}_{(i-1,k)}, \pi_{(i-1,k)}]\right\}_{k=1}^{K}$ with respect to the value of $g_\mathrm{l}(\boldsymbol{\theta}_{(i-1,k)}, \pi_{(i-1,k)})$ in ascending order.
      ii. Set $h_i = \frac{g_\mathrm{l}\left(\boldsymbol{\theta}_{(i-1,p_\mathrm{t}\cdot K)}\right) + g_\mathrm{l}\left(\boldsymbol{\theta}_{(i-1,p_\mathrm{t}\cdot K+1)}\right)}{2}$; i.e., set h_i as the p_t-percentile of the ordered set.
      iii. Select n as the number of samples in $\left\{[\boldsymbol{\theta}_{(i-1,k)}, \pi_{(i-1,k)}]\right\}_{k=1}^{K}$ with $g_\mathrm{l}(\boldsymbol{\theta}_{(i-1,k)}, \pi_{(i-1,k)}) \leq \max(h_i, 0)$.
      iv. **if** $(h_i < 0)$ **then**: Set $h_i = 0$, and $p_i = \frac{n}{K}$.
      else: Set $p_i = p_\mathrm{t}$.
   (c) Generate samples conditional on domain Z_i:
      i. Randomize the ordering of the samples in the set $\left\{[\boldsymbol{\theta}_{(i-1,k)}, \pi_{(i-1,k)}]\right\}_{k=1}^{n}$; i.e., thereafter, the n samples are no longer ordered.
      ii. Generate the samples $[\boldsymbol{\theta}_{(i,k)}, \pi_{(i,k)}]$ by means of n Markov chains; e.g., by means of the CS algorithm (*Algorithm (3.6)*) applied inside *Algorithm (3.3)*. The n samples $\left\{[\boldsymbol{\theta}_{(i-1,k)}, \pi_{(i-1,k)}]\right\}_{k=1}^{n}$ are already within Z_i and are used as seeds for the n Markov chains. The length of each Markov chain is K/n. Thus, the total number of MCMC samples generated in one level is $K - n$.
      Algorithm (3.9) is applied during the MCMC sampling to adopt the spread of the MCMC proposal distribution.
4. Set $N = i$
5. Estimate $p_\Omega = \prod_{i=1}^{m} p_i$
6. Evaluate the evidence $c_{\mathrm{E}|\mathcal{M}} = p_\Omega / c$

Note that we actually perform the generation of samples by means of MCMC in *step 3c(ii)* of *Algorithm (7.2)* in the underlying *standard Normal* space. Thus, the standard Normal transform $\mathbf{u}$ of each generated sample $[\boldsymbol{\theta}, \pi]$ is ideally stored as well. For the sake of convenience,

this is not explicitly explained in *Algorithm (7.2)*; the procedure in standard Normal space is explained in more detail in *Algorithm (5.2)*. Furthermore, note that MCMC sampling needs to be performed in the $M+1$-dimensional augmented parameter space.

The principle behind *Algorithm (7.2)* is exemplified in Fig. 7.3. Note that *step 3c(i)* in *Algorithm (7.2)* is not a standard step in Subset Simulation; *step 3c(i)* corresponds to *step 4* in *Algorithm (5.2)*. This step is introduced to tune the spread of the MCMC proposal distribution during the MCMC sampling – see *Algorithm (3.9)*. Without this step, *Algorithm (3.9)* would possibly introduce a bias.

Algorithm (7.2) is presented so that the employed number of samples per level is equivalent to the number K of samples in the final level of Subset Simulation. In general, the number of samples in the final level of *BUS-SuS* can be chosen larger than the number of samples in the intermediate levels, simply by generating more MCMC samples if the threshold level h_i in SuS becomes *zero*. However, in this contribution we only investigate the case where the number of samples in each level of SuS is the same as the final number K of posterior samples generated.

7.3.9 Correcting the results of BUS simulations with c^{-1} selected too small

7.3.9.1 The constant c in BUS

An appropriate selection of the constant c is crucial in BUS: On the one hand, if c^{-1} is selected larger than the maximum $L_{\max}$ of the likelihood function, the efficiency of the approach decreases; the acceptance probability p_Ω of BUS decreases linearly with c. On the other hand, if c^{-1} is selected smaller than the maximum $L_{\max}$ of the likelihood function, BUS does not produce samples that follow the posterior distribution. Instead, in such case BUS produces samples that follow distribution $p^{(c)}_{\text{trunc}}$:

$$p^{(c)}_{\text{trunc}}(\boldsymbol{\theta}|\mathcal{D}) = \frac{L^{(c)}_{\text{trunc}}(\boldsymbol{\theta}|\mathcal{D}) \cdot p(\boldsymbol{\theta})}{{c_{\mathbf{E}|\mathcal{M}}}^{(c)}} = \frac{c \cdot L^{(c)}_{\text{trunc}}(\boldsymbol{\theta}|\mathcal{D}) \cdot p(\boldsymbol{\theta})}{p^{(c)}_\Omega} \tag{7.20}$$

where ${c_{\mathbf{E}|\mathcal{M}}}^{(c)} = p^{(c)}_\Omega / c$ and $p^{(c)}_\Omega$ is the associated evidence and acceptance probability that belongs to the selected $c^{-1} < L_{\max}$, respectively. Furthermore, $L^{(c)}_{\text{trunc}}(\boldsymbol{\theta}|\mathcal{D})$ is defined as:

$$L^{(c)}_{\text{trunc}}(\boldsymbol{\theta}|\mathcal{D}) = \min\left(L(\boldsymbol{\theta}|\mathcal{D}), c^{-1}\right) \tag{7.21}$$

Let $w^{(c)}$ be a correction factor such that

$$c_{\mathbf{E}|\mathcal{M}} = w^{(c)} \cdot {c_{\mathbf{E}|\mathcal{M}}}^{(c)} \tag{7.22}$$

The correction factor $w^{(c)}$ can be expanded as:

$$w^{(c)} = \frac{c_{\mathrm{E}|\mathcal{M}}}{{c_{\mathrm{E}|\mathcal{M}}}^{(c)}} \tag{7.23}$$

$$= \frac{\int_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\mathcal{D})\, p(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}}{{c_{\mathrm{E}|\mathcal{M}}}^{(c)}} \tag{7.24}$$

$$= \int_{\boldsymbol{\theta}} \frac{L(\boldsymbol{\theta}|\mathcal{D})}{L^{(c)}_{\mathrm{trunc}}(\boldsymbol{\theta}|\mathcal{D})} \cdot \frac{L^{(c)}_{\mathrm{trunc}}(\boldsymbol{\theta}|\mathcal{D}) \cdot p(\boldsymbol{\theta})}{{c_{\mathrm{E}|\mathcal{M}}}^{(c)}}\, \mathrm{d}\boldsymbol{\theta} \tag{7.25}$$

$$= \int_{\boldsymbol{\theta}} \frac{L(\boldsymbol{\theta}|\mathcal{D})}{L^{(c)}_{\mathrm{trunc}}(\boldsymbol{\theta}|\mathcal{D})} \cdot p^{(c)}_{\mathrm{trunc}}(\boldsymbol{\theta}|\mathcal{D})\, \mathrm{d}\boldsymbol{\theta} \tag{7.26}$$

The quantities involved in Eq. (7.26) can be easily estimated based on the samples $\{\boldsymbol{\theta}^{(c)}_{(k)}\}_{k=1,\ldots,K}$, generated with BUS and $c^{-1} < L_{\max}$ that follow distribution $p^{(c)}_{\mathrm{trunc}}(\boldsymbol{\theta}|\mathcal{D})$:

$$w^{(c)} \approx \frac{1}{K} \sum_{k=1}^{K} \frac{L(\boldsymbol{\theta}^{(c)}_{(k)}|\mathcal{D})}{L^{(c)}_{\mathrm{trunc}}(\boldsymbol{\theta}^{(c)}_{(k)}|\mathcal{D})} \tag{7.27}$$

Using Eq. (7.22) and the estimate for $w^{(c)}$ given in Eq. (7.27), we can correct the evidence computed with $c^{-1} < L_{\max}$. Let p_Ω be the probability of Eq. (7.5) for a choice of $c = 1/L_{\max}$. From Eq. (7.23) it follows that

$$p_\Omega = p^{(c)}_\Omega \cdot w^{(c)} \cdot \frac{c^{-1}}{L_{\max}} \tag{7.28}$$

where ${c_{\mathrm{E}|\mathcal{M}}}^{(c)} = p^{(c)}_\Omega / c$ and $c_{\mathrm{E}|\mathcal{M}} = p_\Omega \cdot L_{\max}$ is used (Eq. (7.6) does not hold if $c^{-1} < L_{\max}$). Note that for $c^{-1} \leq L_{\max}$, we have $p_\Omega \leq p^{(c)}_\Omega$, as the relative size of the failure domain increases with decreasing c^{-1} (for $c^{-1} \to 0$ we have $p^{(c)}_\Omega \to 1$). For $c^{-1} > L_{\max}$, we have $p_\Omega > p^{(c)}_\Omega$, as $p_\Omega = p^{(c)}_\Omega \cdot c^{-1}/L_{\max}$. Moreover, $L(\boldsymbol{\theta}^{(c)}_{(k)}|\mathcal{D}) \geq L^{(c)}_{\mathrm{trunc}}(\boldsymbol{\theta}^{(c)}_{(k)}|\mathcal{D})$ clearly holds independent of c, due to Eq. (7.21). Thus, the values that $w^{(c)}$ can take are bounded: $1 \leq w^{(c)} \leq L_{\max}/c^{-1}$. For $c^{-1} \geq L_{\max}$ the correction factor is *one*. For $c^{-1} \leq L_{\max}$ the correction factor must be smaller than $L_{\max}/c^{-1}$ in order to maintain $p_\Omega \leq p^{(c)}_\Omega$ (see Eq. (7.28)). Additionally, it becomes clear that ${c_{\mathrm{E}|\mathcal{M}}}^{(c)}$ underestimates the actual evidence $c_{\mathrm{E}|\mathcal{M}}$.

The relative error in the evidence associated with selecting c^{-1} too small (i.e., $c^{-1} < L_{\max}$) can be expressed as:

$$\varepsilon^{(c)}_{c_{\mathrm{E}|\mathcal{M}}} = 1 - \frac{1}{w^{(c)}} = 1 - \frac{{c_{\mathrm{E}|\mathcal{M}}}^{(c)}}{c_{\mathrm{E}|\mathcal{M}}} = 1 - \frac{p^{(c)}_\Omega}{p_\Omega} \frac{c^{-1}}{L_{\max}} \tag{7.29}$$

$$= 1 - \int_{\boldsymbol{\theta}} \frac{L^{(c)}_{\mathrm{trunc}}(\boldsymbol{\theta}|\mathcal{D})}{L(\boldsymbol{\theta}|\mathcal{D})} \cdot \frac{L(\boldsymbol{\theta}|\mathcal{D}) \cdot p(\boldsymbol{\theta})}{c_{\mathrm{E}|\mathcal{M}}}\, \mathrm{d}\boldsymbol{\theta} \tag{7.30}$$

$$= 1 - \mathrm{E}_{\boldsymbol{\theta}|\mathcal{D}} \left[\frac{L^{(c)}_{\mathrm{trunc}}(\boldsymbol{\theta}|\mathcal{D})}{L(\boldsymbol{\theta}|\mathcal{D})} \right] \tag{7.31}$$

For $c^{-1} = 0$, the error is $\varepsilon_{c_{\mathrm{E}|\mathcal{M}}}^{(\infty)} = 1$ and the generated samples follow the prior distribution. For $c^{-1} = L_{\max}$, we have $\varepsilon_{c_{\mathrm{E}|\mathcal{M}}}^{(1/L_{\max})} = 0$ and the samples follow the posterior distribution.

7.3.9.2 Post-processing step to correct the posterior distribution

As was discussed in the previous section, BUS does not return samples from the posterior distribution if the constant c^{-1} is selected smaller than $L_{\max}$. Employing Eq. (7.27), the evidence of the investigated problem can be estimated even if c is not selected properly. Additional to that, the distribution of the generated samples needs to be corrected: The samples produced by BUS if $c^{-1} < L_{\max}$ follow distribution $p_{\mathrm{trunc}}^{(c)}(\boldsymbol{\theta}|\mathcal{D})$. However, the aim is to obtain samples from the posterior distribution. The posterior distribution can be corrected by one of the following two strategies:

1. The (equal weighted) posterior samples obtained with c^{-1} selected too small, are corrected by means of importance weights. The kth, $k \in \{1, \ldots, K\}$, posterior sample is associated with importance weight:

$$w_{(k)}^{(c)} = \frac{1}{w^{(c)} \cdot K} \cdot \frac{L(\boldsymbol{\theta}_{(k)}^{(c)}|\mathcal{D})}{L_{\mathrm{trunc}}^{(c)}(\boldsymbol{\theta}_{(k)}^{(c)}|\mathcal{D})} \tag{7.32}$$

   Thus, with this approach, weighted samples of the posterior distribution are obtained.

2. A Metropolis-Hastings [Metropolis et al., 1953; Hastings, 1970] step is added after the BUS simulation with c^{-1} selected too small, to ensure that the generated samples follow the desired posterior distribution. The proposed Metropolis-Hastings post-processing step (*Algorithm (7.3)*) is based on the algorithm presented in [Tierney, 1994].

In general, the strategy listed first should be more efficient compared to the second one: In the re-sampling required for the second strategy, samples with comparatively large weights are likely to appear more often in the list of generated posterior samples, whereas samples with small weights are prone to be repudiated. Nevertheless, in this chapter (Chapter 7), focus is put on sampling methods that produce posterior samples with equal weighs. Therefore, only the second strategy is considered in the following.

Algorithm 7.3. *M-H rejection sampling*:
As input the algorithm requires:

- The K samples $\boldsymbol{\theta}_{(k)}^{(c)}, k \in \{1, \ldots, K\}$ generated with BUS that follow distribution $p_{\mathrm{trunc}}^{(c)}(\boldsymbol{\theta}|\mathcal{D})$.
- Scaling constant c employed in BUS, where c^{-1} may be selected smaller than $L_{\max}$.

The algorithm evaluates the evidence $c_{\mathrm{E}|\mathcal{M}}$ and returns K unweighted but dependent posterior samples $\boldsymbol{\theta}_{(k)}$ with $k = 1, \ldots, K$.

1. Evaluate the evidence $c_{\mathrm{E}|\mathcal{M}}$
   (a) Estimate $p_{\Omega}^{(c)}$ by means of Eq. (7.10) or Eq. (7.13).
   (b) Evaluate the quantity $w^{(c)}$, defined in Eq. (7.27).
   (c) An estimate $\hat{c}_{\mathrm{E},K}$ of $c_{\mathrm{E}|\mathcal{M}}$ is obtained through Eq. (7.22).
2. Pick sample $\boldsymbol{\theta}_{(1)}$:
   (a) Draw index i randomly from the list $\{1, \ldots, K\}$ where the jth element of the list is associated with probability $\frac{L(\boldsymbol{\theta}_{(j)}^{(c)}|\mathcal{D})}{L_{\mathrm{trunc}}^{(c)}(\boldsymbol{\theta}_{(j)}^{(c)}|\mathcal{D})} / \left(w^{(c)} \cdot K\right)$.
   (b) **swap** $\boldsymbol{\theta}_{(i)}^{(c)}$ with $\boldsymbol{\theta}_{(1)}^{(c)}$
   (c) Set $\boldsymbol{\theta}_{(1)} = \boldsymbol{\theta}_{(1)}^{(c)}$
   (d) Set $k = 2$
3. **while** $(k \leq K)$ **do**:
   (a) Draw a sample u randomly from a uniform distribution with support $[0, 1]$.
   (b) Evaluate the accept/reject-ratio r_k:
   $r_k = \min\left(1, \frac{\max\left(L\left(\boldsymbol{\theta}_{(k)}^{(c)}\middle|\mathcal{D}\right), c^{-1}\right)}{L\left(\boldsymbol{\theta}_{(k-1)}\middle|\mathcal{D}\right)}\right)$
   (c) **if** $(u \leq r_k)$ **then** accept the candidate sample:
   Set $\boldsymbol{\theta}_{(k)} = \boldsymbol{\theta}_{(k)}^{(c)}$
   else; i.e., if $(u > r_k)$ then reject the candidate sample:
   Set $\boldsymbol{\theta}_{(k)} = \boldsymbol{\theta}_{(k-1)}$.
   (d) Increase the counter $k = k + 1$.

In the first step of *Algorithm (7.3)*, the evidence is computed based on the correction factor $w^{(c)}$ derived in Section 7.3.9.1. In the second step, an initial seed is selected to initiate the Metropolis-Hastings algorithm. This seed is selected through a re-sampling step such that it asymptotically follows the posterior distribution. Finally, in the last step, the distribution of the samples is corrected by means of MCMC: The Metropolis-Hastings algorithm proposed in [Tierney, 1994] is employed.

The validity of the accept/reject-ratio r_k employed in *Algorithm (7.3)* can be demonstrated as follows: As samples generated with BUS are used as candidate samples of the MCMC algorithm, the employed proposal distribution is $p_{\mathrm{trunc}}^{(c)}(\boldsymbol{\theta}|\mathcal{D})$. The accept/reject-ratio of the kth candidate samples is:

$$r_k = \min\left(1, \frac{\max\left(L\left(\boldsymbol{\theta}_{(k)}^{(c)}\middle|\mathcal{D}\right), c^{-1}\right)}{L\left(\boldsymbol{\theta}_{(k-1)}\middle|\mathcal{D}\right)}\right) \tag{7.33}$$

where $k \in \{2, \ldots, K\}$. The candidate sample is accepted, with probability r_k and rejected with probability $1 - r_k$. This Metropolis-Hastings algorithm has stationary distribution $p(\boldsymbol{\theta}|\mathcal{D})$ and, thus, the generated samples asymptotically follow the posterior distribution. To show that the stationary distribution actually is $p(\boldsymbol{\theta}|\mathcal{D})$, it is sufficient to show that the following equation holds:

$$r_k = \min\left(1, \frac{p\left(\boldsymbol{\theta}_{(k)}^{(c)}|\mathcal{D}\right)}{p\left(\boldsymbol{\theta}_{(k-1)}|\mathcal{D}\right)} \cdot \frac{p_{\text{trunc}}^{(c)}\left(\boldsymbol{\theta}_{(k-1)}|\mathcal{D}\right)}{p_{\text{trunc}}^{(c)}\left(\boldsymbol{\theta}_{(k)}^{(c)}|\mathcal{D}\right)}\right) \tag{7.34}$$

$$= \min\left(1, \frac{L\left(\boldsymbol{\theta}_{(k)}^{(c)}|\mathcal{D}\right)}{L\left(\boldsymbol{\theta}_{(k-1)}|\mathcal{D}\right)} \cdot \frac{L_{\text{trunc}}^{(c)}\left(\boldsymbol{\theta}_{(k-1)}|\mathcal{D}\right)}{L_{\text{trunc}}^{(c)}\left(\boldsymbol{\theta}_{(k)}^{(c)}|\mathcal{D}\right)}\right) \tag{7.35}$$

Four different cases can be distinguished:

Case (1): $L\left(\boldsymbol{\theta}_{(k)}^{(c)}|\mathcal{D}\right) \leq c^{-1}$ and $L\left(\boldsymbol{\theta}_{(k-1)}|\mathcal{D}\right) \leq c^{-1}$
We have $L\left(\boldsymbol{\theta}|\mathcal{D}\right) = L_{\text{trunc}}^{(c)}\left(\boldsymbol{\theta}|\mathcal{D}\right)$ and, consequently, both Eq. (7.33) and Eq. (7.35) become *one*.

Case (2): $L\left(\boldsymbol{\theta}_{(k)}^{(c)}|\mathcal{D}\right) > c^{-1}$ and $L\left(\boldsymbol{\theta}_{(k-1)}|\mathcal{D}\right) \leq c^{-1}$
As in the previous case, Eqs. (7.33) and (7.35) become *one*.

Case (3): $L\left(\boldsymbol{\theta}_{(k)}^{(c)}|\mathcal{D}\right) \leq c^{-1}$ and $L\left(\boldsymbol{\theta}_{(k-1)}|\mathcal{D}\right) > c^{-1}$
Eqs. (7.33) and (7.35) evaluate to $c^{-1}/L\left(\boldsymbol{\theta}_{(k-1)}|\mathcal{D}\right)$.

Case (4): $L\left(\boldsymbol{\theta}_{(k)}^{(c)}|\mathcal{D}\right) > c^{-1}$ and $L\left(\boldsymbol{\theta}_{(k-1)}|\mathcal{D}\right) > c^{-1}$
We have $L_{\text{trunc}}^{(c)}\left(\boldsymbol{\theta}|\mathcal{D}\right) = c^{-1}$. Thus, both Eq. (7.33) and Eq. (7.35) transform to $\min\left(1, L\left(\boldsymbol{\theta}_{(k)}^{(c)}|\mathcal{D}\right)/L\left(\boldsymbol{\theta}_{(k-1)}|\mathcal{D}\right)\right)$.

Consequently, Eqs. (7.33) and (7.34) are indeed equivalent and, thus, the posterior distribution is the stationary distribution of the Markov chain.

The burn-in period of the Markov chain in *step (3)* can be considered negligible for reasonably large K, as the initial seed asymptotically follows the stationary distribution of the chain. Note that the likelihood function needs to be evaluated only in the actual BUS simulation. Consequently, if the evaluation of the model behind the likelihood is computationally demanding, the computational overhead of the post-processing step presented as *Algorithm (7.3)* is negligible.

The principle of *Algorithm (7.3)* combined with rejection sampling (*Algorithm (7.1)*) is illustrated in Fig. 7.4. Contrary to plain *rejection sampling*, this algorithm does not produce K independent samples. Instead, the K generated posterior samples are dependent, because of the acceptance/rejection step in the Metropolis-Hastings algorithm (*steps (3b) and (3c)* in

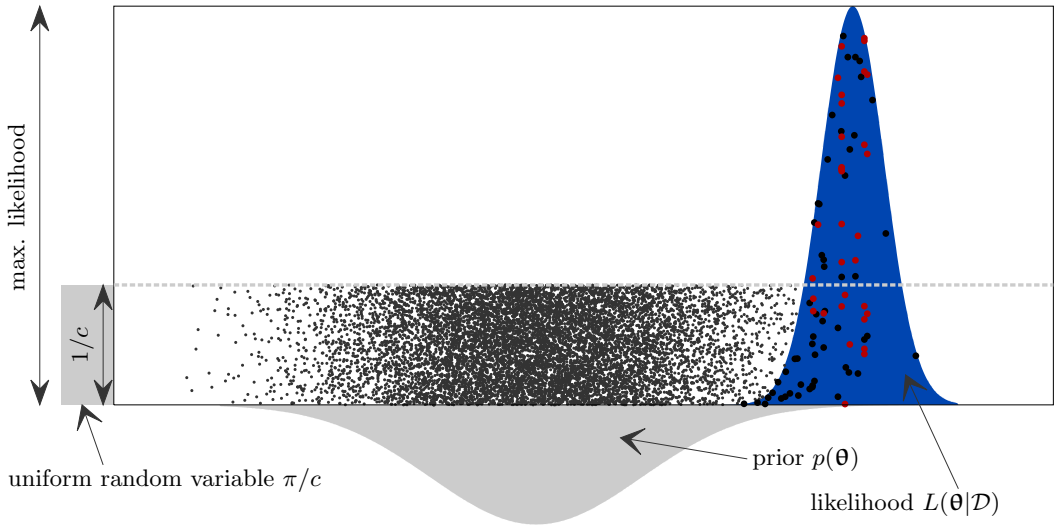


Figure 7.4: Illustration of the principle of *M-H rejection sampling* algorithm – *Algorithm (7.3)* combined with rejection sampling (*Algorithm (7.1)*). Note that only the *black* samples are independent; the *red* samples introduce a dependency (at least one *black* sample has the same $\boldsymbol{\theta}$ as a *red* sample) that decreases the efficiency of the sampling algorithm. For clarity, the ordinate of a sample $\boldsymbol{\theta}$ in Ω is selected randomly between *zero* and $L(\boldsymbol{\theta}|\mathcal{D})$.

Algorithm (7.3)). The smaller the corresponding acceptance rate, the larger is the induced dependency. A particular advantage of *Algorithm (7.3)* compared to *Algorithm (7.1)* is that *Algorithm (7.3)* can be applied without knowing $L_{\max}$. However, the selected c still has a considerable influence on the efficiency of the algorithm.

7.3.9.3 Numerical investigation

The performance of the example problems introduced in Section 7.2 is assessed for *Algorithm (7.3)* combined with rejection sampling (*Algorithm (7.1)*) and different values of $b \in (0, 1]$. The combination of *Algorithm (7.3)* with rejection sampling (*Algorithm (7.1)*) is referred to as *MH-RS* algorithm in the following (*MH-RS* stands for *Metropolis-Hastings rejection sampling*). The number of posterior samples generated per updating run is $K = 10^3$. The smallest value of b investigated in the studies is 10^{-4}. The results are presented in Figs. 7.5 – 7.7. Fig. 7.5 shows the statistics of the estimated evidence, Fig. 7.6 shows the statistics of the generated posterior samples of θ_1, and Fig. 7.7 shows the statistics of both $\mathrm{E}[n_{10^3}]$ and $\sigma[a_{10^3}]$. The data used to plot Figs. 7.5 – 7.7 was generated by solving the updating problem repeatedly, generating $K = 10^3$ posterior samples in each run. The number of times the problem was solved is $2 \cdot 10^5$, 400, $9 \cdot 10^3$ and $8 \cdot 10^4$ for *Example problems 1a*, *1b*, *2* and *3*, respectively. Note that some of the notation used in this section is introduced in Section 7.2.

First, we look at the statistics of the estimated evidence $\hat{c}_{\mathrm{E},10^3}$ as a function of b, presented

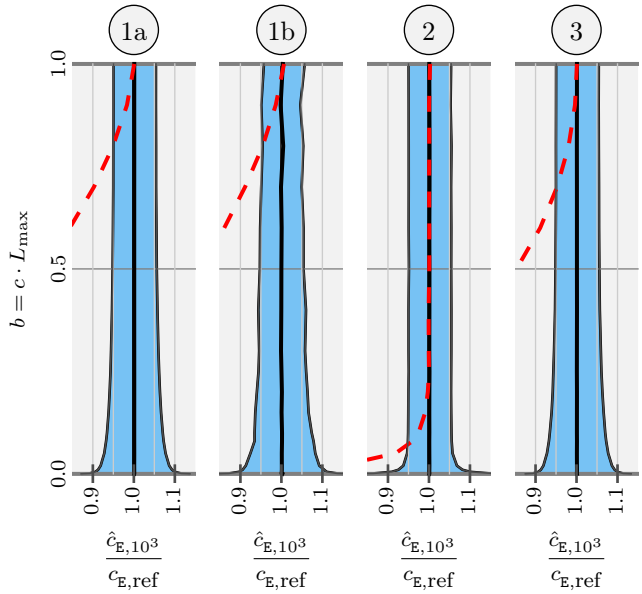


Figure 7.5: Statistics of the estimated evidence $\hat{c}_{\mathrm{E},10^3}$ divided by the reference value of the evidence $c_{\mathrm{E,ref}}$ for different values of $b = L_{\max} \cdot c$ and limit-state functions 1a, 1b, 2 and 3. The **thick black line** represents the average obtained with the *MH-RS* algorithm – as the average is 1.0 with good approximation, the estimate $\hat{c}_{\mathrm{E},10^3}$ can be considered unbiased. The highlighted area shows the 90% confidence interval of the estimated $\hat{c}_{\mathrm{E},10^3}/c_{\mathrm{E,ref}}$ computed with the *MH-RS* algorithm. The dashed red line shows the average obtained with standard rejection sampling; the bias in the estimated evidence clearly increases for decreasing b. The underlying data was generated by solving the updating problem repeatedly, generating $K = 10^3$ posterior samples in each run.

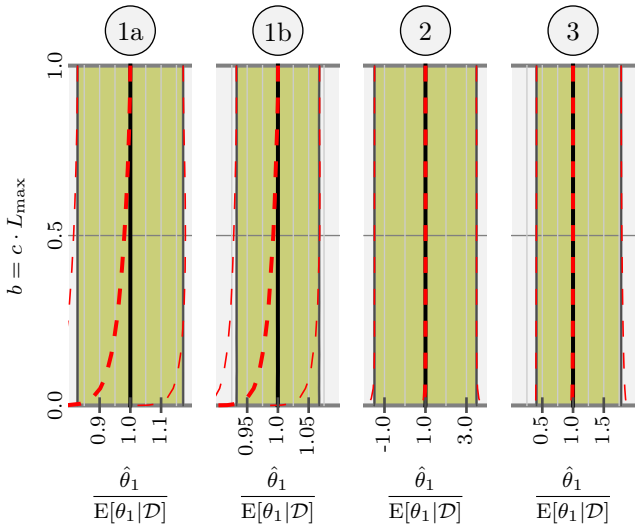


Figure 7.6: Statistics of the estimated posterior samples $\hat{\theta}_1$ divided by the reference mean of θ_1 for different values of $b = L_{\max} \cdot c$ and limit-state functions 1a, 1b, 2 and 3. The **thick black line** represents the average obtained with the *MH-RS* algorithm – as the average is 1.0 with good approximation, the sample mean can be considered unbiased. The highlighted area shows the 90% confidence interval obtained with the *MH-RS* algorithm. The dashed red lines show the average and confidence intervals obtained with standard rejection sampling. The underlying data was generated by solving the updating problem repeatedly, generating $K = 10^3$ posterior samples in each run.

in Fig. 7.5. Independent of the value of b, a bias in the estimate of the evidence cannot be observed. Additional to the mean, the 90% confidence interval of $\hat{c}_{\mathrm{E},10^3}/c_{\mathrm{E,ref}}$ is shown. For $b \in [0.5, 1]$, the influence of b on the interval can be considered negligible, independent of the example problem. However, this does not imply that it is irrelevant whether the analysis is performed with $b = 1$ or, for example, with $b = 0.5$. The reason is that although the discussed confidence interval remains stable, the computational cost decreases with b. The computational cost is expressed in terms of the average number of prior samples $\mathrm{E}[n_{10^3}]$ required to perform the analysis (see left part of Fig. 7.7). For $b \in [0.5, 1]$ the value of $\mathrm{E}[n_{10^3}]$ is – with good approximation – proportional to b. Additional to the mean and 90% confidence interval of quantity $\hat{c}_{\mathrm{E},10^3}/c_{\mathrm{E,ref}}$ obtained by means of the *MH-RS* algorithm, the mean of $\hat{c}^{(c)}_{\mathrm{E},10^3}/c_{\mathrm{E,ref}}$ computed with standard rejection sampling (*Algorithm (7.1)*) and $c^{-1} < L_{\max}$ is shown in Fig. 7.5, where $\hat{c}^{(c)}_{\mathrm{E},K}$ denotes the evidence estimated with standard rejection sampling and scaling constant c. The results clearly show that the estimate of the evidence $\hat{c}^{(c)}_{\mathrm{E},10^3}$ obtained with standard rejection sampling and $c^{-1} < L_{\max}$ underestimates the true evidence $c_{\mathrm{E,ref}}$ of the example problem.

Fig. 7.6 shows that the estimates of both the mean and the 90% confidence interval of the posterior samples $\hat{\theta}_1$ obtained with the *MH-RS* algorithm are unbiased, independent of the choice of b. The same cannot be said about standard rejection sampling and $c^{-1} < L_{\max}$. In this case, the bias in the mean and the deviation from the 90% confidence interval of the

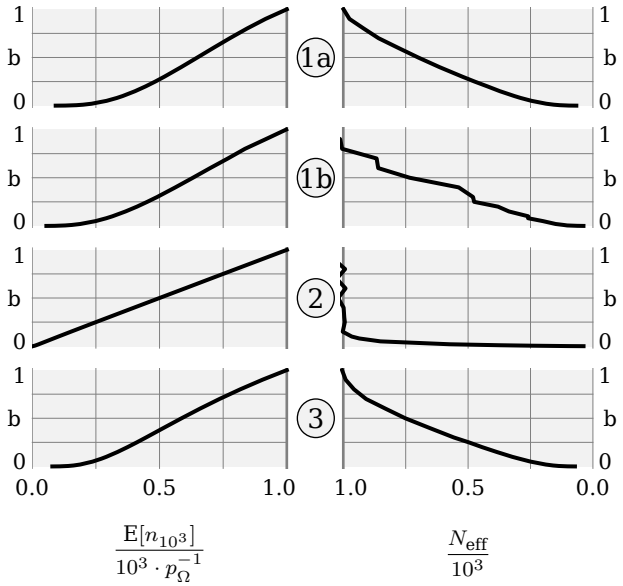


Figure 7.7: The average number of model calls (left side of plot, x-axis) and the standard deviation of the estimate a_{10^3} (right side of plot, x-axis) are shown for $b \in [0, 1]$ (y-axis of sub-plots) and the *MH-RS* algorithm. The quantity b is plotted on the y-axis of the sub-plots and ranges from *zero* to *one*. In horizontal direction, the figure is split in two parts: (**left-part**) On the left-hand side of the figure the average number $\mathrm{E}[n_{10^3}]$ of prior samples required to solve the problem is divided by the average number of prior samples needed in standard rejection sampling; i.e., by $10^3/p_\Omega$. For $b \in [0, 1]$, this quantity is within $[0, 1]$; it measures how many prior samples were needed to solve the problem compared to standard rejection sampling. (**right-part**) On the right-hand side of the figure the effective number of independent posterior samples (N_{eff}) divided by the total number (10^3) of posterior samples is shown.

samples θ_1 generated with the *MH-RS* algorithm increases with decreasing b. This effect is more dominant in *Example problems 1a* and *1b* than in *Example problems 2* and *3*.

The samples produced by the standard rejection sampling are independent, whereas the samples generated with *Algorithm (7.3)* and $b < 1$ are dependent. The dependency of the samples increases with decreasing b. This effect is illustrated in the right part of Fig. 7.7: The effective number of independent posterior samples decreases with b. Thus, the efficiency of *Algorithm (7.3)* with respect to the generated posterior samples decreases with decreasing b. However, the computational cost to generate 10^3 (dependent) posterior samples decreases also with decreasing b – see left part of Fig. 7.7.

7.4 aBUS – adaptive BUS-SuS

This section contains material originally published in [Betz et al., 2017]. Some passages and figures are directly taken from the mentioned reference.

7.4.1 Introduction

Contrary to the standard BUS algorithm presented in Section 7.3.8.2, an algorithm is proposed that does not require the constant c as input. The proposed algorithm is based on the BUS variant originally proposed in [Betz et al., 2014b] that adaptively learns the constant c. The idea from [Betz et al., 2014b] is extended to improve the efficiency of the method. Due to the particular formulation of the limit-state function of the BUS problem given in Eq. (7.17), the method can be considerably simplified, as it requires only minimal modifications of the original *BUS-SuS* algorithm. The proposed method is termed *aBUS*, in accordance with [Betz et al., 2014b].

For the limit-state function introduced in Eq. (7.16) we have: The ith intermediate failure domain Z_i can be expressed as the set of all $\boldsymbol{\theta}$ and π for which the following inequality holds (see Eq. (7.18)):

$$\pi \leq c \cdot L(\boldsymbol{\theta}|\mathcal{D}) \cdot \exp(h_i) \tag{7.36}$$

The particular advantage of this limit-state formulation is as follows. The inequality in Eq. (7.36) can be equivalently stated as:

$$\pi \leq \frac{L(\boldsymbol{\theta}|\mathcal{D}) \cdot \exp(h_i + \delta)}{c^{-1} \cdot \exp(\delta)} \tag{7.37}$$

where $\delta \in \mathbb{R}$ is an arbitrary scalar value. From Eq. (7.37) it follows that the intermediate failure domain Z_i associated with scaling constant c and threshold level h_i can be equivalently expressed by scaling constant c^* and threshold level h^* if h^* is chosen as:

$$h^* = h_i + \ln\left(\frac{c}{c^*}\right) = h_i - \ell + \ell^* \tag{7.38}$$

where $\ell = -\ln(c)$ and $\ell^* = -\ln(c^*)$. Consequently, if we always modify the current threshold value of Subset Simulation after changing the value of c or ℓ, the change in c or ℓ does not affect the distribution of the current samples (i.e., the samples that are in domain Z_i).

7.4.2 Proposed modifications to the basic BUS-SuS algorithm

The first step of the adaptive algorithm is the same as the one of standard *BUS-SuS*; i.e., it consists in drawing K samples from the prior distribution. The likelihood of each sample is evaluated and stored. However, before the value of the first threshold level h_i can be selected, a value has to be assigned to the BUS scaling constant c: The constant c^{-1} is set equal to the value of the largest likelihood within the generated set of samples. Thereafter, each iteration is performed in accordance with *BUS-SuS*: The value of each intermediate threshold is selected based on the limit-state function realizations, and MCMC sampling is performed to generate samples conditional on the current intermediate failure domain. At the end of a Subset level, it is checked whether a likelihood larger than the current value of c^{-1} was observed. If so, the current value of c^{-1} is adapted such that it matches the largest likelihood observed and the value of h_i is modified according to Eq. (7.38). Note that c^{-1} can only increase and, thus, the threshold h_i increases as well. The iteration over the subset levels is performed until the current threshold value h_i is *zero* at the end of a subset level. The thus obtained intermediate failure domains are clearly nested – which is a prerequisite for the application of SuS. The evidence is estimated based on the last value of c^{-1} at the end of Subset Simulation according to Eq. (7.6); i.e. with c^{-1} equal to the value of the largest likelihood observed during the simulation.

The general structure of the proposed algorithm is given in the following – changes compared to *Algorithm (7.2)* are highlighted. The following algorithm employs the log-transform of the likelihood $\ln L(\boldsymbol{\theta}|\mathcal{D})$, which is beneficial from a numerical point of view compared to working with the likelihood function directly.

Algorithm 7.4. *aBUS – adaptive BUS-SuS*:
As input the algorithm requires:

- K, the total number of samples to draw from the posterior distribution.
- p_t, the probability of the intermediate subsets. p_t needs to be selected such that $p_\mathrm{t} \cdot K$ is an integer number.

The algorithm evaluates the evidence $c_{\mathrm{E}|\mathcal{M}}$ and returns K unweighted but dependent posterior samples $\boldsymbol{\theta}_{(k)}$ with $k = 1, \ldots, K$.

1. Draw K samples $[\boldsymbol{\theta}_{(0,k)}, \pi_{(0,k)}]$, with $k = 1, \ldots, K$, from the prior distribution.
2. Initialize $i = 0$ and $h_0 = \infty$.
3. Set $\ell = \max\left(\left\{\ln L\left(\boldsymbol{\theta}_{(0,k)}|\mathcal{D}\right)\right\}_{k=1}^{K}\right)$, where ℓ is defined as in Eq. (7.17).
4. **while** $(h_i > 0)$ **do**:
   (a) Increase counter i by *one*: $i = i + 1$.

(b) Select the threshold level h_i:

i. Sort the K samples $\left\{[\boldsymbol{\theta}_{(i-1,k)}, \pi_{(i-1,k)}]\right\}_{k=1}^{K}$ with respect to the value of $g_{\mathrm{l}}(\boldsymbol{\theta}_{(i-1,k)}, \pi_{(i-1,k)})$ in ascending order.
Note that $g_{\mathrm{l}}(\cdot,\cdot)$ as defined in Eq. (7.17) is used.

ii. Set $h_i = \frac{g_{\mathrm{l}}\left(\boldsymbol{\theta}_{(i-1,p_{\mathrm{t}}\cdot K)}\right)+g_{\mathrm{l}}\left(\boldsymbol{\theta}_{(i-1,p_{\mathrm{t}}\cdot K+1)}\right)}{2}$; i.e., set h_i as the p_{t}-percentile of the ordered set.

iii. Select n as the number of samples in $\left\{[\boldsymbol{\theta}_{(i-1,k)}, \pi_{(i-1,k)}]\right\}_{k=1}^{K}$ with $g_{\mathrm{l}}(\boldsymbol{\theta}_{(i-1,k)}, \pi_{(i-1,k)}) \leq \max(h_i, 0)$.

iv. **if** $(h_i < 0)$ **then**: Set $h_i = 0$, and $p_i = \frac{n}{K}$.
else: Set $p_i = p_{\mathrm{t}}$.

(c) Generate samples conditional on domain Z_i:

i. Randomize the ordering of the samples in the set $\left\{[\boldsymbol{\theta}_{(i-1,k)}, \pi_{(i-1,k)}]\right\}_{k=1}^{n}$; i.e., thereafter, the n samples are no longer ordered.

ii. Generate the samples $[\boldsymbol{\theta}_{(i,k)}, \pi_{(i,k)}]$ by means of n Markov chains; e.g., by means of the CS algorithm (*Algorithm (3.6)*) applied inside *Algorithm (3.3)*. The n samples $\left\{[\boldsymbol{\theta}_{(i-1,k)}, \pi_{(i-1,k)}]\right\}_{k=1}^{n}$ are already within Z_i and are used as seeds for the n Markov chains. The length of each Markov chain is K/n. Thus, the total number of MCMC samples generated in one level is $K - n$.
Algorithm (3.9) is applied during the MCMC sampling to adopt the spread of the MCMC proposal distribution.

(d) Update the value of the scaling constant:

i. Set $\ell_{\mathrm{new}} = \max\left(\ell, \left\{\ln L\left(\boldsymbol{\theta}_{(i,k)}|\mathcal{D}\right)\right\}_{k=1}^{K}\right)$.

ii. Modify $h_i = h_i - \ell + \ell_{\mathrm{new}}$.

iii. Set $\ell = \ell_{\mathrm{new}}$.

(e) Decrease dependence of the K samples:
For $(k = 1, \ldots, K)$ **do**:

i. Draw $\tilde{\pi}$ as a sample from a uniform distribution with support $\left[0, \min\left(1, \exp\left(\ln L(\boldsymbol{\theta}_{(i,k)}|\mathcal{D}) - \ell + h_i\right)\right)\right]$.

ii. Set $[\boldsymbol{\theta}_{(i,k)}, \pi_{(i,k)}] = [\boldsymbol{\theta}_{(i,k)}, \tilde{\pi}]$

5. Set $N = i$

6. Estimate $p_\Omega = \prod_{i=1}^{m} p_i$

7. Evaluate the evidence $c_{\mathrm{E}|\mathcal{M}} = p_\Omega \cdot \exp(\ell)$

In a conventional reliability problem, we typically have very limited knowledge about the shape of the failure domain. Contrary to that, in reliability problems that stem from BUS, we know that for a sample $[\boldsymbol{\theta}, \pi]$ with associated likelihood $L(\boldsymbol{\theta}|\mathcal{D})$ each π that is smaller or equal than $cL(\boldsymbol{\theta}|\mathcal{D})$ means that the sample $[\boldsymbol{\theta}, \pi]$ is a posterior sample and is located within the failure domain – and vice versa. In addition to that, we know that for fixed $\boldsymbol{\theta}$, all π that maintain $\pi \leq cL(\boldsymbol{\theta}|\mathcal{D})$ are distributed uniformly on the interval $[0, cL(\boldsymbol{\theta}|\mathcal{D})]$. At the

intermediate levels of SuS, a sample $[\boldsymbol{\theta}, \pi]$ is in domain Z_i if $\pi \leq cL(\boldsymbol{\theta}|\mathcal{D})\exp(h_i)$. Thus, we can easily modify the component π of sample $[\boldsymbol{\theta}, \pi]$. This is what is done in *Step 4(e)* of *Algorithm (7.4)*.

Step 4(e) in *Algorithm (7.4)* is a re-sampling strategy that comes at – practically – no additional cost (because it does not involve additional evaluations of the likelihood function). Its particular appeal is: In the MCMC sampling procedure each sample that is rejected means that an existing sample is duplicated. For example, with a target acceptance rate of $\alpha_{\text{opt}} = 0.44$ we aim at rejecting 56% percent of all proposed samples. Thus, a considerable number of the generated samples will not be unique. *Step 4(e)* distributes the π-components of all samples with the same $\boldsymbol{\theta}$ uniformly on the interval $\left[0, \min\left(1, c \cdot L(\boldsymbol{\theta}_{(i,k)}|\mathcal{D}) \cdot \exp(h_i)\right)\right]$. This step is added to decrease the dependency of the generated MCMC samples and, consequently, to increase the overall performance of SuS.

7.4.3 Comments on the final value of c^{-1} in aBUS and $L_{\max}$

The final value of $c^{-1} = \exp(\ell)$ in *aBUS*, corresponds to the largest likelihood observed during the simulation. Consequently, we have $\exp(\ell) \leq L_{\max}$. Asymptotically, $\exp(\ell)$ approaches $L_{\max}$ for large K. However, for finite K, $\exp(\ell)$ is very likely smaller than $L_{\max}$. Therefore, *aBUS* works with values of c^{-1} that are on average smaller than $L_{\max}$. However, this does not prevent *aBUS* from producing samples that follow the posterior distribution, as is explained in the following.

The difference to rejection sampling (and BUS approaches in general) performed with $c^{-1} \leq L_{\max}$ is that rejection sampling works with a prespecified fixed c^{-1}, whereas *aBUS* does not work with a fixed c^{-1}. The final $c^{-1} = \exp(\ell)$ in *aBUS* varies; it is a stochastic quantity that is equivalent to the largest likelihood value observed during the entire simulation. Let $P_L\left(L(\boldsymbol{\theta})|\mathcal{D}\right)$ denote the cumulative distribution function of likelihood values evaluated for samples of the posterior distribution. If aBUS produces posterior samples, the quantity c^{-1} is a realization from a distribution that has CDF $P_{c^{-1}}(c^{-1}|\mathcal{D}) = \left(P_L\left(c^{-1}|\mathcal{D}\right)\right)^K$.

In the following, we show that the CDF $P_L\left(L(\boldsymbol{\theta})|\mathcal{D}\right)$ can be approximated well with posterior samples generated with *aBUS*. This is demonstrated numerically by means of *Example problem 2* defined in Section 7.2: In this example problem, the probability that we will observe a likelihood larger than $0.8 \cdot L_{\max}$ in a set of 10^3 independent posterior samples is $1 \cdot 10^{-4}$ (see Table 7.1). Therefore, it is unlikely that the value of c^{-1} in *aBUS* is close to the theoretical $L_{\max}$ for 10^3 generated samples.

The decisive parameter in *aBUS* is the number K of samples employed in each level of SuS. For the method to generate posterior samples, K must be selected large enough such that the final K samples can propagate in the entire domain Ω. The bulk of the generated posterior

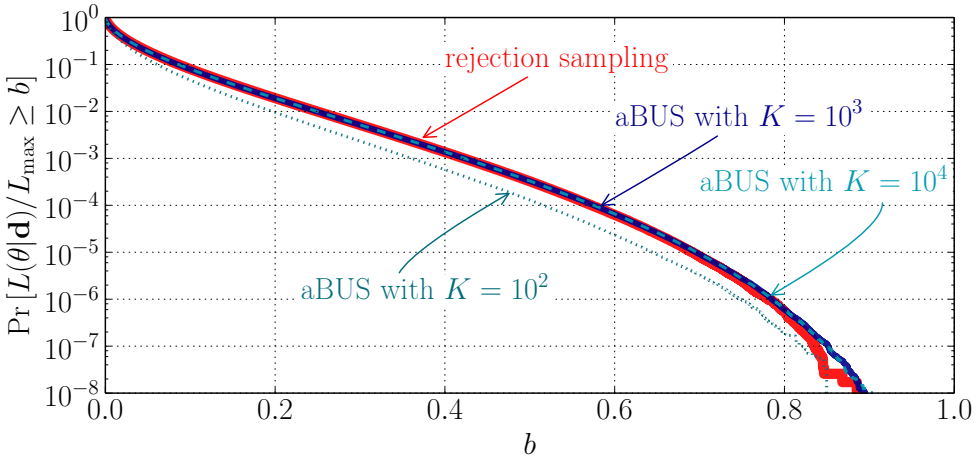


Figure 7.8: The posterior probability $\Pr[L(\boldsymbol{\theta}|\mathcal{D})/L_{\max} \geq b]$ is plotted for different values of b (for *Example problem 2*). Results are shown for posterior samples obtained by means of repeated runs of *aBUS* with $K = 10^2$, $K = 10^3$ and $K = 10^4$, where K denotes the number of samples used in each level of Subset Simulation. The reference solution is evaluated numerically by means of 10^9 statistically independent posterior samples obtained with rejection sampling.

samples will be in the "high probability region" of the posterior distribution – which does not necessarily mean that many samples will fall in the region that has large likelihood.

We investigate the distribution of likelihood values associated with the generated posterior samples for *Example problem 2*. The Bayesian inference problem is solved with *aBUS* for $K = 10^2$, $K = 10^3$ and $K = 10^4$ samples per subset level. The reference distribution was obtained by means of rejection sampling and $c = L_{\max}^{-1}$. The posterior probability $\Pr[L(\boldsymbol{\theta}|\mathcal{D})/L_{\max} \geq b]$ obtained with *aBUS* and rejection sampling is depicted in Fig. 7.8 for different values of b and K; the definition of b is according to Section 7.2. For $K = 10^2$, the resulting posterior distribution of the likelihood values deviates from the reference solution. However, already for $K = 10^3$, the resulting posterior distribution matches the reference solution well. Therefore, even if *aBUS* selects c^{-1} for this example problem on average considerably smaller than $L_{\max}$: The distribution of the likelihood values associated with the generated posterior samples is not biased – provided that K is selected large enough. The performance of *aBUS* with respect to K is investigated in detail in Section 7.4.4.

Note that the average number of likelihood function calls in *aBUS* is at most as large as in standard *BUS-SuS*. This is because $c_N^{-1} \leq L_{\max}$ in *aBUS*, whereas for *BUS-SuS* $c^{-1} \geq L_{\max}$ is required. Thus, as p_Ω is proportional to c^{-1}, the average total number of subset levels required to solve the inference problem with *aBUS* is at most as large as the one of *BUS-SuS*. Additional to that, if the limit-state formulation given in Eq. (7.16) is used, the distribution of samples produced at the intermediate levels of SuS is invariant to the selected c – compare Section 7.4.1 and Eq. (7.37) in particular. As a consequence, the *aBUS* algorithm should be preferred over standard *BUS-SuS* even if the theoretical maximum of the likelihood function

is known in advance.

7.4.4 Numerical investigation of the performance of aBUS

7.4.4.1 Some notes on the notation employed

The notation introduced in Section 7.2 is used. Additional to that, the following quantitative measures are defined:

- $\text{bias}\left[\hat{c}_{\text{E},K}\right]$: the bias in the estimated evidence.

$$\text{bias}\left[\hat{c}_{\text{E},K}\right] = \left| \frac{\text{E}\left[\hat{c}_{\text{E},K}\right] - c_{\text{E}|\mathcal{M}}}{c_{\text{E}|\mathcal{M}}} \right| \tag{7.39}$$

- $\text{CoV}\left[\hat{c}_{\text{E},K}\right]$: the coefficient of variation of the estimated evidence.

$$\text{CoV}\left[\hat{c}_{\text{E},K}\right] = \frac{\sigma\left[\hat{c}_{\text{E},K}\right]}{\text{E}\left[\hat{c}_{\text{E},K}\right]} \tag{7.40}$$

- $\text{bias}\left[a_K\right]$: the bias in the estimated posterior mean of θ_1.

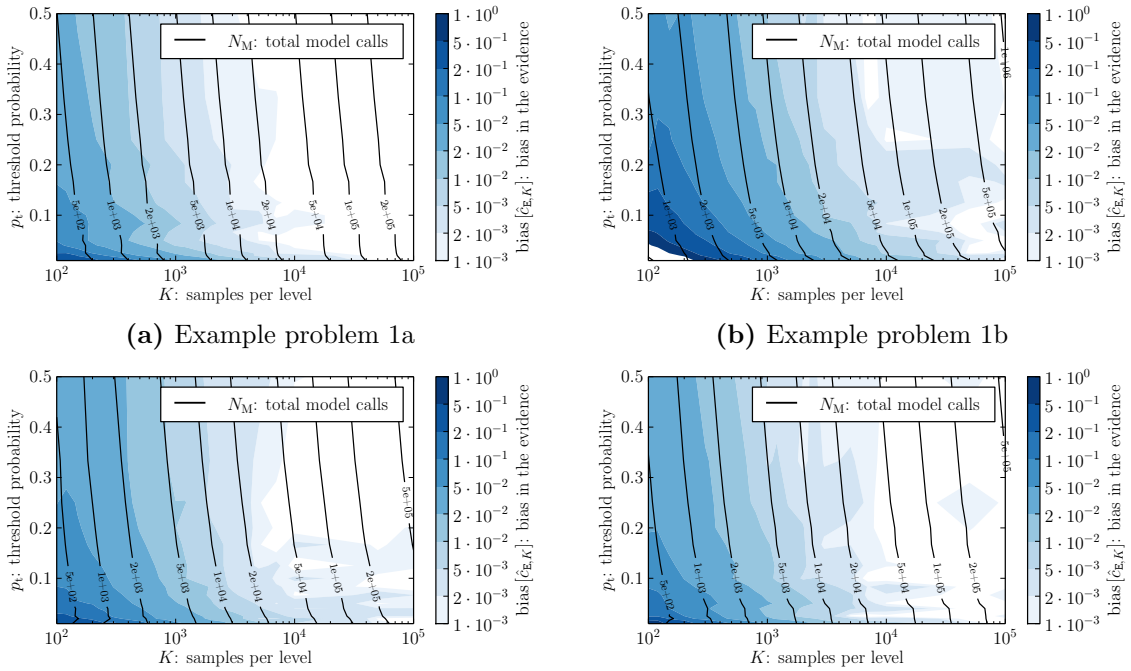
$$\text{bias}\left[a_K\right] = \left| \frac{\text{E}\left[a_K\right] - \text{E}[\theta_1|\mathcal{D}]}{\text{E}[\theta_1|\mathcal{D}]} \right| \tag{7.41}$$

- $\text{bias}\left[s_K\right]$: the bias in the estimated posterior standard deviation of θ_1.

$$\text{bias}\left[s_K\right] = \left| \frac{\text{E}\left[s_K\right] - \sigma[\theta_1|\mathcal{D}]}{\sigma[\theta_1|\mathcal{D}]} \right| \tag{7.42}$$

7.4.4.2 Performance of aBUS for different p_t and K

The performance of *aBUS* is assessed for different p_t and K. The probability of the intermediate subsets p_t is analyzed for values within $[1\%, 50\%]$. The number K of samples per level is modified between 10^2 and 10^5. The four example problems already introduced in Section 7.2 are investigated. The aim is to determine which values of p_t lead to a (near-)optimal performance for the investigated example problems, where optimality is measured with respect to the number N_M of total required model calls; i.e. N_M is the total number of likelihood evaluations in Subset Simulation. For the MCMC sampling in the subset levels, the CS algorithm (*Algorithm (3.6)* applied inside *Algorithm (3.3)*) is employed. The spread of the proposal distribution is modified during the simulation as described in *Algorithm (3.9)*, with a target acceptance rate of $\alpha_\text{opt} = 0.44$.



(a) Example problem 1a

(b) Example problem 1b

(c) Example problem 2

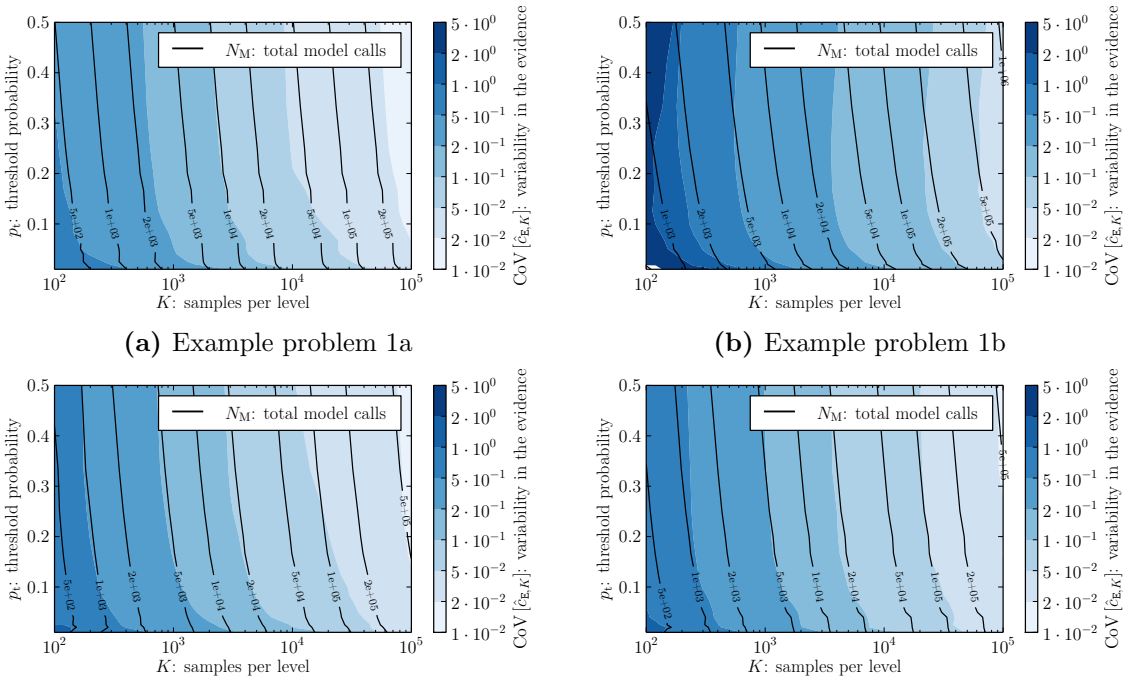
(d) Example problem 3

Figure 7.9: Bias in the evidence estimated with *aBUS* for different p_t and K.

First, we look at the bias in the evidence estimated with *aBUS* (by means of measure bias $[\hat{c}_{E,K}]$ introduced in Eq. (7.39)). The results for the four example problems are shown in Fig. 7.9. The bias in the evidence decreases with an increasing number of samples per level in all investigated example problems. We observe that $p_t = 10\%$ is clearly not an optimal choice. Especially for $K < 10^3$, the bias is smaller for large p_t than for small p_t. An intermediate probability p_t between 20% and 40% is a good choice for all investigated problems. Among all investigated example problems, the largest bias is observed in *Example problem 1b*; the smallest bias is observed in *Example problem 1a*. This suggests that the bias in the evidence computed with *aBUS* (and probably *BUS-SuS* in general) increases with an increasing number of subset levels.

Overall, the bias in the estimated evidence of *aBUS* is, however, negligible compared to the coefficient of variation of the estimate. The coefficient of variation CoV $[\hat{c}_{E,K}]$ in the estimated evidence is depicted in Fig. 7.10. The CoV $[\hat{c}_{E,K}]$ decreases with an increasing number of samples per level. For p_t between 10% and 30%, *aBUS* performs robustly with respect to a fixed number N_M of total model calls in all investigated example problems.

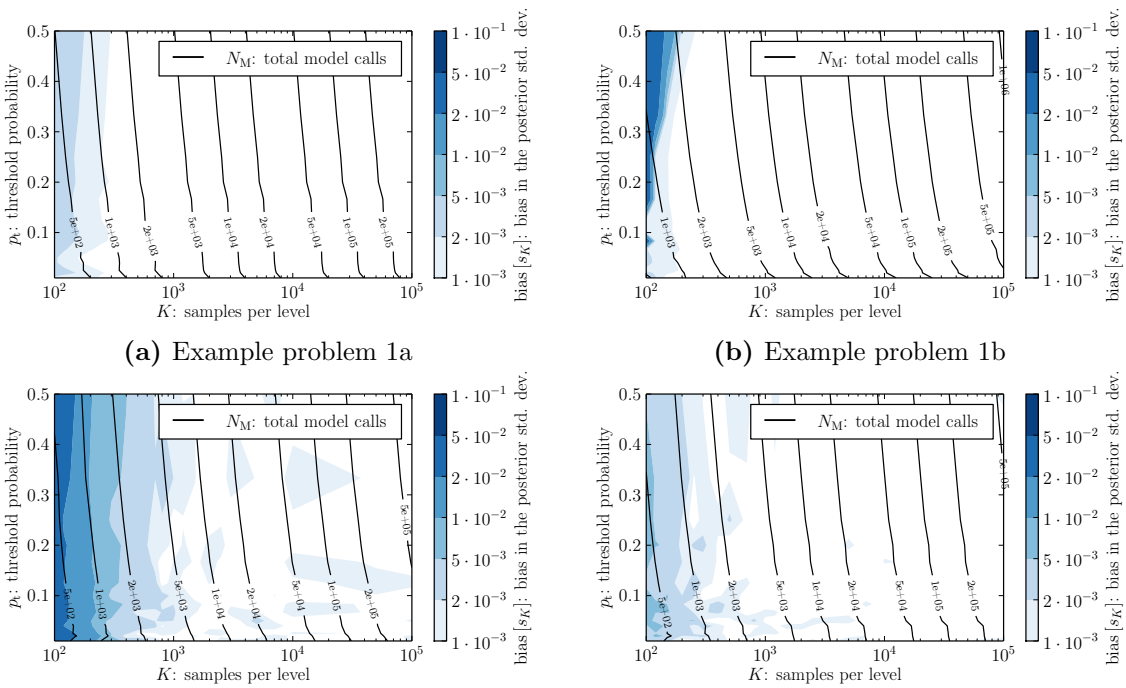
Next, we look at the mean and standard deviation of the posterior samples produced with *aBUS*. The bias in the estimated posterior mean and standard deviation is depicted in Fig. 7.11 and Fig. 7.12, respectively. For $K \geq 5 \cdot 10^2$, the bias in both mean and standard deviation is smaller than 0.5% and is, thus, considered negligible.



(a) Example problem 1a **(b)** Example problem 1b

(c) Example problem 2 **(d)** Example problem 3

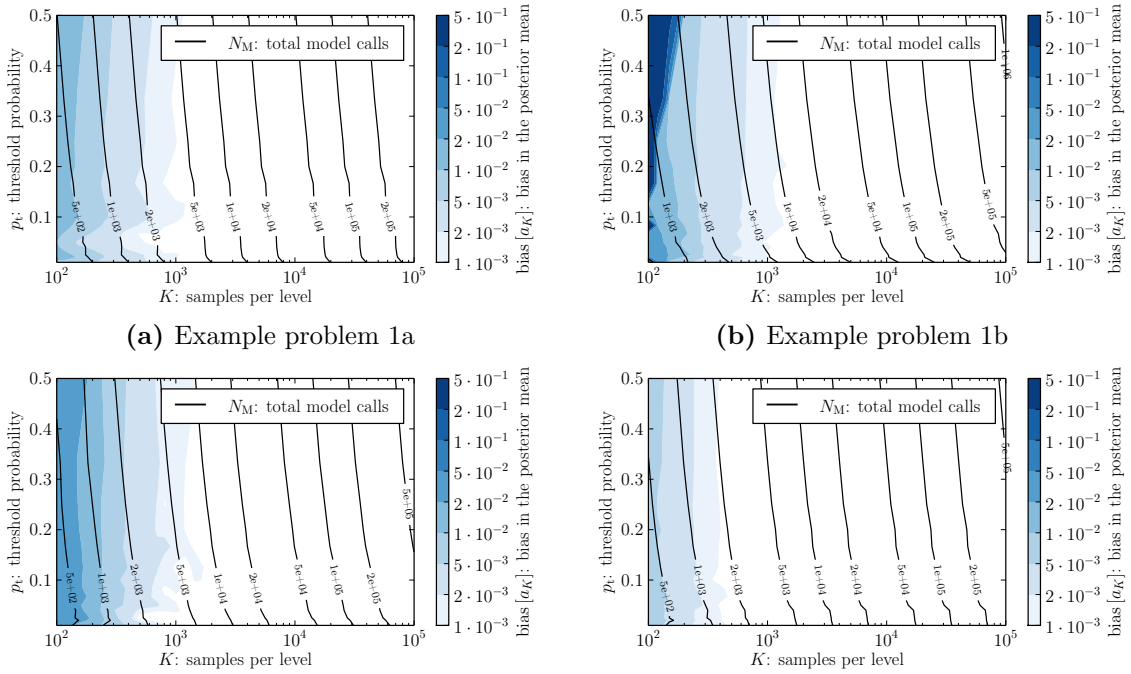
Figure 7.10: Coefficient of variation of the evidence estimated with *aBUS* for different p_t and K.



(a) Example problem 1a **(b)** Example problem 1b

(c) Example problem 2 **(d)** Example problem 3

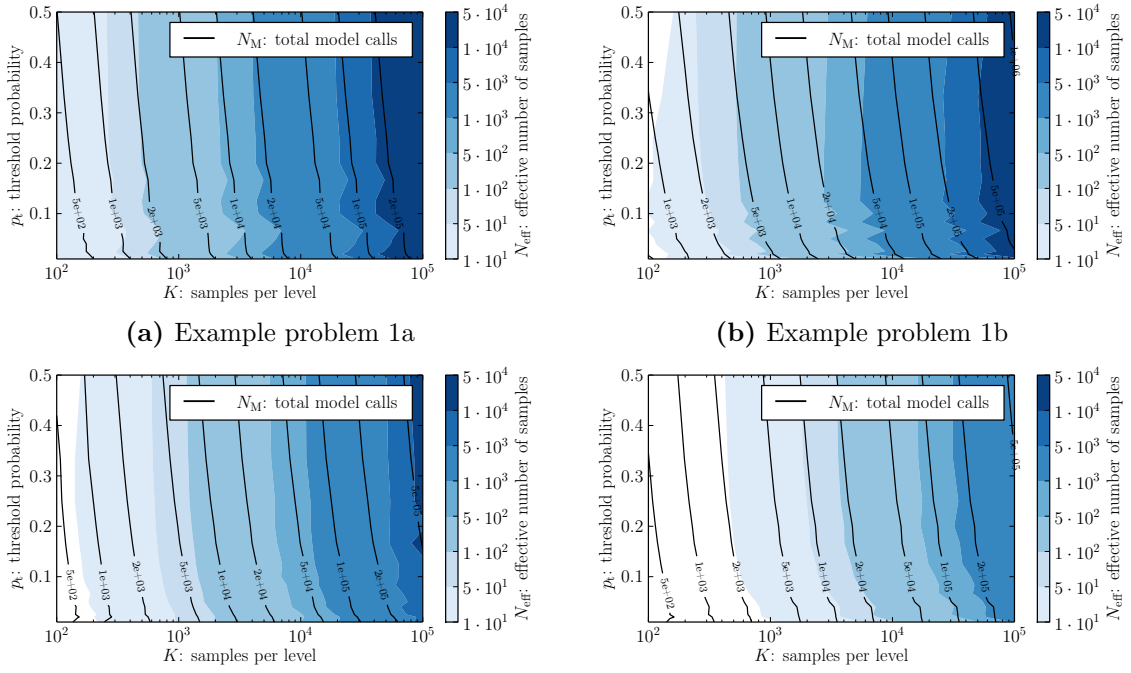
Figure 7.11: Bias in the mean of posterior samples generated with *aBUS* for different p_t and K.



(a) Example problem 1a **(b)** Example problem 1b

(c) Example problem 2 **(d)** Example problem 3

Figure 7.12: Bias in the standard deviation of posterior samples generated with *aBUS* for different p_t and K.



(a) Example problem 1a **(b)** Example problem 1b

(c) Example problem 2 **(d)** Example problem 3

Figure 7.13: Number N_{eff} of effectively independent samples obtained with *aBUS* for different p_t and K.

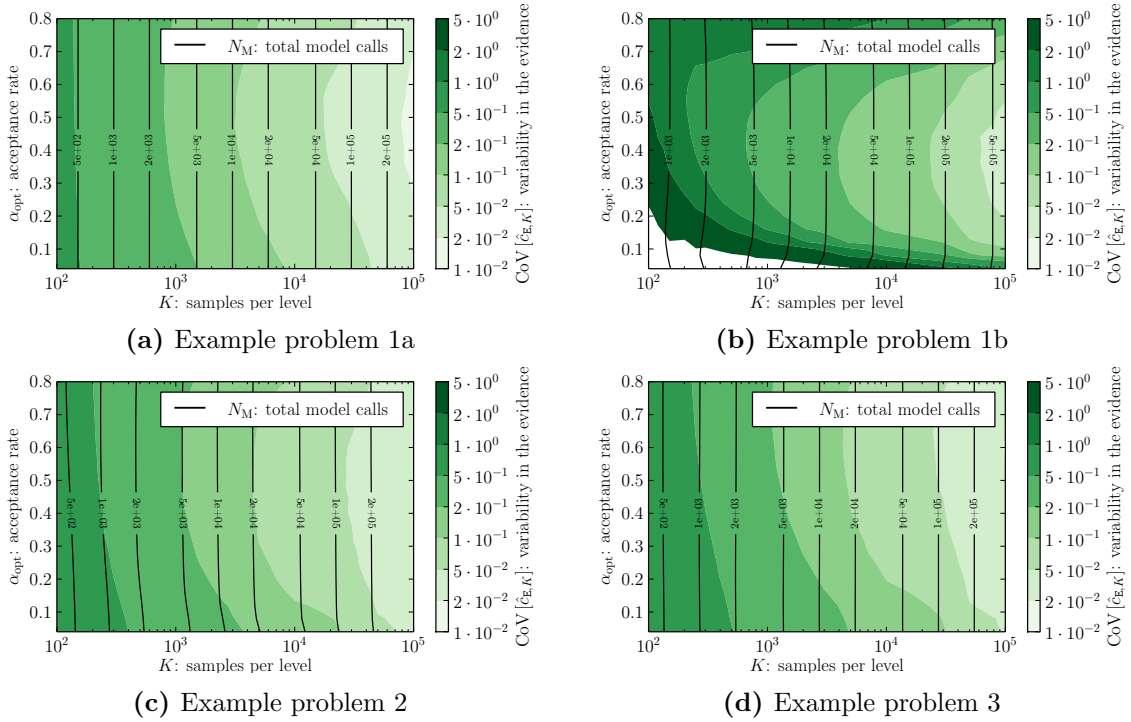
Finally, we look at the number N_{eff} of effectively independent samples in the generated set of K posterior samples. The results are shown in Fig. 7.13 for the four example problems. N_{eff} increases with increasing K. For a fixed number N_{M} of total model calls, *aBUS* exhibits the best performance for $p_{\text{t}} = 10\%$. However, N_{eff} is always considerably smaller than K: For $K = 10^3$ and $p_{\text{t}} = 10\%$ we obtain only 210, 150, 70 and 20 efficiently independent posterior samples in *Example problem 1a*, *1b*, *2* and *3*, respectively. In particular, $N_{\text{eff}} \approx 20$ in *Example problem 3* is a relatively small value. The poor performance in this example problem can be attributed to the bimodal shape of the posterior distribution: The standard deviation of quantity a_K that governs N_{eff} (see Eq. (7.2)) is relatively large in this example problem, because it is difficult for the intermediate samples in Subset Simulation to alternate between the two modes. If the fraction of samples in the separate modes at initial subset levels is amiss, this error will most probably propagate at the higher levels of Subset Simulation. However, *Example problem 3* demonstrates also the flexibility of *BUS-SuS* based approaches: They are able to produce posterior samples even if the target distribution is multi-modal.

To summarize the findings obtained in this section: The potential bias in the evidence estimated with *aBUS* is negligible compared to the variability of the estimate. Furthermore, the bias in the mean and standard deviation of posterior samples produced with *aBUS* is insignificant for reasonable sample sizes. For the investigated example problems, $K \geq 5 \cdot 10^2$ was large enough. However, as a general rule of thumb, it is recommended to use at least 10^3 samples per level in Subset Simulation. Therefore, the parameter p_{t} of SuS should be selected such that for a given number K of posterior samples, N_{eff} is maximized and CoV $[\hat{c}_{\text{E},K}]$ is minimized. For the investigated example problems, a near-optimal performance can be achieved with $p_{\text{t}} = 10\%$.

7.4.4.3 Performance of aBUS for different α_{opt} and K

In this study, the target acceptance rate α_{opt} of *aBUS* and the number K of samples per level is modified: α_{opt} is changed between 0.04 and 0.80, and K is modified between 10^2 and 10^5. As the bias of *aBUS* in the estimated evidence, the posterior mean and the posterior standard deviation was found to be negligible in the previous study (Section 7.4.4.2), we only investigate the performance in terms of CoV $[\hat{c}_{\text{E},K}]$ and N_{eff}. Again, we assess the performance of *aBUS* for combinations of α_{opt} and K that result in the same number N_{M} of total likelihood evaluations during Subset Simulation. In this study, the total number of required likelihood evaluations is approximately proportional to K. The probability of the intermediate subsets p_{t} is kept constant; p_{t} is set to 10%.

The coefficient of variation of the evidence estimated with *aBUS* is shown in Fig. 7.14 for different α_{opt} and K. For fixed N_{M}, a comparatively good performance is achieved for all investigated example problems if α_{opt} is selected between 0.4 and 0.6; where *Example problem*



(a) Example problem 1a **(b)** Example problem 1b

(c) Example problem 2 **(d)** Example problem 3

Figure 7.14: Coefficient of variation of the evidence estimated with *aBUS* for different α_{opt} and K.

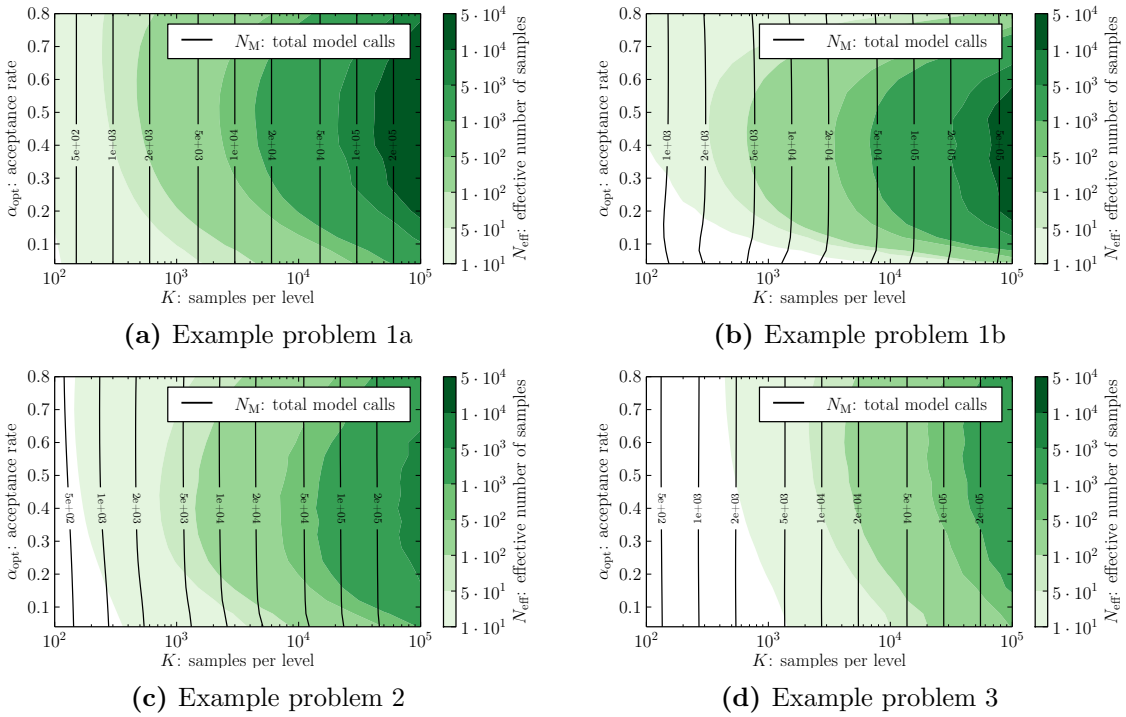
1b favors slightly smaller α_{opt} for large K and *Example problem 3* favors slightly larger α_{opt}. For *Example problem 1b* that has the smallest p_Ω amongst all investigated example problems, the dependency of CoV $[\hat{c}_{\text{E},K}]$ on α_{opt} is more pronounced than in the other problems.

The number N_{eff} of effectively independent posterior samples in the generated set of K posterior samples is depicted in Fig. 7.15. For *Example problems 1a*, *1b* and *2*, *aBUS* exhibits a near-optimal performance for $0.3 \leq \alpha_{\text{opt}} \leq 0.5$ (with respect to N_{M} fixed). For *Example problem 3*, a slightly better performance is achieved for larger values of α_{opt}; i.e. for α_{opt} selected around 0.6. Again, it is evident that for *Example problem 3*, *aBUS* produces a relatively small N_{eff}.

In a nutshell, the choice of $\alpha_{\text{opt}} = 0.44$ proposed in [Papaioannou et al., 2015] for Subset Simulation works reasonably well for *aBUS*. The number N_{eff} of effectively independent posterior samples in the generated set of K posterior samples depends strongly on the problem at hand.

7.4.4.4 Comparison of aBUS with cBUS

An alternative BUS-based approach, referred to as *cBUS*, that does not require the scaling constant c as input is proposed in [DiazDelaO et al., 2017]: The structural reliability problem



(a) Example problem 1a

(b) Example problem 1b

(c) Example problem 2

(d) Example problem 3

Figure 7.15: Effective number N_{eff} of independent samples obtained with *aBUS* for different α_{opt} and K.

is rephrased as

$$\Omega = \left\{ -\ln(c) \leq \ln\left(\frac{L(\boldsymbol{\theta}|\mathcal{D})}{\pi}\right) \right\}. \tag{7.43}$$

If Eq. (7.43) is tackled by means of Subset Simulation, in this approach, only the stopping criterion of SuS depends on c, but not the intermediate failure domains. For threshold levels smaller than $-\ln(c)$, the algorithm produces samples that follow the posterior distribution. By stopping the algorithm if the logarithm of the failure probability displays a slope of -1, the constant c is not required anymore. To this end, subset levels with associated thresholds smaller than $-\ln(c)$ need to be generated in order to assess whether the stopping criterion is met.

As a consequence, *cBUS* requires more evaluations of the likelihood function than *aBUS*. This can be demonstrated by means of *Example problem 3*, which is also investigated in [DiazDelaO et al., 2017]: *aBUS* requires on average $m = 3$ subset levels, because $p_\Omega = 1.52 \cdot 10^{-3}$ and $\mathrm{E}\left[b_{10^3,\max}\right] = 0.999$. With *cBUS* two additional subset levels are required to verify that the stopping criterion is maintained [DiazDelaO et al., 2017]. After verifying that the stopping criterion is met, it is suggested to use the samples produced in level *three* [DiazDelaO et al., 2017], because (1) all subset levels larger or equal than *three* produce posterior samples and (2) the samples at level *three* are less dependent than the samples produced at higher subset levels [DiazDelaO et al., 2017]. This means that for the same inference problem, *cBUS* requires more evaluations of the likelihood function than *aBUS*.

Additional to that, for *Example problem 2*, the logarithm of the failure probability associated with the individual levels of Subset Simulation in *cBUS* will exhibit a slope that is practically -1 already for threshold levels larger than $-\ln(c)$. The reason is: It is unlikely that likelihood values close to $L_{\max}$ are observed, as $\Pr\left[b_{10^3,\max} > 0.8\right] = 10^{-4}$ (see also Fig. 7.8). Thus, the stopping rule employed in *cBUS* does not guarantee that the threshold of the final subset level is smaller than $-\ln(c)$.

7.5 Nested sampling

7.5.1 Introduction

Nested sampling [Skilling et al., 2006] is a Bayesian inference method whose "prime target" is to compute the *evidence* $c_{\mathbb{E}|\mathcal{M}}$. As by-product, the method produces a set of weighted samples from the posterior distribution.

In *nested sampling*, the integral in Eq. (6.2) is re-written as:

$$c_{\mathbb{E}|\mathcal{M}} = \int_0^1 \mathfrak{L}(X)\,\mathrm{d}X \tag{7.44}$$

where $X \in [0,1]$ is defined as:

$$\begin{aligned} X(\lambda) &= \Pr_{\boldsymbol{\theta}|\mathcal{M}}\left[L(\boldsymbol{\theta}|\mathcal{D},\mathcal{M}) \geq \lambda|\mathcal{D},\mathcal{M}\right] \\ &= \int_{L(\boldsymbol{\theta}|\mathcal{D},\mathcal{M})\geq\lambda} p(\boldsymbol{\theta}|\mathcal{M})\,\mathrm{d}\boldsymbol{\theta} \end{aligned} \tag{7.45}$$

and $\mathfrak{L}(X)$ is the inverse function of $X(\lambda)$; i.e., $\mathfrak{L}(X(\lambda)) \equiv \lambda$; i.e., $\mathfrak{L} \in [0, L_{\max}]$.

The evidence $c_{\mathbb{E}|\mathcal{M}}$ is estimated based on Eq. (7.44) as the weighted sum of values $\mathfrak{L}(X_i)$, where the X_i are an ordered sequence $1 = X_0 > X_1 > X_2 > \ldots > X_N > X_{N+1} = 0$, with $\mathfrak{L}(X_0) = 0$ and $\mathfrak{L}(X_{N+1}) = L_{\max}$. For example, with a trapezoidal integration rule, the evidence is approximated as:

$$c_{\mathbb{E}|\mathcal{M}} \approx \sum_{i=1}^{N+1} (X_i - X_{i-1}) \cdot \frac{1}{2} \cdot (\mathfrak{L}(X_i) + \mathfrak{L}(X_{i-1})) \tag{7.46}$$

The difficulty in numerical *Nested sampling* is to estimate the quantities X_i, $i = 1, \ldots, N$ that belong to the values $\mathfrak{L}_i$ of the likelihood function.

7.5.2 Nested Sampling and Subset Simulation

Nested sampling works on shrinking intermediate subsets of the prior domain that are nested. At the ith intermediate level, samples $\boldsymbol{\theta}$ from the prior distribution need to be generated that are conditional on $L(\boldsymbol{\theta}|\mathcal{D},\mathcal{M}) \geq \mathfrak{L}_i$, where $\mathfrak{L}_i$ is a likelihood threshold. The $\mathfrak{L}_i$ are picked adaptively such that the probability of $\boldsymbol{\theta}$ being in the ith nested subset is approximately p_t, conditional on $\boldsymbol{\theta}$ being in the $i-1$th nested subset. The X_i belonging to the selected $\mathfrak{L}_i$ is estimated as $X_i = p_\mathrm{t} \cdot X_{i-1}$.

Nested sampling has considerable similarities with *Subset Simulation*. Most notably, sampling is performed by means of nested subsets, starting from the prior distribution. The limit-state function employed is $g(\boldsymbol{\theta}) = \mathfrak{L}_i - L(\boldsymbol{\theta}|\mathcal{D},\mathcal{M})$. The main difference between *nested sampling* and *Subset Simulation* is: In *nested sampling* the intermediate probabilities are used as weighting factors to approximate the integral in Eq. (7.45). In *Subset Simulation* the target quantity of interest is the product of the intermediate conditional probabilities.

Another method for Bayesian inference that is based on *Subset Simulation* is *BUS-SuS* (Section 7.3.8). The main difference between *nested Sampling* and *BUS-SuS* is: (i) The evidence in *BUS-SuS* is computed as the scaled product of the intermediate probabilities in SuS, whereas the evidence in *nested sampling* is computed as the sum of functions that depend on the intermediate probabilities. (ii) *BUS-SuS* works in a $(M+1)$-dimensional augmented space of random variable, whereas *nested sampling* directly operates in the space spanned by the uncertain parameter vector $\boldsymbol{\theta}$.

Specialized MCMC algorithms are available for *Subset Simulation* (in particular for Subset Simulation formulated in standard Normal space [Papaioannou et al., 2015]) whose efficiency is independent of the number of random variables in the problem. As mentioned above, the problem solved in *nested sampling* can be interpreted as a reliability problem with limit-state function $g(\boldsymbol{\theta}) = \mathfrak{L}_i - L(\boldsymbol{\theta}|\mathcal{D},\mathcal{M})$. Moreover, it is typically straight-forward to transform the uncertain parameter vector $\boldsymbol{\theta}$ to an underlying independent standard Normal space (see Section 3.2). Therefore, the efficient MCMC algorithms available for Subset Simulation can directly be employed within *nested sampling*.

7.5.3 Standard nested sampling algorithm

There is no single method that can be uniquely identified as *nested sampling*, as different algorithms based on the original variant [Skilling et al., 2006] have been proposed under the term *nested sampling*. The implementation given as *Algorithm (7.5)* is based on [Skilling et al., 2006] and [Skilling, 2012]; it was modified compared to standard implementations of *nested sampling* to accentuate the similarities to *Subset Simulation*.

Algorithm 7.5. *Nested sampling*:
As input, the algorithm requires:

- K, the number of samples to propagate.
- p_t, the target multiplier for the X_i; i.e., $X_i = p_\mathrm{t} \cdot X_{i-1}$. [Skilling et al., 2006] propose to set $p_\mathrm{t} = K^{-1}$. p_t should ideally be selected such that $(1 - p_\mathrm{t}) \cdot K$ is a positive integer number.
- ε, a quantity that controls the termination of the *nested sampling* iteration; e.g., $\varepsilon = 10^{-4}$. The smaller ε, the longer the iteration continues.

The algorithm returns an estimate Z for the evidence $c_{\mathrm{E}|\mathcal{M}}$ and a weighted set of samples that follow the posterior distribution. The weighted samples are denoted as $\boldsymbol{\theta}_i$, the associated weights as w_i, where $i \in \{1, \ldots, n\}$, and n denotes the total number of weighted samples generated by the algorithm.

1. Generate K samples $\boldsymbol{\theta}_j^{(\mathrm{p})}$, with $j = 1, \ldots, K$, from the prior distribution, and set $L_j^{(\mathrm{p})} = L(\boldsymbol{\theta}_j^{(\mathrm{p})}|\mathcal{D}, \mathcal{M})$.
2. Initialize $Z = 0$, $i = 0$, $X_0 = 1$, $n = 0$.
3. while $\left(X_i \cdot \frac{1}{K} \sum_{j=1}^{K} L_j^{(\mathrm{p})} > \varepsilon \cdot Z\right)$ do:
   (a) Sort the K likelihood values $L_j^{(\mathrm{p})}$ and the associated samples $\boldsymbol{\theta}_j^{(\mathrm{p})}$ according to the value of $L_j^{(\mathrm{p})}$ in descending order.
   (b) Increase i by *one*: $i = i + 1$.
   (c) Set $X_i = p_\mathrm{t} \cdot X_{i-1}$.
   (d) Set $k = p_\mathrm{t} \cdot K$.
   (e) Set $\mathfrak{L}_i = \frac{1}{2}\left(L_k^{(\mathrm{p})} + L_{k+1}^{(\mathrm{p})}\right)$
   (f) Set $Z = Z + (X_{i-1} - X_i) \cdot \frac{1}{K-k} \cdot \sum_{j=k+1}^{K} L_j^{(\mathrm{p})}$
   (g) **for** $j = k + 1$ to K do:
      i. Increase n by *one*; i.e., $n = n + 1$.
      ii. Set $\boldsymbol{\theta}_n = \boldsymbol{\theta}_j^{(\mathrm{p})}$ and $w_n = (X_{i-1} - X_i) \cdot \frac{1}{K-k} \cdot L_j^{(\mathrm{p})}$.
      iii. Generate integer number l randomly from the set $\{1, \ldots, k\}$ with equal probability.
      iv. Use sample $\boldsymbol{\theta}_l^{(\mathrm{p})}$ as seed to generate a realization of $\boldsymbol{\theta}_j^{(\mathrm{p})}$ by means of MCMC sampling. The target distribution is proportional to the prior distribution, and conditional on $L\left(\boldsymbol{\theta}_j^{(\mathrm{p})}|\mathcal{D}, \mathcal{M}\right) \geq \mathfrak{L}_i$. As MCMC algorithm, e.g., the CS algorithm (*Algorithm (3.6)*) can be applied inside *Algorithm (3.3)*. To adopt the spread of the MCMC proposal distribution, *Algorithm (3.9)* can be used. Note that the samples $\boldsymbol{\theta}_l^{(\mathrm{p})}$ with $l \in \{1, \ldots, k\}$ already follow the target distribution.
      v. Set $L_j^{(\mathrm{p})} = L\left(\boldsymbol{\theta}_j^{(\mathrm{p})}|\mathcal{D}, \mathcal{M}\right)$
4. Set $Z = Z + X_i \cdot \frac{1}{K} \cdot \sum_{j=1}^{K} L_j^{(\mathrm{p})}$.
5. **for** $j = 1$ to K do:
   (a) Increase n by *one*; i.e., $n = n + 1$.

(b) Set $\boldsymbol{\theta}_n = \boldsymbol{\theta}_j^{(\mathrm{p})}$ and $w_n = X_i \cdot \frac{1}{K-k} \cdot L_j^{(\mathrm{p})}$.

The following steps where modified in *Algorithm (7.5)* compared to standard *nested sampling* as in [Skilling et al., 2006; Skilling, 2012]:

- Termination condition in *step 3*:
  There is no clear stopping criterion in *nested sampling*. The employed stopping criterion terminates the iteration if the expected pending contribution to Z is smaller than a specified fraction ε of the preliminary$Z$ ([Keeton, 2011]).
- Estimation of the intermediate probabilities in *step 3(c)*:
  The given estimate is based on the one typically employed in SuS. In [Skilling et al., 2006], the X_i are estimated assuming that the distribution quantifying our uncertainty about the actual value of p_t is known. The distribution for p_t is modeled by a likelihood and a prior. As likelihood for p_t, the observation "$p_\mathrm{t} \cdot K$ out of K" samples can be interpreted as an observation from a Binomial process (this is similar to what is done in Section 5.2.3.1). As prior for p_t, [Skilling et al., 2006; Skilling, 2012] assume Jeffrey's prior (see Section 5.2.3.2). Based on such assumptions, the distribution for p_t is a beta distribution.
  [Skilling et al., 2006; Skilling, 2012] suggest to either sample realizations from the thus obtained distribution for p_t, or to compute the X_i based on the expectation of the log-transform of the product of the individual p_t that contribute to the X_i.
- MCMC sampling in *step 3(g)(iv)*:
  The same algorithms that are efficient for SuS should also perform well for *nested sampling*. To the best of our knowledge, MCMC algorithms specifically designed for Subset Simulation have not yet been applied in *nested sampling*.

7.5.4 Proposed modifications to the nested sampling algorithm

As in *Subset Simulation*, in standard *nested sampling*, the target conditional probability p_t is kept constant during the simulation. In *BUS-SuS*, the evidence is estimated as the scaled product of the intermediate conditional probabilities. Working with the same target p_t is a reasonable choice if one assumes the same coefficient of variation for each estimated conditional probability. However, contrary to *BUS-SuS*, the evidence in *nested sampling* is estimated based on the weighted sum of likelihood values of the samples that do not fall in the next conditional failure domain. Thus, each subset contributes differently to the integral that describes the evidence.

Working with larger p_t in regions that contribute most to the estimate Z, and with smaller

p_t in regions that contribute little to Z seems a reasonable choice. Such a strategy can be easily integrated in *Algorithm (7.5)*. For example, the p_t can be chosen prior to *step 3(c)* such that the portion added to Z in *step 3(f)* will be approximately $p_\mathrm{e} \cdot \hat{Z}$, where $\hat{Z}$ is an estimate for the final value of Z and p_e is the fraction of $\hat{Z}$ to cover in each subset level; i.e., the targeted total number of subset levels is $1/p_\mathrm{e}$. The quantity $\hat{Z}$ can be computed in the beginning of *step 3* as:

$$\hat{Z} = Z + X_i \cdot \frac{1}{K} \cdot \sum_{j=1}^{K} L_j^{(\mathrm{p})} \tag{7.47}$$

As an additional requirement, the value of p_e can be selected such that the conditional probability p_t is bounded; e.g., $p_\mathrm{t} \in [5\%, 95\%]$.

In this work only the idea of how the standard nested sampling algorithm could be enhanced is presented. Numerical performance investigations have not yet been conducted.

7.6 Transitional Markov chain Monte Carlo (TMCMC)

This section contains material originally published in [Betz et al., 2016b,a]. Some passages and figures are directly taken from the mentioned reference.

7.6.1 Introduction

The Transitional Markov Chain Monte Carlo (TMCMC) method, proposed by [Ching and Chen, 2007], belongs to the class of sequential particle filter methods [Chopin, 2002] and is based on MCMC sampling. The method tries to overcome the issues of MCMC mentioned in Section 7.1 by gradually pushing the samples from the prior to the posterior distribution. The method has become popular in both research and practice: recent contributions include [Zheng and Chen, 2014], [Jensen et al., 2014], [Ortiz et al., 2015], [Hadjidoukas et al., 2015], [Angelikopoulos et al., 2015]. In addition to the posterior samples generated by TMCMC, the method returns an estimate of the evidence of the Bayesian model class, which is needed for Bayesian model class selection and Bayesian model averaging (Section 6.3).

In this section, properties of the TMCMC method are discussed and potential improvements are identified. In particular, it is observed that the TMCMC method tends to produce estimates of the evidence that contain a considerable bias. Three potential modifications to the TMCMC method are proposed: (1) Adjusting the sample weights after each MCMC step tends to improve the performance of the method and reduces the bias in the estimated evidence. (2) A burn-in period in the MCMC sampling step can improve the posterior approximation. (3) The scale of the proposal distribution can be adjusted adaptively such

that the MCMC algorithm maintains a specified near-optimal acceptance rate.

7.6.2 The principle behind TMCMC

The TMCMC method starts with independent samples from the prior distribution. In subsequent steps, the sampling distribution is gradually transformed such that it approaches the posterior distribution. For this purpose, Eq. (6.1) is modified to:

$$p_j(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}) \cdot L(\boldsymbol{\theta}|\mathcal{D})^{q_j} \tag{7.48}$$

where $j = 0, \ldots, m$ denotes the level, and the $q_j \in [0, 1]$ are chosen such that $q_0 = 0 < q_1 < \cdots < q_m = 1$. Consequently, for $j = 0$, $p_0(\boldsymbol{\theta})$ is equal to the prior distribution $p(\boldsymbol{\theta})$; and for $j = m$, $p_m(\mathbf{x})$ matches the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$.

The principle behind the TMCMC method is to gradually push the samples from the prior distribution to the posterior distribution. The speed of this gradual transition is controlled by the coefficients q_j. [Ching and Chen, 2007] proposed to select q_{j+1} based on q_j such that the coefficient of variation of $L(\boldsymbol{\theta}|\mathcal{D})^{q_{j+1}-q_j}$ approximately equals v_t, where $v_\mathrm{t} = 100\%$ was suggested. The value of q_{j+1} can then be determined based on the samples of the previous level as:

$$q_{j+1} = \arg\min_{q} \left(\left| \mathrm{CV}_j(q) - v_\mathrm{t} \right| \right) \tag{7.49}$$

where $\mathrm{CV}_j(q)$ with $q \in (q_j, 1]$ is the sample coefficient of variation of the set $\left\{ L(\boldsymbol{\theta}_{(j,k)}|\mathbf{d})^{q-q_j} \right\}_{k=1}^{K}$, K is the number of samples generated at each level, $\boldsymbol{\theta}_{(j,k)}$ denotes the kth sample at level j, and $L(\boldsymbol{\theta}_{(j,k)}|\mathcal{D})$ is the likelihood value that is associated with $\boldsymbol{\theta}_{(j,k)}$.

The evidence of the stochastic model class $c_{\mathbf{E}|\mathcal{M}}$ Eq. (6.2) can be rewritten as follows:

$$c_{\mathbf{E}|\mathcal{M}} = \int_{\mathbf{x}} p(\boldsymbol{\theta}) \cdot \prod_{j=1}^{m} L(\boldsymbol{\theta}|\mathcal{D})^{q_j - q_{j-1}} \, \mathrm{d}\boldsymbol{\theta} \tag{7.50}$$

$$= \prod_{j=1}^{m} \int_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\mathcal{D})^{q_j - q_{j-1}} \cdot p_{j-1}(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \tag{7.51}$$

$$= \prod_{j=1}^{m} \mathrm{E}_{p_{j-1}(\boldsymbol{\theta})} \left[L(\boldsymbol{\theta}|\mathcal{D})^{q_j - q_{j-1}} \right] \tag{7.52}$$

where $\mathrm{E}_{p_{j-1}(\boldsymbol{\theta})}[\cdot]$ denotes the expectation with respect to distribution p_{j-1}, which can be estimated based on the generated samples:

$$\mathrm{E}_{p_j(\boldsymbol{\theta})} \left[L(\boldsymbol{\theta}|\mathcal{D})^{q_{j+1} - q_j} \right] \approx \frac{1}{K} \sum_{k=1}^{K} L(\boldsymbol{\theta}_{(j,k)}|\mathbf{d})^{q_{j+1} - q_j} \tag{7.53}$$

7.6.3 The TMCMC algorithm

The TMCMC algorithm can be summarized as follows:

Algorithm 7.6. *TMCMC algorithm*:
For the initial $j = 0$, all K samples $\boldsymbol{\theta}_{(0,1)}, \ldots, \boldsymbol{\theta}_{(0,K)}$ are drawn from the prior distribution, and j is set to *one* thereafter. For all $j > 0$, [Ching and Chen, 2007] propose the following scheme:

1. Find q_j through solving Eq. (7.49). If $q_j > 1$, then set $q_j = 1$.
2. For all samples $k = 1, \ldots, K$ compute a weighting coefficient $w_{(j,k)}$:

$$w_{(j,k)} = \left(L(\boldsymbol{\theta}_{(j-1,k)}|\mathcal{D})\right)^{q_j - q_{j-1}} \tag{7.54}$$

3. Compute the mean of the weighting coefficients:

$$S_j = \frac{1}{K} \sum_{k=1}^{K} w_{(j,k)} \tag{7.55}$$

4. Compute the covariance matrix of the Normal proposal distribution:

$$\boldsymbol{\Sigma}_j = \beta^2 \cdot \sum_{k=1}^{K} \left[\frac{w_{(j,k)}}{S_j \cdot K} \cdot \left(\boldsymbol{\theta}_{(j-1,k)} - \overline{\boldsymbol{\theta}}_j\right) \cdot \left(\boldsymbol{\theta}_{(j-1,k)} - \overline{\boldsymbol{\theta}}_j\right)^{\mathrm{T}} \right] \tag{7.56}$$

   with

$$\overline{\boldsymbol{\theta}}_j = \frac{\sum_{l=1}^{K} w_{(j,l)} \cdot \boldsymbol{\theta}_{(j-1,l)}}{\sum_{l=1}^{K} w_{(j,l)}} \tag{7.57}$$

   The coefficient β scales the proposal distribution. [Ching and Chen, 2007] suggest to set $\beta = 0.2$.
5. For each l in $\{1, \ldots, K\}$ set: $\boldsymbol{\theta}^{\mathsf{c}}_{(j,l)} = \boldsymbol{\theta}_{(j-1,l)}$. Thereafter, for $k = 1, \ldots, K$ do:
   - Select index l from the set $\{1, \ldots, K\}$ at random, where each l is assigned probability $\frac{w_{(j,l)}}{\sum_{n=1}^{K} w_{(j,n)}}$.
   - Propose a new sample: draw $\boldsymbol{\theta}^{\mathsf{c}}$ from a Normal distribution that is centered at $\boldsymbol{\theta}^{\mathsf{c}}_{(j,l)}$ and has covariance matrix $\boldsymbol{\Sigma}_j$.
   - Generate a sample r from a uniform distribution on $[0, 1]$.
   - If $r \leq \frac{p_j(\boldsymbol{\theta}^{\mathsf{c}})}{p_j(\boldsymbol{\theta}^{\mathsf{c}}_{(j,l)})}$ then set $\boldsymbol{\theta}^{\mathsf{c}}_{(j,l)} = \boldsymbol{\theta}^{\mathsf{c}}$, otherwise do nothing.
   - Set $\boldsymbol{\theta}_{(j,k)} = \boldsymbol{\theta}^{\mathsf{c}}_{(j,l)}$.
6. If $q_j = 1$ then stop the iteration, otherwise set $j = j + 1$ and continue with 1.

An estimate for the evidence of the assumed model class that is based on Eq. (7.52) is:

$$\hat{c}_{\mathtt{E},} = \prod_{j=1}^{m} S_j \tag{7.58}$$

with S_j defined according to Eq. (7.55).

7.6.4 Observations and potential improvements

7.6.4.1 Observation 1: sample weights

At each level j in the TMCMC method, samples $\{\boldsymbol{\theta}_{(j,1)}, \ldots, \boldsymbol{\theta}_{(j,K)}\}$ are generated that (approximately) follow distribution p_j based on samples $\{\boldsymbol{\theta}_{(j-1,1)}, \ldots, \boldsymbol{\theta}_{(j-1,K)}\}$ that follow distribution p_{j-1}. To each sample a weight is attached according to Eq. (7.54). Instead of performing only a simple resampling step based on the weights, the weighted sampling is combined with a MCMC step: In an iterative process, one randomly picks a sample according to its weight, uses the selected sample as seed to perform a MCMC step with stationary distribution p_j, and replaces the randomly picked sample with the sample that the MCMC step produced. The principle is that the samples that the MCMC step produced already follow (asymptotically) the target distribution p_j, because the seeds of the MCMC step are picked according to their importance weights.

If the proposed sample is accepted, the chain moves on. Therefore, the absolute weight of the current chain should change. This is not taken into account by the original TMCMC method. Asymptotically, the samples of the individual Markov chains that initially follow distribution p_{j-1} approach the target distribution p_j, and the weights of all chains should asymptotically equalize. However, in practice the intermediate distributions of the transition are difficult to obtain. Therefore, the original TMCMC method assumes that for the finite number of samples that are drawn from a single chain, no transition takes place. Consequently, the absolute weight of a Markov chain can be updated after each accepted sample by means of Eq. (7.54). To consider the transition to the target distribution when updating the weights remains an area of future research.

7.6.4.2 Proposed modification (1):

In the original TMCMC method, the weights $w_{(j,k)}$ attached to each Markov chain are computed at the beginning of each TMCMC level (see Eq. (7.54)) and then kept constant. We propose to adapt the weight of a Markov chain each time the chain moves on. In order to do so, step (5) in the TMCMC algorithm given above needs to be modified as follows:

5. For each l in $\{1, \ldots, K\}$ set: $\boldsymbol{\theta}^{\mathsf{c}}_{(j,l)} = \boldsymbol{\theta}_{(j-1,l)}$. Thereafter, for $k = 1, \ldots, K$ do:
   - ...
   - Set $\boldsymbol{\theta}_{(j,k)} = \boldsymbol{\theta}^{\mathsf{c}}_{(j,l)}$.
   - Set $w_{(j,l)} = \left(L(\boldsymbol{\theta}^{\mathsf{c}}_{(j,l)}|\mathbf{d})\right)^{q_j - q_{j-1}}$

Note that S_j and $\mathbf{\Sigma}_j$ are computed only once at the beginning of each level.

As we will demonstrate by means of numerical examples, this modification considerably reduces the average bias in the estimate of the evidence. The statistics of the posterior samples is improved marginally by the modification.

7.6.4.3 Observation 2: burn-in

In the TMCMC method one uses samples from distribution p_{j-1} to generate samples that asymptotically follow distribution p_j. The sampling is based on MCMC, where weighted samples of distribution p_{j-1} are taken as seeds. The weighted samples follow distribution p_j only asymptotically. Therefore, the MCMC sampling performed in the TMCMC method does not possess the property of *perfect sampling* [Cheung and Beck, 2009], where *perfect sampling* implies that the initial distribution of the seeds equals the stationary distribution of the Markov chains [Robert and Casella, 2004].

For practical applications, this is usually not an issue, as is demonstrated by means of numerical examples in Section 7.6.5. Consequently, the TMCMC method does not usually require a burn-in period. However, one should be cautious if TMCMC is used with only a small number of samples per level: In this case a burn-in period might actually be required for the Markov chains to converge.

7.6.4.4 Proposed modification (2):

It is straight-forward to introduce a burn-in period of length N_b to the MCMC sampling of each TMCMC level. Again, only step (5) in the TMCMC algorithm has to be modified:

5. For each l in $\{1,\dots,K\}$ set: $\boldsymbol{\theta}^{\mathsf{c}}_{(j,l)} = \boldsymbol{\theta}_{(j-1,l)}$. Thereafter, for $k = 1,\dots,(K+N_\mathrm{b})$ do:
   - ...
   - If $r \leq \frac{L(\boldsymbol{\theta}^{\mathsf{c}}|\mathbf{d})}{L(\boldsymbol{\theta}^{\mathsf{c}}_{(j,l)}|\mathbf{d})}$ then set $\boldsymbol{\theta}^{\mathsf{c}}_{(j,l)} = \boldsymbol{\theta}^{\mathsf{c}}$, otherwise do nothing.
   - If $k > N_\mathrm{b}$ then set $\boldsymbol{\theta}_{(j,k-N_\mathrm{b})} = \boldsymbol{\theta}^{\mathsf{c}}_{(j,l)}$, otherwise do nothing.
   - Set $w_{(j,l)} = \left(L(\boldsymbol{\theta}^{\mathsf{c}}_{(j,l)}|\mathbf{d})\right)^{q_j - q_{j-1}}$

However, our numerical investigations demonstrate that N_b can usually be set to *zero*.

7.6.4.5 Observation 3: scaling of the proposal

The optimal value of the constant β that is used to scale the proposal distribution depends considerably on the problem at hand. On the one hand, a β that is selected too small leads to

a proposal distribution that accepts many samples; however, subsequent samples in a Markov chain are close to each other. Thus, the correlation in the chain is large and the produced samples will not properly propagate into the relevant domain for reasonable K. On the other hand, a β that is selected too large leads to many rejected samples and, thus, also results in a large chain correlation. The larger the correlation in a Markov chain, the less efficient is the sampling procedure, because the effective number of independent samples in the chain is reduced and, thus, the variance of applied estimators is bound to increase.

Setting $\beta = 0.2$ as proposed in [Ching and Chen, 2007] works well for some problems, but is, in our experience, far from optimal for other problems.

7.6.4.6 Proposed modification (3):

In the original TMCMC algorithm, the proposal distribution is set-up in the space spanned by $\boldsymbol{\theta}$: The proposal is a multivariate Normal distribution that is centered at the current state of the Markov chain and whose variance is defined according to Eq. (7.56).

Without loss of generality, we propose to represent the joint prior PDF of the uncertain parameter vector $\boldsymbol{\theta}$ in terms of an underlying vector $\mathbf{u} \in \mathbb{R}^M$ of independent standard Normal random variables. If the components of $\boldsymbol{\theta}$ are a-priori independent, then the transformation $u_i \to \theta_i$ of the ith component of the uncertain parameter vector is given as: $\theta_i = F_{\theta_i}^{-1}(\Phi(u_i))$, where $F_{\theta_i}^{-1}(\cdot)$ denotes the inverse CDF of the prior distribution of the ith component of $\boldsymbol{\theta}$, and $\Phi(\cdot)$ denotes the CDF of the standard Normal distribution. If the components of $\boldsymbol{\theta}$ are dependent, the marginal transformation based on the Nataf model [Der Kiureghian and Liu, 1986] or the Rosenblatt transformation [Hohenbichler and Rackwitz, 1981] can be used.

The updating problem is then solved in terms of $\mathbf{u}$. Performing the Bayesian inference in an underlying standard Normal space has numerical advantages: (1) The uncertainty in u_i is normalized, whereas the uncertainty in θ_i is usually not. (2) The support of $\boldsymbol{\theta}$ may be bounded, whereas the support of $\mathbf{u}$ is not bounded.

Based on this, the proposal is then set-up in the space spanned by $\mathbf{u}$: it is a multivariate Normal distribution that is centered at the current state of the Markov chain and whose covariance matrix is defined as the sample covariance in terms of $\mathbf{u}$ (and not in terms of $\boldsymbol{\theta}$). We propose to select the initial scaling factor β of the thus obtained proposal distribution as: $\beta = 2.4/\sqrt{M}$ based on [Gelman et al., 2004b; Andrieu and Thoms, 2008].

Additionally, the performance of the MCMC algorithm can be enhanced, by adjusting the scaling factor β adaptively during the simulation. Often the scaling factor β is tuned such that the average acceptance probability of the MCMC algorithm approaches a specified target acceptance-rate t_{acr} [Andrieu and Thoms, 2008; Papaioannou et al., 2015]. We suggest to adaptively modify β such that the monitored average acceptance-rate approaches the follow-

ing target acceptance-rate $t_{\text{acr}} = 0.21/M + 0.23$, i.e., $t_{\text{acr}} = 0.44$ for $M = 1$, $t_{\text{acr}} = 0.27$ for $M = 5$ and $t_{\text{acr}} = 0.23$ for large M. This rule is based on the findings published in [Roberts et al., 1997, 2001].

The algorithm to adaptively update β is as follows: At the initial sampling level set $\beta_{(\text{old})} = 2.4/\sqrt{M}$; in all other levels use the last value of $\beta_{(\text{old})}$ from the previous sampling level. At the beginning of each sampling level, set $N_{\text{adapt}} = 1$. Perform N_{a} MCMC steps. Thereafter, evaluate the coefficient $c_{\text{a}} = (p_{\text{acr}} - t_{\text{acr}}) / \sqrt{N_{\text{adapt}}}$, where p_{acr} is the mean acceptance-rate of the last N_{a} MCMC steps (i.e., the number of accepted samples divided by N_{a}), and t_{acr} denotes the target acceptance-rate. Modify β based on the value of c_{a}: set $\beta_{(\text{new})} = \beta_{(\text{old})} \cdot \exp(c_{\text{a}})$. Increase the value of N_{adapt} by one, set $\beta_{(\text{old})} = \beta_{(\text{new})}$, perform another N_{a} MCMC steps and evaluate c_{a} again. Repeat this until the required number of samples is generated. We suggest to set $N_{\text{a}} = 100$.

Note: In principle, the algorithm to choose the scaling factor β adaptively works also if the problem is solved directly in the space spanned by $\boldsymbol{\theta}$. However, for the following reasons, we expect a reduced efficiency in this case: On the one hand, working in the underlying $\mathbf{u}$-space facilitates the choice of an initial scaling factor that leads to a robust behavior of the algorithm for a large variety of problems. Moreover, if the support of $\boldsymbol{\theta}$ is bounded and a Normal proposal distribution is used, some proposed samples have to be rejected simply because the proposed sample is not within the support of $\boldsymbol{\theta}$. On the other hand, solving the updating problem in terms of $\mathbf{u}$ adds an additional layer of complexity. Working in the underlying standard normal space, the transformed target distribution is more likely to obey the conditions for which the suggested target acceptance rate $t_{\text{acr}} = 0.21/M + 0.23$ is the optimal one [Roberts et al., 1997]. This is especially the case in the initial sampling levels of iTMCMC where the exponent of the likelihood is small and hence the transformed target density is closer to the independent standard normal prior.

7.6.4.7 iTMCMC algorithm

The TMCMC algorithm that takes the first and the third proposed modification into account is referred to as *iTMCMC* and presented in the following.

Algorithm 7.7. *iTMCMC algorithm*:
Initially, set $j = 0$ and $\beta = 2.4/\sqrt{M}$, where M denotes the dimension of vector $\boldsymbol{\theta}$. Furthermore, set $t_{\text{acr}} = 0.21/M + 0.23$ and $N_{\text{a}} = 100$, where N_{a} denotes the number of MCMC steps after which the value of β is modified Additionally, set $N_{\text{adapt}} = 1$.

For the initial $j = 0$, all K samples $\mathbf{u}_{(0,1)}, \ldots, \mathbf{u}_{(0,K)}$ are drawn from the M-dimensional independent standard normal distribution. For $k = 1, \ldots, K$, the standard normal samples $\mathbf{u}_{(0,k)}$ are transformed to samples $\boldsymbol{\theta}_{(0,k)}$ in the original parameter space, using e.g. the *Nataf* transformation or the *Rosenblatt*

transformation. Note that the samples $\mathbf{u}_{(0,k)}$ follow the prior distribution. Thereafter, j is set to *one*. For all $j > 0$, the following scheme is applied:

1. Find q_j through solving the minimization problem

$$q_{j+1} = \arg\min_{q} \left(\left| \mathrm{CV}_j(q) - v_\mathrm{t} \right| \right) \tag{7.59}$$

   where $\mathrm{CV}_j(q)$ with $q \in (q_j, 1]$ is the sample coefficient of variation of the current samples, and v_t is the target coefficient of variation (typically set to $v_\mathrm{t} = 100\%$). If $q_j > 1$, then set $q_j = 1$.

2. For all samples $k = 1, \ldots, K$ compute a weighting coefficient $w_{(j,k)}$:

$$w_{(j,k)} = \left(L(\boldsymbol{\theta}_{(j-1,k)} | \mathcal{D}) \right)^{q_j - q_{j-1}} \tag{7.60}$$

3. Compute the mean of the weighting coefficients:

$$S_j = \frac{1}{K} \sum_{k=1}^{K} w_{(j,k)} \tag{7.61}$$

4. Compute the sample covariance matrix of the Normal proposal distribution:

$$\boldsymbol{\Sigma}_{\mathrm{sample},j} = \sum_{k=1}^{K} \left[\frac{w_{(j,k)}}{S_j \cdot K} \cdot \left(\mathbf{u}_{(j-1,k)} - \overline{\mathbf{u}}_j \right) \cdot \left(\mathbf{u}_{(j-1,k)} - \overline{\mathbf{u}}_j \right)^\mathrm{T} \right] \tag{7.62}$$

   with

$$\overline{\mathbf{u}}_j = \frac{\sum_{l=1}^{K} w_{(j,l)} \cdot \mathbf{u}_{(j-1,l)}}{\sum_{l=1}^{K} w_{(j,l)}} \tag{7.63}$$

5. For $k = 1, \ldots, K$, set: $\mathbf{u}^\mathsf{c}_{(j,k)} = \mathbf{u}_{(j-1,k)}$ and $\boldsymbol{\theta}^\mathsf{c}_{(j,k)} = \boldsymbol{\theta}_{(j-1,k)}$.

6. Set $n_\mathrm{a} = 0$ and $\mathrm{acr} = 0$.

7. For $k = 1, \ldots, K$ do:

   (a) Select index l from the set $\{1, \ldots, K\}$ at random, where each index $i \in \{1, \ldots, K\}$ of the set is assigned probability

$$\frac{w_{(j,i)}}{\sum_{n=1}^{K} w_{(j,n)}} \tag{7.64}$$

   (b) Propose a new sample: draw $\mathbf{u}^*$ from a normal distribution that is centered at $\mathbf{u}^\mathsf{c}_{(j,l)}$ and has covariance matrix $\boldsymbol{\Sigma}_j = \beta^2 \cdot \boldsymbol{\Sigma}_{\mathrm{sample},j}$.

   (c) Transform the sample $\mathbf{u}^*$ to sample $\boldsymbol{\theta}^*$ in original parameter space.

   (d) Generate a sample r from a uniform distribution on $[0, 1]$.

   (e) If $r \leq \frac{p_j(\boldsymbol{\theta}^*)}{p_j(\boldsymbol{\theta}^\mathsf{c}_{(j,l)})}$ then set $\mathbf{u}^\mathsf{c}_{(j,l)} = \mathbf{u}^*$, $\boldsymbol{\theta}^\mathsf{c}_{(j,l)} = \boldsymbol{\theta}^*$ and $\mathrm{acr} = \mathrm{acr} + 1$, otherwise do nothing.

   (f) Set $\mathbf{u}_{(j,k)} = \mathbf{u}^\mathsf{c}_{(j,l)}$ and $\boldsymbol{\theta}_{(j,k)} = \boldsymbol{\theta}^\mathsf{c}_{(j,l)}$.

   (g) Set $w_{(j,l)} = \left(L(\boldsymbol{\theta}^\mathsf{c}_{(j,l)} | \mathcal{D}) \right)^{q_j - q_{j-1}}$

   (h) Increase n_a by *one*; i.e., $n_\mathrm{a} = n_\mathrm{a} + 1$.

(i) If $n_\mathrm{a} \geq N_\mathrm{a}$, then

- Compute the average acceptance-rate p_acr of the last N_a MCMC steps; i.e., $p_\mathrm{acr} = \mathrm{acr}/N_\mathrm{a}$
- Evaluate coefficient $c_\mathrm{a} = (p_\mathrm{acr} - t_\mathrm{acr}) \,/\, \sqrt{N_\mathrm{adapt}}$.
- Update the value of $\beta = \beta \cdot \exp(c_\mathrm{a})$.
- Set $n_\mathrm{a} = 0$, $N_\mathrm{adapt} = N_\mathrm{adapt} + 1$ and $\mathrm{acr} = 0$.

8. If $q_j = 1$ then stop the iteration, otherwise set $j = j + 1$ and continue with 1.

Remark: iTMCMC proposes an adaptive choice of the scaling factor β in contrast to the use of a fixed β suggested in [Ching and Chen, 2007]. If the sought posterior distribution has certain known properties, it is possible to theoretically derive a fixed optimal scaling factor. In particular, if the target distribution has iid components and obeys certain regularity conditions, the optimal scaling factor is $2.38/\sqrt{M \cdot F}$, where M is the dimension and F is a Fisher's information measure of the component target density with $F = 1$ for the Normal case [Gelman et al., 1996; Roberts et al., 1997]. The corresponding optimal acceptance rate of the candidate sample is 0.23 for sufficiently large M. For such situations, β should be directly selected as the optimal value. However, the crux is that in most real applications, these conditions are not met and the optimal β cannot be found directly. Ching & Wang [Ching and Wang, 2016] deal with this situation by postulating a fixed value of β that works well in a variety of cases, whereas iTMCMC determines a near-optimal β case-specifically in an adaptive manner.

7.6.5 Numerical Investigations

Numerical investigations of the proposed modifications to the TMCMC algorithm are performed in [Betz et al., 2016b] by means of three example problems. Additional performance studies are conducted in [Ching and Wang, 2016; Betz et al., 2016a].

Chapter 8

Conclusions and Outlook

8.1 Concluding remarks

Uncertainty is best viewed as being *personal*, and not as either *subjective* or *objective*. *Cox-Jaynes interpretation of probability* provides a basis for *Probability Theory* that is more appropriate for uncertainty quantification in engineering models than an interpretation based on the *Kolmogorov axioms*, as all probabilities are viewed as conditional probabilities. Within a Bayesian framework, all probabilities are conditional on the stochastic model class that contains the assumptions made.

Within a Bayesian analysis, the evidence is a quantitative measure for the plausibility of a stochastic model class. The evidence can be used to compare competing stochastic model classes. The use of weakly- or non-informative prior distributions in stochastic engineering models is best avoided, as the evidence is rendered meaningless and the posterior distribution might be improper. Uncertain parameters of engineering models are frequently represented through uniform prior distributions. Typically, this modeling choice is made out of convenience, and such an assumption should be scrutinized critically. Prior distributions based on the principle of maximum information entropy are a simple and objective choice. However, other distribution choices can be more appropriate depending on the problem at hand. Neglecting the dependence in error structures and assuming statistical independence based on the principle of maximum information entropy is often an inappropriate assumption for engineering models. The efficiency of numerical inference methods depends on how the likelihood is expressed in terms of prediction- and observation-errors.

Bayesian updating using structural reliability methods (BUS) [Straub and Papaioannou, 2015] converts sampling from the posterior into sampling from the failure domain of a structural reliability problem. The use of Subset Simulation within BUS, referred to as BUS-SuS, is a particularly interesting strategy for Bayesian inference. BUS-SuS can handle inference

problems with many uncertain parameters, and efficiently generates posterior samples even if the posterior differs considerably from the prior. In order to apply the BUS approach, the maximum that the likelihood function can take must be known. The proposed adaptive variant of BUS-SuS, referred to as *aBUS*, does not require the maximum of the likelihood to be known, in order to generate samples that follow the posterior distribution and provide an unbiased estimate of the evidence.

Both aBUS and BUS-SuS employ Subset Simulation (SuS) as reliability method. Within SuS, conditional samples are generated by means of Markov chain Monte Carlo (MCMC) simulation. In this context, the conditional sampling in standard Normal space algorithm proposed in [Papaioannou et al., 2015] is an efficient and easy to implement MCMC algorithms that works well in problems with many uncertain parameters. The uncertainty about the estimated probability of failure in SuS depends on the formulation of the limit-state function; i.e., on the shape of the final and the intermediate failure domains. For unfavorable limit-state functions and small probabilities of failure, the distribution describing the uncertainty about the estimated probability of failure can be considerably right-skewed.

8.2 Main contributions of this thesis

New developments and finding that emerged from this thesis are:

1. A modified variant of the BUS approach combined with Subset Simulation, referred to as *aBUS*, is proposed that does not require the scaling constant c^{-1} as input (see Section 7.4). It is argued that aBUS is computationally at least as efficient as standard BUS-SuS.
2. Contributions to the BUS approach:
   (a) A numerically more beneficial variant of the BUS limit-state function is suggested in Section 7.3.8.1. The suggested limit-state function is based on a log-transform that was first used in [DiazDelaO et al., 2017] to propose a variant of the BUS approach.
   (b) A post-processing step is proposed to correct the posterior distribution and the estimated evidence if BUS was performed with the constant c^{-1} selected smaller than the maximum of the likelihood function, denoted $L_{\max}$ (see Section 7.3.9.1).
3. Observations regarding Subset Simulation (SuS) made: The probability of failure in the estimated probability of failure in Subset simulation is denoted by $p_{f,\mathrm{SuS}}$. The actual underlying (possibly unknown) probability of failure is denoted by P_f.
   (a) It is shown that the distribution of the estimated $p_{f,\mathrm{SuS}}$ can be considerably right-skewed (*positive* skewness) for small P_f (Example 5.4). The distribution of the

log-transformed $p_{f,\mathrm{SuS}}$ exhibits a strong skewness as well. A consequence of this strongly asymmetric distribution of $p_{f,\mathrm{SuS}}$ for small P_f is that the coefficient of variation is not a good measure to quantify the uncertainty about the probability of failure estimated obtained with Subset Simulation.

(b) Moreover, the estimate for the upper bound of the coefficient of variation of $p_{f,\mathrm{SuS}}$ proposed in [Au and Beck, 2001] neglects the influence of correlated seeds. This can result in the actual coefficient of variation being above the estimated upper bound for small P_f. This effect is demonstrated by means of numerical examples (Example 5.1) in Section 5.3.4.2.

(c) The beta distribution is not a good choice in describing the uncertainty about $p_{f,\mathrm{SuS}}$. The log-Normal distribution is usually a better choice. However, also the log-Normal distribution is not optimal in quantifying the uncertainty about $p_{f,\mathrm{SuS}}$ (see Example 5.3 in Section 5.3.4.2).

(d) The bias in the estimated $p_{f,\mathrm{SuS}}$ is negligible compared to the spread of $p_{f,\mathrm{SuS}}$ in all investigated example problems (Section 5.3.4.2).

(e) The median of $p_{f,\mathrm{SuS}}$ tends to be smaller than P_f for small P_f. This is especially true if the shape of the final or the intermediate failure domains is not optimal for the selected MCMC sampling strategy in Subset Simulation. As a consequence, for small P_f, Subset Simulation returns an estimate $p_{f,\mathrm{SuS}}$ that is in the majority of cases smaller than P_f.

(f) The statistics of $p_{f,\mathrm{SuS}}$ estimated by means of a large number of samples per level is slightly better than performing Subset Simulation repeatedly with a smaller number of samples per level at a comparable number of total limit-state function calls. However, by performing Subset Simulation repeatedly (with a smaller number of samples per level), the uncertainty about the estimated probability of failure can be quantified (at least approximately)[1]. Contrary to that, a satisfying and conservative measure to quantify the uncertainty about $p_{f,\mathrm{SuS}}$ from a single run of Subset Simulation does not exist at present. The main reason is that the uncertainty about $p_{f,\mathrm{SuS}}$ depends to a large degree on the formulation of the limit-state function; i.e., the shape of the final and the intermediate failure domains (see e.g. Example 5.5).

(g) By performing multiple independent runs of Subset Simulation, the average of the estimated probability of failure can be evaluated. The distribution of the average converges asymptotically to a Normal distribution for an increasing number of repeated SuS runs, according to the central limit theorem. However, it is demonstrated that even for a large number of repeated SuS runs (10^4), the uncertainty about the estimated average cannot be approximated well by a Normal distribution if P_f is small (Example 5.6).

[1] It is recommended to use at least 10^3 samples per subset level to keep the bias in $p_{f,\mathrm{SuS}}$ small.

4. Mathematical reasoning is given as to why the MCMC algorithm *conditional sampling in standard Normal space* (CS) proposed in [Papaioannou et al., 2015] is very efficient for problems with many uncertain parameters (Section 3.5.6.2). In the standard version, the CS algorithm has a single parameter that can be expressed as the correlation ρ between the current state of the Markov chain and the proposed sample. Let M denote the number of uncertain parameters in the problem. It is shown that for $M \to \infty$, the parameter ρ is essentially equal to the cosine of the angle between the proposed sample and the current state, and the distance between the proposed sample and the current state is $\sqrt{2M(1-\rho)}$.

5. A MCMC algorithm is proposed in Section 3.5.7 that is specifically designed to generate samples from standard Normal target distributions. The proposed algorithm is called directional conditional sampling (DCS). However, it is found that CS is more efficient than DCS.

6. A modified variant of the TMCMC method is proposed that reduces the bias in the estimate of the evidence (Section 7.6.4).

7. Similarities between nested sampling and Subset Simulation are highlighted. Based on the experiences with Subset Simulation, a modified version of nested sampling that works with different "intermediate threshold probabilities" is suggested.

8. Choosing prior distributions based on the principle of maximum information entropy (MEP) represents an objective probabilistic modeling approach, if some probabilistic constraints can be specified and the associated distribution is unknown. However, it is found that MEP priors do not guarantee conservative posterior results (e.g., a large posterior spread in the uncertain model parameters, or a large posterior probability of failure). Other probabilistic models that also maintain the specified constraints can have a larger associated evidence, and at the same time be more conservative with respect to the posterior results (Example 6.7).

9. Different ways to formulate the likelihood function in terms of prediction- and observation-error are shown and rated with respect to their computational complexity (Section 6.5).

10. It is shown that the expansion optimal linear estimation (EOLE) method proposed in [Li and Der Kiureghian, 1993] constitutes (under rather general conditions) a special case of the Nyström method and, thus, the EOLE method numerically approximates the Karhunen–Loève expansion of random fields (see [Betz et al., 2014c] and Section D.7).

8.3 Outlook

8.3.1 Models of different resolution in combination with aBUS

In this work, numerical Bayesian inference schemes based on an underlying fixed deterministic "black box" model are investigated. Alternatively, the system of interest could be represented by models of different mesh resolution (see e.g., [Koutsourelakis, 2009a,b]). By combining e.g. *aBUS* with deterministic models that have different mesh resolutions – by starting with a coarse mesh resolution and by successively refining the mesh, an additional computational speed-up can be achieved. Especially for computational demanding models, this appears to be a promising strategy. For aBUS and BUS-SuS, the main difficulty is to account for non-nested subsets when modifying the mesh resolution (see [Ullmann and Papaioannou, 2015]). The fact that the intermediate failure levels in *aBUS* are flexible might prove helpful in this context.

Other potential improvements regarding *aBUS* and *BUS-SuS* are mainly coupled to improvements in Subset Simulation. This is discussed in the following.

8.3.2 Uncertainty in the estimated probability of failure in Subset Simulation

For some (favorable) problems and large P_f, the distribution of the estimated probability of failure $p_{f,\mathrm{SuS}}$ in Subset Simulation can be almost symmetric. For other (unfavorable) problems and small P_f, the distribution of $p_{f,\mathrm{SuS}}$ can be considerably skewed, such that the median is much smaller than the mean and P_f. Therefore, quantifying the uncertainty about $p_{f,\mathrm{SuS}}$ in terms of the coefficient of variation or in terms of a Normal approximation is generally not a good choice. The uncertainty about the average $p_{f,\mathrm{SuS}}$ is best expressed in terms of credible intervals[1]. However, this is difficult, as the distribution of $p_{f,\mathrm{SuS}}$ can not be described well through a standard distribution model. The beta distribution is an inappropriate choice (Example 5.3). The log-Normal distribution describes the uncertainties about $p_{f,\mathrm{SuS}}$ better, but is still not an optimal choice.

Some future research should be dedicated to quantifying the uncertainty about $p_{f,\mathrm{SuS}}$, as no satisfying and conservative measure has yet been proposed. Ideally, the uncertainty about $p_{f,\mathrm{SuS}}$ is expressed through a Bayesian post-processing step. This quest can be approached by two fundamentally different strategies:

[1] When expressing $p_{f,\mathrm{SuS}}$ in terms of credible intervals for very small P_f and a relatively small number of samples per level, one hast to be careful: The mean of $p_{f,\mathrm{SuS}}$ can be well above the 95% confidence interval of $p_{f,\mathrm{SuS}}$. Nevertheless, the use of credible intervals is still better than the use of the coefficient of variation, which can – in this context – be misinterpreted very easily.

Strategy 1 The uncertainty about $p_{f,\mathrm{SuS}}$ is quantified based on a single run of Subset Simulation.

This approach is rather difficult, as the potential dependency of MCMC seeds has to be quantified and its influence on $p_{f,\mathrm{SuS}}$ must be assessed. At present, no technique to account for the influence of correlated seeds has been proposed. The main difficulty is that the uncertainty about $p_{f,\mathrm{SuS}}$ can depend strongly on the formulation of the limit-state function; namely the shape of the final and the intermediate failure domains. Moreover, how the limit-state formulation performs in SuS, is conditional on the employed MCMC algorithm.

Existing approaches tend to underestimate the uncertainty about $p_{f,\mathrm{SuS}}$: In [Au and Beck, 2001], an estimate for the upper bound of the coefficient of variation of $p_{f,\mathrm{SuS}}$ is given, assuming that the influence of correlated MCMC seeds is negligible. In [Zuev et al., 2012], the uncertainty about $p_{f,\mathrm{SuS}}$ is expressed based on a beta distribution, also assuming that the influence of correlated MCMC seeds is negligible.

The author of this thesis tried to express the uncertainty about $p_{f,\mathrm{SuS}}$ through the product of the conditional failure probabilities p_i, where the p_i were assumed to follow a beta distribution. The correlation of the Markov chains was taken into account by estimating an efficient number of samples according to [Au and Beck, 2001]. The correlation of the MCMC seeds was approximated by employing the Nataf distribution and setting up a correlation matrix for the individual beta-distributed p_i. The corresponding correlation coefficients where estimated based on monitoring the history of the samples used as seeds during the preceding SuS levels. The uncertainty about $p_{f,\mathrm{SuS}}$ was quantified by generating a large number of samples from the Nataf distribution model and by computing $p_{f,\mathrm{SuS}}$ as the product of the individual sampled p_i. However, so far an appropriate and conservative measure could not be obtained. Moreover, the algorithms to monitor seed dependencies tend to become rather complex.

An alternative strategy could be to set up the previously mentioned correlation matrix through approximate relations based on the effective number of samples in a single SuS level. The relation could somehow be learned by means of a Bayesian approach and a large number of exemplary reliability problems.

Strategy 2 The uncertainty about $p_{f,\mathrm{SuS}}$ is quantified based on the average of multiple runs of Subset Simulation (with a possibly small number of samples per level).

Compared to the previous strategy, it is much easier to repeatedly perform Subset Simulation multiple times, and to quantify the uncertainty about the average $p_{f,\mathrm{SuS}}$ based on the obtained outcomes of $p_{f,\mathrm{SuS}}$. The estimates from the individual SuS runs are statistically independent. However, it is usually not valid to assume that the average $p_{f,\mathrm{SuS}}$ follows a Normal distribution due to the central limit theorem, because of the potentially highly skewed shape of the distribution of $p_{f,\mathrm{SuS}}$.

Based on the estimated outcomes of $p_{f,\mathrm{SuS}}$ in n performed SuS runs, the CDF of $p_{f,\mathrm{SuS}}$ can be approximated. The statistical uncertainty due to a finite number of repeated runs should ideally be accounted for when setting up the CDF. Using this approximated CDF, the CDF of the average probability of failure estimated from n repeated SuS runs can be computed. Form this, credible intervals for the value of P_f can be obtained.

Instead of approximating the CDF for $p_{f,\mathrm{SuS}}$ entirely based on sample statistics, an appropriate distribution model could be sought and fitted. Instead of fitting with respect to sample mean and sample standard deviation, a fit based on a maximum likelihood or on a Bayesian approach could be investigated. The log-Normal distribution is a good starting point. However, possibly different distribution models might have to be combined to achieve a good fit also in the tails. The challenge is to come up with a distribution that provides a good fit for a large variety of example problems, and can be fitted well already with only a small number of repeated SuS runs.

8.3.3 Efficiency of MCMC in Subest Simulation

If independent samples conditional on the intermediate failure domains could be generated, the performance of Subset Simulation would, as the performance of Monte Carlo simulation, depend only on the target probability of failure[1]. Moreover, the statistical uncertainty about the estimated probability of failure could be expressed through the product of beta distributed random variables. However, we cannot generated independent samples conditional on the intermediate failure domains, and commonly use MCMC simulation instead. MCMC introduces the dependency of the performance of SuS on the shape of the intermediate failure domain.

As a consequence, if the performance of SuS is to be improved also for unfavorably formulated limit-state functions, one needs to enhance the performance of the employed MCMC sampling strategy. The difficulty is that the chosen MCMC strategy should perform also well for problems with many uncertain parameters. With the MCMC algorithm proposed in [Papaioannou et al., 2015], an efficient and elegant sampling technique is available that does not depend on the number of uncertain parameters in the problem. However, MCMC is such a critical part in SuS that future research would be desirable. Especially the adaptive learning of the spread of the MCMC proposal distribution offers potential for further investigations, either in terms of the acceptance rate, or based on alternative performance measures like the expected squared jumping distance (ESJD).

[1] In the entire paragraph it is neglected that in practice, the intermediate failure domains are estimated conditional on fixed p_i, and not the probabilities p_i for fixed limit-state thresholds.

8.3.4 Computational challenges

Increasing speedup of computer cores by increasing the clock rate of computer cores causes considerable technical problems. Therefore, processors with multiple cores are used in modern hardware instead of a processor with a single but very powerful core. As a consequence, software that can run in parallel on multiple cores is on the advance.

One has typically limited potential to influence the software of the models that one works with – often they are so-called black box models. However, the numerical algorithms that one works with could be adopted to work in parallel. Moreover, simulation methods are usually *embarrassingly parallel*[1]. For example Monte Carlo simulation: One can start the simulation on different machines/cores. In the end, all one needs to do is to count the total number of samples generated and the number of samples that were in the failure domain. The same is in principle true when looking at the individual levels of Subset Simulation. If, however, the spread of the proposal distribution is learned adaptively, then communication between different jobs becomes a bit more important – but the problem is still considered embarrassingly parallel. Some research on how to learn the spread adaptively when SuS is run in parallel could improve the practical applicability of the method. A very promising alternative is to have independent runs of SuS on the different cores. As mentioned previously, this allows to quantify the uncertainty in the estimated probability of failure.

All such research attempts should investigate challenging limit-state functions; e.g., limit-state function g_5 with $m = 4$. For the commonly employed limit-state functions, most MCMC methods for SuS behave similar. Also the optimal acceptance rate and the optimal intermediate probabilities could be studied especially for very challenging limit-state functions.

8.3.5 Conservative assumptions in engineering models

It is the task of engineers to design structures such that they are reliable. The employed engineering models used to approximate the response of the system of interest are inevitably simplified compared to reality. Simplifying assumptions are commonly made such that the model predicts the response on the conservative/unfavorable side; i.e., the model tends to predict displacements, stresses, discharges too large. Consequently, engineering models are often deliberately selected such that they are biased to the conservative side.

In a forward model, this is not an issue, as conservative modeling assumptions result usually in conservative estimates. However, if uncertainties in modeling parameters are to be reduced based on observed data, such conservative modeling assumptions are better avoided.

[1] *Embarrassingly parallel* means that the problem can easily be split in smaller tasks that can be computed almost independently of each other. The obtainable speedup is essentially the number of cores available.

Observed data comprises larger surprises in conservative models than in realistic models. The larger the surprises in the data, the stronger the learning effect, and the more uncertainties can be reduced. Thus, the posterior uncertainty in a conservative model can possibly be smaller than in a more realistic model. Moreover, such conservative assumptions can result in an artificial shift of the mean of posterior quantities to the "unconservative" side. For example, assume a model that tends to predict displacements on average too large. If an observed displacement is used to learn the Young's modulus of the structure, the value of the Young's modulus needs to be artificially increased in order to compensate for the model predicting displacements too large.

The influence of such assumptions has not been thoroughly investigated so far. Uniform prior distributions in engineering models are most often motivated by conservative assumptions. Also modeling and observation errors are often considered as independent, as this maximizes the entropy and is viewed as a conservative modeling approach. Studies performed in this thesis indicated that his can be a very critical assumption. A proper analysis of this relations should be the scope of future research.

8.3.6 Requirements for future engineering standards

At present, most of engineering design is based on partial safety factors; and so are the relevant technical standards. This is, for example, the case in civil engineering. [Eurocode 0, 2015] allows for a probabilistic design approach, but specific recommendations for distributions and parameters to select are missing in existing standards. This means that the probabilistic modeling choices have all to be made on an individual basis.

Materials that are used and loads that are applied on the structures are all described in technical standards. The decision on how to model the associated uncertainties should not be put in the hand of the individual engineer. The problem is less that the engineer might not be expert enough to quantify this uncertainties probabilistically. The problem is more that the so performed reliability analyses lack comparability.

Future engineering standards should specify distributions and the parameters of the distributions that represent the uncertainties about different material properties and the applied loads. They do not have to "exactly" represent the uncertainties at hand (such a thing does not exist), because the analysis is conditional on the assumption. A first attempt on a probabilistic model code is made in [JCSS, 2001–2015]. However, such an approach should be intrinsic to every technical standard.

Appendix A

Numerical descriptors of random variables

In the following, some numerical descriptors for random variables X with density $p_X(x)$ are listed. This includes *location parameters* like the *mean*, *median* and *mode*, as well as *dispersion parameters* like the *variance*, *standard deviation* and *interquartile range*.

Note: The definitions are given with respect to continuous probability distributions. However, the transformation in case of discrete probability distributions should be straight-forward.

A.1 Expectation

The *expectation* of random variable X is defined as:

$$\mu_X = \mathrm{E}_X[X] = \int_{-\infty}^{\infty} x \cdot p_X(x)\,\mathrm{d}x \tag{A.1}$$

The *expectation* of X is also called the *mean* or *first moment* of X. It describes the central tendency of a distribution – loosely speaking, the long-run average of X. More generally, the expectation of an arbitrary function of X, $h(X)$, is defined as:

$$\mathrm{E}_X[h(X)] = \int_{-\infty}^{\infty} h(x) \cdot p_X(x)\,\mathrm{d}x \tag{A.2}$$

The X in $\mathrm{E}_X[\cdot]$ denotes that the expectation is taken with respect to X. If it is clear that the expectation is to be taken with respect to e.g. X, the notation can be simplified to $\mathrm{E}[\cdot]$.

In the following, some properties of the expectation operator $\mathrm{E}[\cdot]$ are listed:

Expectation of constants The expectation of a constant $c \in \mathbb{R}$ is c; i.e., $\mathrm{E}[c] = c$.

Proof A.1.

$$\mathrm{E}[c] = c \cdot \int_{-\infty}^{\infty} p_X(x)\,\mathrm{d}x = c$$

□

Linear operator The expectation $\mathrm{E}(\cdot)$ is a linear operator, because:

$$\mathrm{E}[aX + bY] = a\,\mathrm{E}[X] + b\,\mathrm{E}[Y] \tag{A.3}$$

where $a, b \in \mathbb{R}$ are constants and X, Y are random variables.

Proof A.2. Let $p_{X,Y}(x, y)$ be the joint probability density function of X and Y, and $p_X(x)$, $p_Y(y)$ be the respective marginal densities. In this case we can write:

$$\begin{aligned}
\mathrm{E}[aX + bY] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (aX + bY) \cdot p_{X,Y}(x, y)\,\mathrm{d}x\,\mathrm{d}y \\
&= a\int_{-\infty}^{\infty} X \int_{-\infty}^{\infty} p_{X,Y}(x, y)\,\mathrm{d}y\,\mathrm{d}x + b\int_{-\infty}^{\infty} Y \int_{-\infty}^{\infty} p_{X,Y}(x, y)\,\mathrm{d}x\,\mathrm{d}y \\
&= a\int_{-\infty}^{\infty} X \cdot p_X(x)\,\mathrm{d}x + b\int_{-\infty}^{\infty} Y \cdot p_Y(y)\,\mathrm{d}y \\
&= a\,\mathrm{E}[X] + b\,\mathrm{E}[Y]
\end{aligned}$$

□

From Eq. (A.3) it follows that:

$$\mathrm{E}[aX + b] = a\,\mathrm{E}[X] + b \tag{A.4}$$

Iterated expectation

$$\mathrm{E}\left[\mathrm{E}\left[X|Y\right]\right] = \mathrm{E}[X] \tag{A.5}$$

Proof A.3. With $p_{X,Y}(x, y) = p_{X|Y}(x|y) \cdot p_Y(y)$ we can write:

$$\begin{aligned}
\mathrm{E}\left[\mathrm{E}\left[X|Y\right]\right] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} X \cdot p_{X|Y}(x|y)\,\mathrm{d}x\,p_Y(y)\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} X \cdot p_{X|Y}(x|y) \cdot p_Y(y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty} X \cdot p_X(x)\,\mathrm{d}x \\
&= \mathrm{E}[X]
\end{aligned}$$

□

Non-multiplicativity In general, $\mathrm{E}[XY] \neq \mathrm{E}[X] \cdot \mathrm{E}[Y]$ for random variables X and Y; the difference is by definition the covariance between X and Y (see Section **??**). However, if X and Y are independent, then $\mathrm{E}[XY] = \mathrm{E}[X] \cdot \mathrm{E}[Y]$.

Proof A.4. If X and Y are independent, we can write $p_{X,Y}(x,y) = p_X(x) \cdot p_Y(y)$, and, thus:

$$\begin{aligned}\mathrm{E}\left[XY\right] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} XY \cdot p_X(x) \cdot p_Y(y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty} X \cdot p_X(x)\,\mathrm{d}x \cdot \int_{-\infty}^{\infty} Y \cdot p_Y(y)\,\mathrm{d}y \\ &= \mathrm{E}[X] \cdot \mathrm{E}[Y]\end{aligned}$$

□

Expectation of a matrix If $\mathbf{X}$ is a $m \times n$ matrix whose coefficients are random variables, then the expectation of $\mathbf{X}$ is defined as:

$$\mathrm{E}\left[\mathbf{X}\right] = \mathrm{E}\left(\begin{bmatrix} x_{1,1} & \cdots & X_{1,n} \\ x_{2,1} & \cdots & X_{2,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \cdots & X_{m,n} \end{bmatrix}\right) = \begin{bmatrix} \mathrm{E}(X_{1,1}) & \cdots & \mathrm{E}(X_{1,n}) \\ \mathrm{E}(X_{2,1}) & \cdots & \mathrm{E}(X_{2,n}) \\ \vdots & \ddots & \vdots \\ \mathrm{E}(X_{m,1}) & \cdots & \mathrm{E}(X_{m,n}) \end{bmatrix} \tag{A.6}$$

Transformation of the probabilistic basis Let $\mathrm{T} : Z \to X$ be the transformation of random variable Z to random variable X, and let $\mathrm{T}^{-1} : X \to Z$ be the corresponding inverse transformation. We can write:

$$\mathrm{E}_X[h(X)] = \mathrm{E}_Z[h(T(Z))] \tag{A.7}$$

The proof makes use of *integration by substitution.*

Proof A.5. Let $\Gamma_X = [a_X, b_X]$ be the support of X, and $\Gamma_Z = [a_Z, b_Z]$ be the support of Z.

$$\begin{aligned}\mathrm{E}_X[h(X)] &= \int_{a_X}^{b_X} h(x) \cdot p_X(x)\,\mathrm{d}x \\ &= \int_{\mathrm{T}^{-1}(a_X)}^{\mathrm{T}^{-1}(b_X)} h(\mathrm{T}(z)) \cdot p_X(\mathrm{T}(z)) \cdot \frac{p_Z(z)}{p_X(\mathrm{T}(z))}\,\mathrm{d}z \\ &= \int_{\mathrm{T}^{-1}(a_X)}^{\mathrm{T}^{-1}(b_X)} h(\mathrm{T}(z)) \cdot p_Z(z)\;\mathrm{d}z\end{aligned}$$

□

A.2 Variance

The *variance* of random variable X is defined as:

$$\sigma_X{}^2 = \mathrm{Var}[X] = \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] = \int_{-\infty}^{\infty} (x - \mu_X)^2 \cdot p_X(x)\,\mathrm{d}x \tag{A.8}$$

The *variance* of X is sometimes also referred to as the *second central moment* of X. It describes how dispersed the distribution of X is; the larger the variance the more dispersed it is.

In the following some properties of the variance are given. Let X be a random variable, and $c \in \mathbb{R}$ denote a constant. Then, we can write:

- $\mathrm{Var}[X] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$

  Proof A.6.

$$\begin{aligned}\mathrm{Var}[X] &= \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] \\ &= \mathrm{E}\left[X^2 - 2 \cdot X \cdot \mathrm{E}[X] + (\mathrm{E}[X])^2\right] \\ &= \mathrm{E}[X^2] - 2 \cdot (\mathrm{E}[X])^2 + (\mathrm{E}[X])^2 \\ &= \mathrm{E}[X^2] - (\mathrm{E}[X])^2\end{aligned}$$

□

- $\mathrm{Var}[c] = 0$

  Proof A.7.

$$\mathrm{Var}[c] = \mathrm{E}[c^2] - (\mathrm{E}[c])^2 = 0$$

□

- $\mathrm{Var}[X + c] = \mathrm{Var}[X]$

  Proof A.8.

$$\begin{aligned}\mathrm{Var}[X + c] &= \mathrm{E}[(X + c)^2] - (\mathrm{E}[X + c])^2 \\ &= \mathrm{E}[X^2 + 2 \cdot c \cdot X + c^2] - (\mathrm{E}[X] + c)^2 \\ &= \mathrm{E}[X^2] + 2 \cdot c \cdot \mathrm{E}[X] + c^2 - (\mathrm{E}[X])^2 - 2 \cdot c \cdot \mathrm{E}[X] - c^2 \\ &= \mathrm{E}[X^2] - (\mathrm{E}[X])^2 \\ &= \mathrm{Var}[X]\end{aligned}$$

□

- $\mathrm{Var}[cX] = c^2 \, \mathrm{Var}[X]$

  Proof A.9.

$$\begin{aligned}\mathrm{Var}[cX] &= \mathrm{E}[(cX)^2] - (\mathrm{E}[cX])^2 \\ &= c^2 \, \mathrm{E}[X^2] - c^2 \, (\mathrm{E}[X])^2 \\ &= c \left(\mathrm{E}[X^2] - (\mathrm{E}[X])^2\right) \\ &= c^2 \, \mathrm{Var}[X]\end{aligned}$$

□

Furthermore, if random variables X and Y are independent, we can write:

- $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$

  Proof A.10. For independent X and Y:

  $$\begin{aligned}\mathrm{Var}[X+Y] &= \mathrm{E}[(X+Y)^2] - (\mathrm{E}[X+Y])^2 \\ &= \mathrm{E}[X^2 + 2XY + Y^2] - (\mathrm{E}[X] + \mathrm{E}[Y])^2 \\ &= \mathrm{E}[X^2] + \mathrm{E}[Y^2] + 2\,\mathrm{E}[XY] - 2\,\mathrm{E}[X]\,\mathrm{E}[Y] - (\mathrm{E}[X])^2 - (\mathrm{E}[Y])^2 \\ &= \mathrm{E}[X^2] - (\mathrm{E}[X])^2 + \mathrm{E}[Y^2] - (\mathrm{E}[Y])^2 \\ &= \mathrm{Var}[X] + \mathrm{Var}[Y]\end{aligned}$$

  □

- $\mathrm{Var}[XY] = \mathrm{E}[X]^2\,\mathrm{Var}[Y] + \mathrm{E}[Y]^2\,\mathrm{Var}[X] + \mathrm{Var}[X]\,\mathrm{Var}[Y]$

  Proof A.11. For independent X and Y:

  $$\begin{aligned}\mathrm{Var}[XY] &= \mathrm{E}[(XY - \mathrm{E}[X]\,\mathrm{E}[Y])^2] \\ &= \mathrm{E}[X^2Y^2] - 2\,\mathrm{E}[X]^2\,\mathrm{E}[Y]^2 + \mathrm{E}[X]^2\,\mathrm{E}[Y]^2 \\ &= \mathrm{E}[X^2Y^2] - \mathrm{E}[X]^2\,\mathrm{E}[Y]^2 \\ &= \mathrm{E}[X^2]\,\mathrm{Var}[Y] \\ &= \big(\mathrm{Var}[X] + \mathrm{E}[X]^2\big)\,\mathrm{Var}[Y] \\ &= \mathrm{E}[X]^2\,\mathrm{Var}[Y] + \mathrm{E}[Y]^2\,\mathrm{Var}[X] + \mathrm{Var}[X]\,\mathrm{Var}[Y]\end{aligned}$$

  □

A.3 Standard deviation

The square root of the variance of random variable X is its *standard deviation* σ_X:

$$\sigma_X = \sqrt{\mathrm{Var}[X]} \tag{A.9}$$

As is the variance, the standard deviation is a measure for the variability of X. A small standard deviation indicates that realizations of X tend to be close to the mean of X; whereas a large standard deviation indicates that realizations of X have a large variability around the mean. Compared to the variance, the physical meaning of the standard deviation is often more intuitively interpretable, because the units of σ_X are the same as the units of X.

A.4 Coefficient of variation

The coefficient of variation c_v is defined as:

$$c_{v,X} = \frac{\sigma_X}{|\mu_X|} \tag{A.10}$$

It is a normalized measure for the dispersion of the distribution of X. Note that Eq. (A.10) is only defined for $\mu_X \neq 0$, and is particularly useful if all outcomes of X have the same sign.

A.5 Moments about zero

The nth moment of X about zero is defined as, for $n \in \{1, 2, \ldots\}$:

$$\mathrm{E}[X^n] = \int_{-\infty}^{\infty} x \cdot p_X(x)\,\mathrm{d}x \tag{A.11}$$

Note that the first moment about zero is the *mean* (see A.1). The second moment about zero is called the *mean square*.

A.6 Central moments

The nth central moment of X is defined as, for $n \in \{1, 2, \ldots\}$:

$$\mathrm{E}\left[(X - \mathrm{E}[X])^n\right] = \int_{-\infty}^{\infty} (x - \mathrm{E}[X])^2 \cdot p_X(x)\,\mathrm{d}x \tag{A.12}$$

Central moments are also referred to as *moments about the mean*; they are usually more interesting than the *moments about zero*: *central moments* can be used as measures for the dispersion of a distribution. Note: The first central moment is *zero*; the second central moment is called the variance (see A.2).

A.7 Normalized central moments

The normalized nth central moment of X is the nth central moment of X divided by $(\sigma_X)^n$, where σ_X is the standard deviation of X:

$$\frac{\mathrm{E}\left[(X - \mathrm{E}[X])^n\right]}{(\sigma_X)^n} \tag{A.13}$$

The normalized central moments are dimensionless quantities that are invariant to any linear change of scale. Note that the normalized first central moment is *zero*, and the normalized second central moment is *one* (if $\mathrm{E}[X]$ and $\mathrm{Var}[X]$ are defined for random variable X).

A.8 Skewness

The *skewness* γ_X of random variable X that has mean μ_X and standard deviation σ_X is often defined as the normalized 3rd central moment:

$$\gamma_X = \frac{\mathrm{E}\left[(X-\mu_X)^3\right]}{(\sigma_X)^3} = \frac{\mathrm{E}[X^3] - 3\cdot\mu_X\cdot{\sigma_X}^2 - {\mu_X}^3}{{\sigma_X}^3} \tag{A.14}$$

The skewness measures the asymmetry of a probability distribution. Loosely speaking, for unimodal distributions: If the left tail is longer, the skewness is negative. If the right tail is longer, the skewness is positive. However, if the other tail is heavy, this rule of thumb might not hold. For symmetric distributions, the skewness is *zero*; the reverse, however, is not true: *zero* skewness does not imply that the distribution is symmetric.

A.9 Kurtosis

The *excess kurtosis* κ_X of random variable X that has mean μ_X and standard deviation σ_X is defined as:

$$\kappa_X = \frac{\mathrm{E}\left[(X-\mu_X)^4\right]}{(\sigma_X)^4} - 3 \tag{A.15}$$

The -3 is used to set the kurtosis of the Normal distribution to *zero*. Loosely speaking, *kurtosis* measures the peakedness or tail weight of a distribution.

A.10 Percentile

The *p-percentile* of random variable X is the value X_p for which p (percent) of the outcomes of X are smaller or equal than X_p. The $10\%, 20\%, \ldots, 90\%$-percentiles are called the *deciles*.

Appendix B

Probability distributions

B.1 Common discrete probability distributions

B.1.1 Bernoulli distribution

The *Bernoulli* distribution is a discrete probability distribution. A Bernoulli distributed random variable is also referred to as *Bernoulli trial*. The realization of a Bernoulli trial can have exactly two states: 1 or 0, *success* or *failure*, *yes* or *no*, ...

Let p be the parameter of a Bernoulli distribution; it specifies the probability that the outcome of a Bernoulli trial is in the first state (e.g., the outcome is 1 or *success* or *yes*).

Properties

parameter $p \in (0, 1)$

support $k \in \{1, 0\}$

mean $= p$

standard deviation $= \sqrt{p \cdot (1-p)}$

entropy $= -p\ln(p) - (1-p)\ln(1-p)$

PMF

$$\Pr(k) = \begin{cases} p & \text{if } k = 1 \\ (1-p) & \text{if } k = 0 \end{cases} \tag{B.1}$$

B.1.2 Binomial distribution

Suppose we have a Bernoulli trial with success rate p (see Section B.1.1). The number K of successes in N independent homogeneous Bernoulli trials follows a *binomial distribution*.

Properties

parameters $p \in (0, 1)$, $N \in \mathbb{N}$
support $k \in \{0, \ldots, N\}$
mean $= p \cdot N$
standard deviation $= \sqrt{N \cdot p \cdot (1-p)}$
coefficient of variation $= \sqrt{\frac{1-p}{Np}}$
2nd moment about zero $\mathrm{E}[K^2] = Np\,((N-1)p+1)$
skewness $= 1/\sqrt{\lambda}$
excess kurtosis $= 1/\lambda$

PMF

$$\Pr(k) = \binom{N}{k} p^k (1-p)^{N-k} \tag{B.2}$$

where $\binom{N}{k}$ is the binomial coefficient. For the PMF of the Binomial distribution, the following property holds:

$$\Pr(k|N, p) = \Pr(N-k|N, 1-p) \tag{B.3}$$

Relation to other distributions

Bernoulli distribution For $N = 1$ the binomial distribution becomes a Bernoulli distribution.

Normal approximation A Normal distribution with mean Np and standard deviation $\sqrt{Np(1-p)}$ can be used to approximate the binomial distribution. However, this approximation is only reasonable if N is large enough and p is not close to 0 or 1. If p is close to either 0 or 1, the distribution can be highly skewed even for large N, and the Normal approximation might be not sufficient [Brown et al., 2001].

Poisson distribution The N goes to infinity and the expected number of successes remains fixed, the Poisson distribution with parameter $\lambda = Np$ can be derived as limiting case to the binomial distribution. If $N > 100$ and $Np \leq 10$ the quality of the approximation can typically be considered very good.

B.1.3 Negative binomial distribution

Suppose we have a Bernoulli trial with success rate p (see Section B.1.1). The required number N of independent homogeneous Bernoulli trials until K successes are reached follows a *negative binomial distribution*. When $K = 1$, the distribution becomes the *Geometric distribution*.

Properties

parameters $p \in (0, 1)$, $K \in \{1, 2, \ldots\}$

support $n \in \{K, K + 1, \ldots\}$

mean $= \frac{K}{p}$

standard deviation $= \frac{\sqrt{K(1-p)}}{p}$

coefficient of variation $= \sqrt{\frac{1-p}{K}}$

PMF

$$\Pr(n) = \binom{n-1}{K-1} p^K (1-p)^{n-K} \tag{B.4}$$

where $\binom{\cdot}{\cdot}$ is the binomial coefficient.

B.1.4 Poisson distribution

Suppose there are events which occur with a known average rate and the occurrence of an event is independent of the time since the last occurrence. The *Poisson distribution* is a discrete probability distribution that expresses the probability of a number K of events occurring in a specified period of time. The notion of time is arbitrary in this context; the number of events in other measures such as distance, area or volume can also be modeled by the Poisson distribution.

The Poisson distribution is described by a single parameter λ, where $\lambda \in \mathbb{R}^+$ equals the expected number of occurrences during the given interval. For example, if an event occurs on average 5 times per hour, and we are interested in the probability of the event occurring k times within 4 hours, the corresponding Poisson distribution has parameter $\lambda = 5 \cdot 4 = 20$.

Properties

parameter $\lambda \in \mathbb{R}^+$

support $k \in \mathbb{N}^0$

mean $= \lambda$

standard deviation $= \sqrt{\lambda}$

skewness $= \frac{1-2p}{\sqrt{np(1-p)}}$

excess kurtosis $= \frac{1-6p(1-p)}{np(1-p)}$

PMF

$$\Pr(k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{B.5}$$

If λ is sufficiently large, the Poisson distribution can be approximated by a Normal distribution that has mean λ and standard deviation $\sqrt{\lambda}$. For $\lambda > 1000$, the approximation can typically be considered very good; and for $\lambda > 10$ the approximation is decent. The probability that K is smaller or equal than k can be approximated based on the CDF of the standard Normal distribution as:

$$\Pr(K \leq k) \approx \Phi\left(\frac{k + 0.5 - \lambda}{\sqrt{\lambda}}\right) \tag{B.6}$$

where $k + 0.5$ is used instead of k as continuity correction.

B.2 Common continuous probability distributions

B.2.1 Standard Normal distribution

The *standard Normal* distribution is a symmetric continuous probability distribution and a special case of the Normal distribution (Section B.2.2): it has a mean of *zero* and a standard deviation of *one*.

In the following, let U be a random variable that follows a *standard Normal* distribution, and let u denote a particular outcome of U. The letter U is often used to denote a *standard Normal* random variable.

Properties

notation $U \sim \mathcal{N}(0, 1)$

support $u \in \mathbb{R}$

mean $\mu_U = 0$

standard deviation $\sigma_U = 1$

median $= 0$

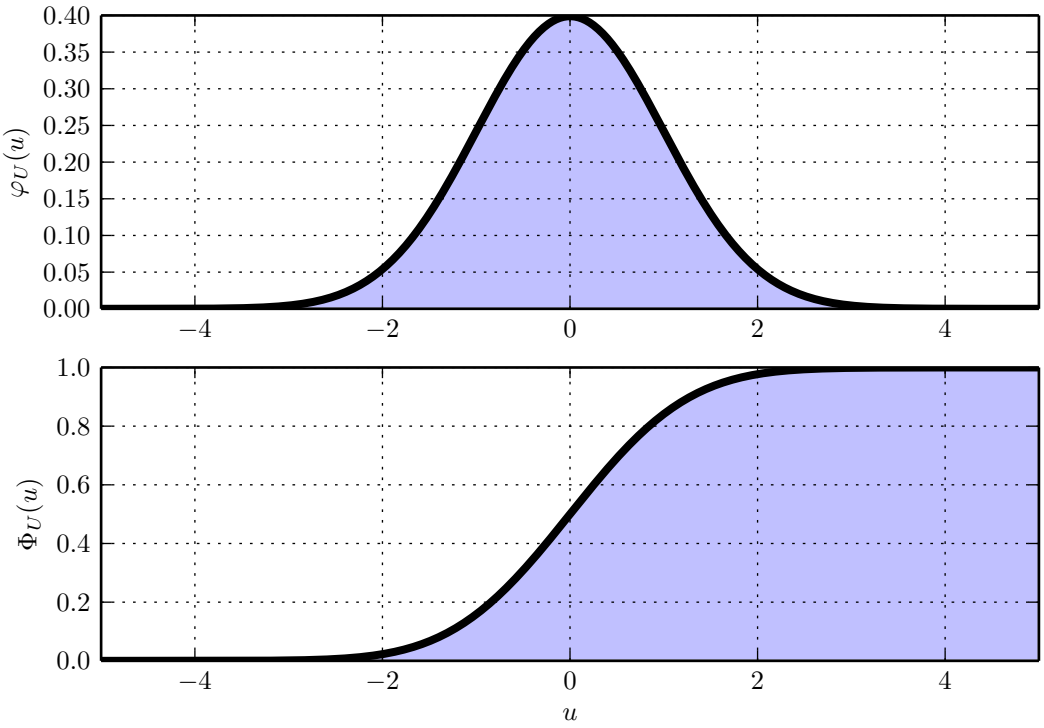


Figure B.1: PDF $\varphi_U(u)$ and CDF $\Phi_U(u)$ of the standard Normal distribution.

mode $= 0$

skewness $= 0$

excess kurtosis $= 0$

entropy $= \frac{1}{2}\ln(2\pi e)$

PDF

The PDF of the *standard Normal* distribution is commonly denoted by the Greek letter φ:

$$\varphi_U(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \tag{B.7}$$

The density has its maximum at $\varphi_U(0) \approx 0.4$; its inflection points are $\varphi_U(\pm 1) \approx 0.24$.

CDF

The CDF of the *standard Normal* distribution is commonly denoted by the Greek letter Φ:

$$\Phi_U(u) = \int_{-\infty}^{u} \varphi_U(t)\,\mathrm{d}t = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{u}{\sqrt{2}}\right)\right] \tag{B.8}$$

Here $\operatorname{erf}(\cdot)$ denotes the *error function*. The *standard Normal* has the following property:

- $\Phi_U(-u) = 1 - \Phi_U(u)$

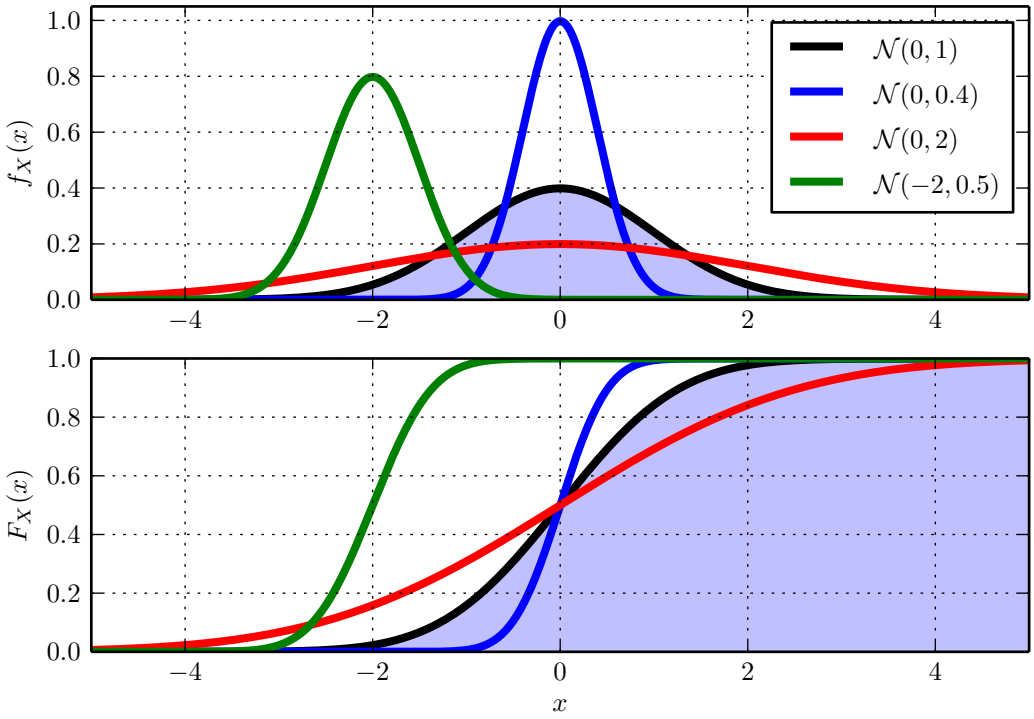


Figure B.2: Sample PDFs and CDFs of the Normal distribution.

Quantile function

The *quantile function* of the *standard Normal* distribution is expressed in terms of the *inverse error function* $\mathrm{erf}^{-1}(\cdot)$:

$$\Phi_U^{-1}(p) = \sqrt{2} \cdot \mathrm{erf}^{-1}(2p-1) \ , \quad p \in (0,1) \tag{B.9}$$

The PDF and CDF of the standard Normal distribution are illustrated in Fig. B.1.

B.2.2 Normal distribution

The *Normal* (or *Gaussian*) distribution is a widely used symmetric continuous probability distribution.

In the following, let X be a random variable that follows a *Normal* distribution with mean μ_X and standard deviation σ_X, and let x denote a particular outcome of X.

Properties

notation $X \sim \mathcal{N}(\mu_X, \sigma_X)$

parameters $\mu_X \in \mathbb{R}$, $\sigma_X \in (0, \infty)$

support $x \in \mathbb{R}$

mean μ_X

standard deviation σ_X

median $= \mu_X$

mode $= \mu_X$

skewness $= 0$

excess kurtosis $= 0$

entropy $= \frac{1}{2} \ln \left(2\pi e {\sigma_X}^2\right)$

PDF

The PDF $f_X(x)$ of the *Normal* distribution is commonly expressed in terms of the PDF of the standard Normal distribution $\varphi(\cdot)$:

$$f_X(x) = \frac{1}{\sigma_X} \cdot \varphi\left(\frac{x - \mu_X}{\sigma_X}\right) \tag{B.10}$$

The PDF of a *Normal* distribution can also be expressed as:

$$f_X(x) = c \cdot \exp\left(-\lambda_1 x - \lambda_2 x^2\right) \tag{B.11}$$

where the mean, standard deviation and scaling constant are:

$$\mu = -\frac{\lambda_1}{2\lambda_2} \qquad \lambda_1 = -\frac{\mu}{\sigma^2}$$
$$\sigma = \frac{1}{\sqrt{2\lambda_2}} \qquad \lambda_2 = \frac{1}{2\sigma^2}$$
$$c = \sqrt{\frac{\lambda_2}{\pi}} \cdot \exp\left(-\frac{\lambda_1^2}{4\lambda_2}\right)$$

From the equations above it follows that λ_2 must be positive.

CDF

The CDF $P_X(x)$ of the *Normal* distribution is typically expressed in terms of the CDF of the standard Normal distribution $\Phi(\cdot)$:

$$P_X(x) = \Phi\left(\frac{x - \mu_X}{\sigma_X}\right) \tag{B.12}$$

Quantile function

The *quantile function* of the *Normal* distribution is:

$$P_X^{-1}(p) = \mu_X + \sigma_X \cdot \Phi^{-1}(p)\,, \quad p \in (0, 1) \tag{B.13}$$

where $\Phi^{-1}(\cdot)$ is the quantile function of the standard Normal distribution.

Standardizing Normal random variables The Normal random variable X can be transformed to a standard Normal random variable Y through:

$$Y = \frac{X - \mu_X}{\sigma_X} \tag{B.14}$$

Conversely, a Normal random variable X with mean μ_X and standard deviation σ_X can be generated from a standard Normal variable as:

$$X = \mu_X + \sigma_X \cdot Y \tag{B.15}$$

Some PDFs and CDFs of the Normal distribution are shown in Fig. B.2.

Example B.1. *Normalized percentile range c_p as a function of the coefficient of variation δ_X:* We are interested in the range c_p that is spanned by the $1 - p/2$ and the $p/2$ percentile of X and that is normalized by $\mu_X = \mathrm{E}[X]$. Consequently, c_p is defined as:

$$\begin{aligned} c_p &= \frac{P_X^{-1}(1 - p/2) - P_X^{-1}(p/2)}{\mu_X} \\ &= \frac{\sigma_X \cdot \Phi^{-1}(1 - p/2) - \sigma_X \cdot \Phi^{-1}(p/2)}{\mu_X} \\ &= \delta_X \cdot \left[\Phi^{-1}(1 - p/2) - \Phi^{-1}(p/2)\right] \end{aligned} \tag{B.16}$$

where δ_X denotes the coefficient of variation of normal random variable X, and p is the probability contained in the normalized range spanned by the $1 - p/2$ and $p/2$ percentile of X. The above expression of c_p does not depend on μ_X. Moreover, the dependence between c_p and δ_X is linear conditionally on fixed p. Eq. (B.16) can be stated as:

$$c_p = \delta_X \cdot c_{p,1}(p) \tag{B.17}$$

with

$$c_{p,1}(p) = \Phi^{-1}(1 - p/2) - \Phi^{-1}(p/2) \tag{B.18}$$

The relation $c_{p,1}(p)$ is shown in Fig. B.3; some selected values of $c_{p,1}(p)$ are listed in Table B.1.

B.2.3 Truncated Normal distribution

The *truncated Normal* distribution is a *Normal* distribution whose support does not span the entire real line $\mathbb{R}$.

In the following, let X be a random variable that follows a *truncated Normal* distribution with parameters m and s, and support $[a, b]$. Furthermore, let x denote a particular outcome

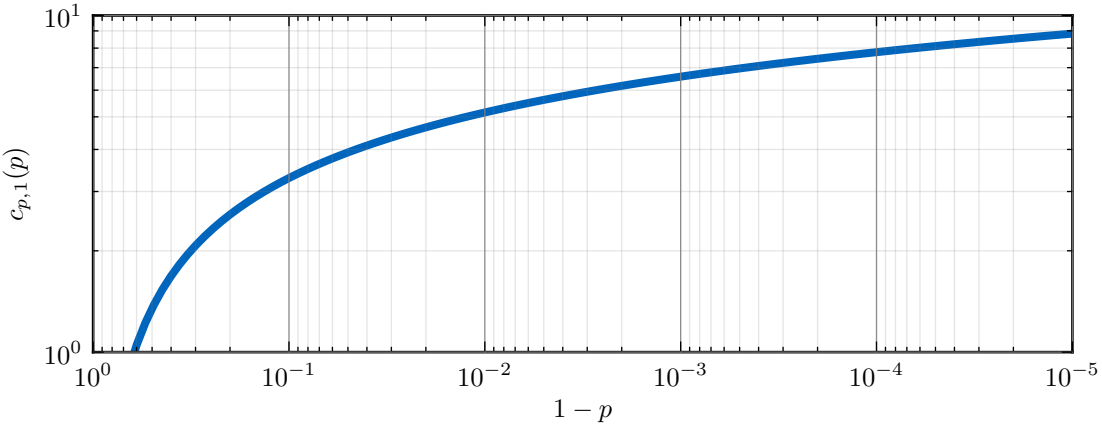


Figure B.3: Plot of Eq. (B.18). (Example B.1)

Table B.1: Values of $c_{p,1}(p)$ for selected values of p. (Example B.1)

p	$c_{p,1}(p)$
10%	0.25
25%	0.64
50%	1.3
75%	2.3
90%	3.3
95%	3.9
99%	5.2
99.9%	6.6

of X. We will use $\alpha = \frac{a-m}{s}$, $\beta = \frac{b-m}{s}$, $\xi = \frac{x-m}{s}$, and $q = \Phi(\beta) - \Phi(\alpha)$ to shorten the following equations.

Properties

parameters $m \in \mathbb{R}$, $s \in (0, \infty)$, $-\infty \leq a < b \leq \infty$

support $a \leq x \leq b$

mean $\mu_X = m + \frac{\varphi(\alpha) - \varphi(\beta)}{q} \cdot s$

standard deviation $\sigma_X = s \cdot \sqrt{1 + \frac{\alpha\varphi(\alpha) - \beta\varphi(\beta)}{q} - \left(\frac{\varphi(\alpha) - \varphi(\beta)}{q}\right)^2}$

mode $= \begin{cases} a & \text{if } m < a \\ m & \text{if } a \leq m \leq b \\ b & \text{if } m > b \end{cases}$

entropy $= \ln\left(\sqrt{2\pi e} s q\right) + \frac{\alpha\varphi(\alpha) - \beta\varphi(\beta)}{2q}$

PDF

The PDF $f_X(x)$ of the *truncated Normal* distribution is conveniently expressed in terms of the PDF of the standard Normal distribution $\varphi(\cdot)$:

$$f_X(x) = \frac{1}{sq} \cdot \varphi(\xi) \tag{B.19}$$

CDF

The CDF $F_X(x)$ of the *truncated Normal* distribution is conveniently expressed in terms of the CDF of the standard Normal distribution $\Phi(\cdot)$:

$$F_X(x) = \frac{\Phi(\xi) - \Phi(\alpha)}{q} \tag{B.20}$$

Parametrization in terms of μ_X and σ_X If the mean and standard deviation are given besides the support, the parameters m and s cannot be evaluated in a straight-forward manner: The values of m and s can be found solving an optimization problem.

The *truncated Normal* distribution approaches the *Normal* distribution if a approaches $-\infty$ and b approaches ∞.

B.2.4 Log-normal distribution

Taking the exponential of a Normal random variable with mean λ and standard deviation ζ gives a *log-normal* random variable; i.e. the natural logarithm of the random variable follows a Normal distribution.

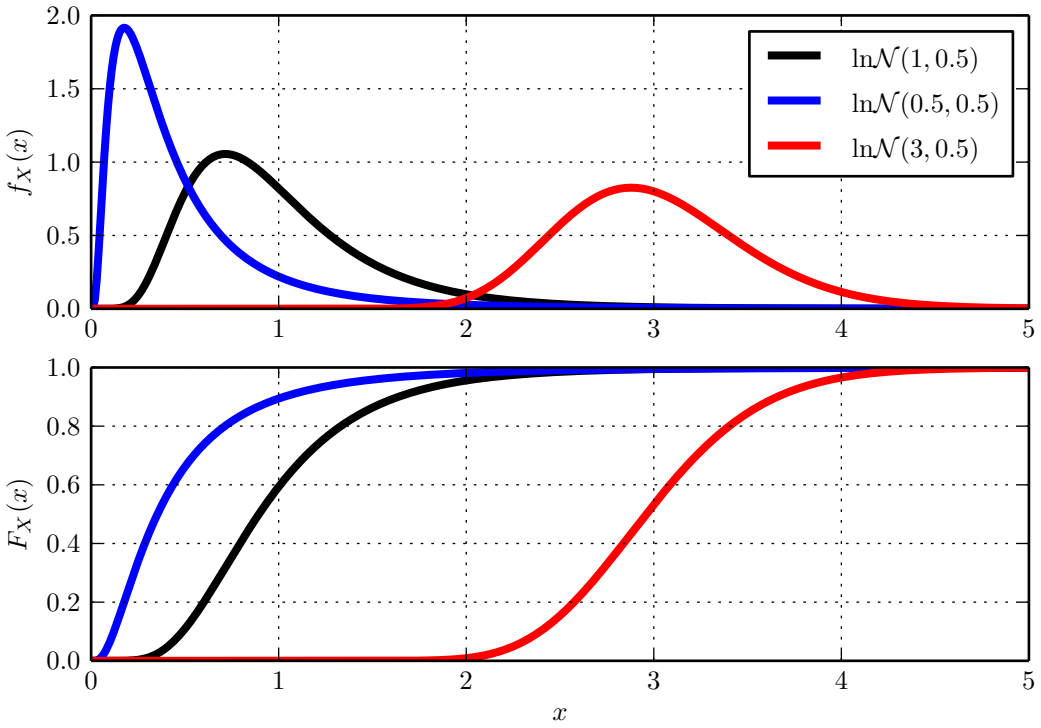


Figure B.4: Sample PDFs and CDFs of the log-normal distribution.

In the following, let X be a random variable that follows a *log-normal* distribution with parameters λ and ζ, and let x denote a particular outcome of X. If the distribution of X is shifted by parameter ε, the distribution is referred to as *shifted log-normal* distribution (If not mentioned explicitly, this parameter equals *zero*).

Properties

notation $X \sim \ln\mathcal{N}(\lambda, \zeta, \varepsilon) = \exp\left[\mathcal{N}\left(\lambda, \zeta\right)\right] + \varepsilon$

parameters $\lambda \in \mathbb{R}$, $\zeta \in (0, \infty)$, $\varepsilon \in \mathbb{R}$

support $x \in (\varepsilon, \infty)$

mean $\mu_X = \exp\left(\lambda + \frac{\zeta^2}{2}\right) + \varepsilon$

standard deviation $\sigma_X = \sqrt{\exp\left(\zeta^2\right) - 1} \cdot \exp\left(\lambda + \frac{\zeta^2}{2}\right)$

median $= \exp\left(\lambda\right) + \varepsilon$

mode $= \exp\left(\lambda - \zeta^2\right) + \varepsilon$

skewness $= \left(\exp\left(\zeta^2\right) + 2\right) \cdot \sqrt{\exp\left(\zeta^2\right) - 1}$

excess kurtosis $= \exp\left(4\zeta^2\right) + 2\exp\left(3\zeta^2\right) + 3\exp\left(2\zeta^2\right) - 6$

entropy $= \frac{1}{2} + \frac{1}{2}\ln\left(2\pi\zeta^2\right) + \lambda$

PDF

The PDF $f_X(x)$ of the *log-normal* distribution is:

$$f_X(x) = \frac{1}{(x-\varepsilon)\cdot\zeta\cdot\sqrt{2\pi}} \cdot \exp\left[-\frac{1}{2}\left(\frac{\ln(x-\varepsilon)-\lambda}{\zeta}\right)^2\right] \tag{B.21}$$

The PDF $f_X(x)$ can also be expressed in terms of the PDF of the standard Normal distribution $\varphi(\cdot)$:

$$f_X(x) = \frac{1}{(x-\varepsilon)\cdot\zeta} \cdot \varphi\left(\frac{\ln(x-\varepsilon)-\lambda}{\zeta}\right) \tag{B.22}$$

CDF

The CDF $F_X(x)$ of the *log-normal* distribution is defined as:

$$F_X(x) = \Phi\left(\frac{\ln(x-\varepsilon)-\lambda}{\zeta}\right) \tag{B.23}$$

where $\Phi(\cdot)$ is the CDF of the standard Normal distribution.

Quantile function

The *quantile function* of the *log-normal* distribution is:

$$F_X^{-1}(p) = \exp\left(\lambda + \zeta\cdot\Phi^{-1}(p)\right) + \varepsilon\ , \quad p\in(0,1) \tag{B.24}$$

where $\Phi^{-1}(\cdot)$ is the quantile function of the standard Normal distribution.

Standardizing log-normal random variables

A log-normal random variable X can be transformed to a standard Normal random variable Y through:

$$Y = \frac{\ln(X-\varepsilon)-\lambda}{\zeta} \tag{B.25}$$

Conversely, a log-normal random variable X with parameters λ and ζ can be generated from a standard Normal variable as:

$$X = \exp\left(\lambda + \zeta\cdot Y\right) + \varepsilon \tag{B.26}$$

Parametrization in terms of μ_X and σ_X

If the mean and standard deviation are given, the parameters λ and ζ can be derived as:

$$\lambda = \ln\left(\mu_X-\varepsilon\right) - \frac{1}{2}\ln\left(\frac{\sigma_X^2}{(\mu_X-\varepsilon)^2}+1\right) \tag{B.27}$$

$$\zeta = \sqrt{\ln\left(\frac{\sigma_X^2}{(\mu_X-\varepsilon)^2}+1\right)} \tag{B.28}$$

Covariance of two correlated log-normal random variables

The correlation coefficient ρ between two correlated log-normal random variables X_1 and X_2 is

$$\rho = \frac{\mathrm{Cov}(X_1, X_2)}{\sigma_1 \sigma_2} = \frac{\exp\left(\rho' \sqrt{\ln(\delta_1^2 + 1)\ln(\delta_2^2 + 1)}\right) - 1}{\delta_1 \delta_2} \tag{B.29}$$

where δ_1 and δ_2 is the coefficient of variation of X_1 and X_2, respectively; ρ' is the correlation coefficient of the underlying standard Normal random variables.

Some PDFs and CDFs of the log-normal distribution are shown in Fig. B.4.

B.2.5 Uniform distribution

The PDF of the uniform distribution is constant on the interval $[a, b]$, where a and b denote the lower and upper bound of plausible values, respectively.

In the following, let X be a random variable that follows a *uniform* distribution with parameters a and b, and let x denote a particular outcome of X.

Properties

notation $X \sim \mathcal{U}(a, b)$

parameters $-\infty < a < b < \infty$

support $x \in [a, b]$

mean $\mu_X = \frac{1}{2}(a + b)$

standard deviation $\sigma_X = \frac{1}{\sqrt{12}}(b - a)$

median $= \mu_X$

mode $= \mu_X$

skewness $= 0$

excess kurtosis $= -\frac{6}{5}$

entropy $= \ln(b - a)$

PDF

The PDF $f_X(x)$ of the *uniform* distribution is:

$$f_X(x) = \frac{1}{b - a}\,, \quad a \le x \le b \tag{B.30}$$

CDF

The CDF $F_X(x)$ of the *uniform* distribution is defined as:

$$F_X(x) = \frac{x-a}{b-a} \tag{B.31}$$

Quantile function

The *quantile function* of the *uniform* distribution is:

$$F_X^{-1}(p) = p \cdot (b-a) + a\,, \quad p \in (0,1) \tag{B.32}$$

Parametrization in terms of μ_X and σ_X

If the mean and standard deviation are given, the parameters a and b can be derived as:

$$a = \mu_X - \sigma_X\sqrt{3} \tag{B.33}$$

$$b = \mu_X + \sigma_X\sqrt{3} \tag{B.34}$$

B.2.6 Beta distribution

In the following, let X be a random variable that follows a *beta* distribution with shape parameters α and β. The support of X is (a, b). A particular outcome of X is denoted as x.

Properties

parameters $\alpha \in (0,\infty)$, $\beta \in (0,\infty)$

support $x \in [a, b]$

mean $\mu_X = \frac{\alpha}{\alpha+\beta} \cdot (b-a) + a$

standard deviation $\sigma_X = \sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}} \cdot (b-a)$;

mode $= \frac{\alpha-1}{\alpha+\beta-2}$ for α and β larger than one

PDF

The PDF $f_X(x)$ of the *beta* distribution is:

$$f_X(x) = \frac{w^{\alpha-1}(1-w)^{\beta-1}}{\mathrm{B}(\alpha,\beta)\cdot(b-a)}\,, \quad a \le x \le b \tag{B.35}$$

where w is defined as $w = \frac{x-a}{b-a}$.

CDF

The CDF $F_X(x)$ of the *beta* distribution is defined as:

$$F_X(x) = I_w\,(\alpha,\beta) \tag{B.36}$$

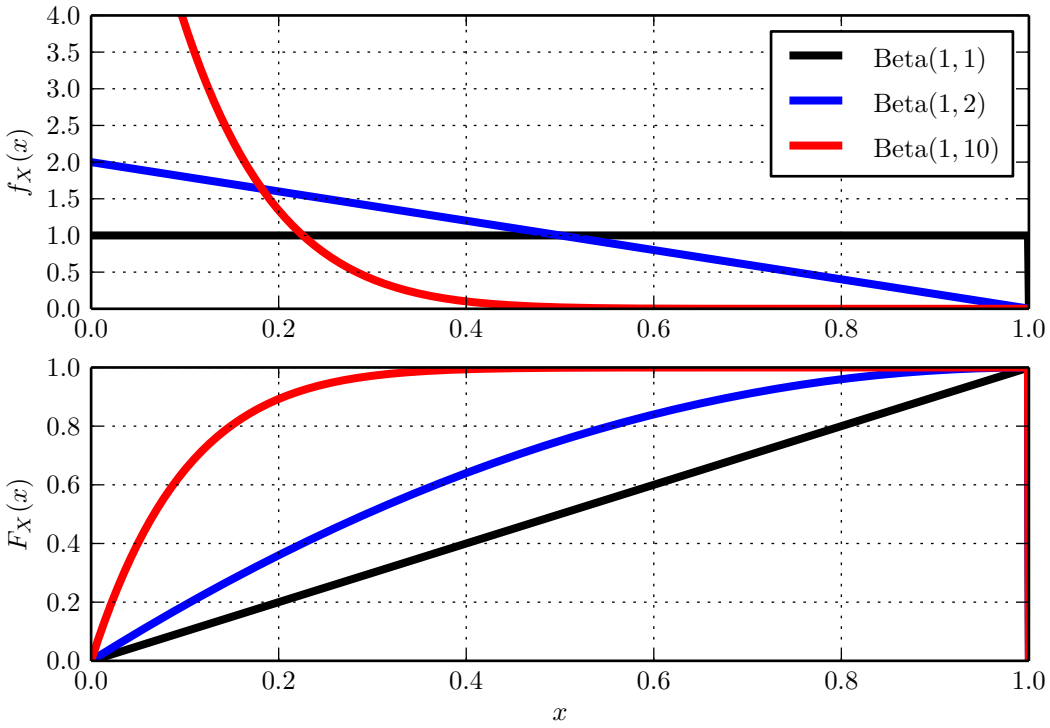


Figure B.5: Sample PDFs and CDFs of the beta distribution.

where $I_w(\alpha, \beta)$ denotes the regularized incomplete beta function, and $w = \frac{x-a}{b-a}$.

Some PDFs and CDFs of the *beta* distribution are shown in Fig. B.5. For $\alpha = \beta = 1$, the beta distribution is equivalent to the uniform distribution.

B.2.7 Extreme value distributions

B.2.7.1 Introduction

Let $\{X_1, \ldots, X_N\}$ denote a set of random variables. Furthermore, the smallest and largest value within this set is denoted as Y_1 and Y_N, respectively:

$$Y_1 = \min(X_1, \ldots, X_N) \tag{B.37}$$

$$Y_N = \max(X_1, \ldots, X_N) \tag{B.38}$$

B.2.7.2 Maxima

If the individual $X_i, i \in \{1, \ldots, N\}$ are independent and identically distributed, the CDF for Y_N, denoted $P_{Y_N}(\cdot)$ can be derived analytically:

$$P_{Y_N}(y) = [P_X(y)]^N \tag{B.39}$$

where P_X denotes the CDF of the X_i. For the more general case that the joint distribution of the set $\{X_1, \ldots, X_N\}$ is known, $P_{Y_N}(\cdot)$ can be expressed as:

$$P_{Y_N}(y) = P_{X_1}(y) \cdot P_{X_2|X_1}(y) \cdot \ldots \cdot P_{X_N|X_1,\ldots,X_{N-1}}(y) \tag{B.40}$$

To derive asymptotic extreme value distributions, a standardization of the Y_n is required (as $Y_N \to \infty$ if the support of the X_i is unbounded for the upper tail):

$$Z_N = \frac{Y_N - b_N}{a_N} \tag{B.41}$$

where $a_N > 0$ is referred to as a *scale parameter* and b_N is a *location parameter*. A distribution for X with CDF P_X is referred to as *extreme value* distribution, if the distribution of its maxima Y_N (for independent and identically distributed samples) has the same type with different location and scale parameters:

$$P_{Y_N}(y) = [P_X(y)]^N = P_X\left(\frac{y - b_N}{a_N}\right) \tag{B.42}$$

This statement is known as *stability postulate*. It can be shown that only three general types of asymptotic extreme value distributions exist: Gumbel distribution (Type I), Frechet distribution (Type II), Weibull distribution (Type III). The upper tail of the underlying distribution for the X_i determines to which of the three types Y_N converges asymptotically.

B.2.7.3 Minima

For the minima Y_1, the same relations as for the maxima can be derived. For independent and identically distributed X_i, the CDF $P_{Y_1}(\cdot)$ for Y_1 can be derived as:

$$P_{Y_1}(y) = [1 - P_X(y)]^N \tag{B.43}$$

In analogy to Y_N, three general forms of extreme value distributions can be derived for Y_1. The distribution types are directly related to the types for maxima:

$$P_{Y_1}(y) = 1 - P_{Y'_N}(-y). \tag{B.44}$$

where Y'_N is defined as the maxima of the set $\{X'_1, \ldots, X'_N\}$, with $X'_i = -X'_i$.

B.2.8 Gumbel distribution (Type I extreme value distribution for maxima)

The maxima of random variables that have an upper tail with exponential decay converges to the *Gumbel* distribution. This is the case for the Normal, log-Normal, exponential, gamma, logistic, Weibull and Gumbel disribution [Straub, 2016].

In the following, let X be a random variable that follows a *Gumbel* distribution with scale parameter a_N and location parameter b_N, and let x denote a particular outcome of X.

Properties

parameters scale: $a_N \in (0, \infty)$, location: $b_N \in \mathbb{R}$

support $x \in \mathbb{R}$

mean $b_N + a_N \cdot \gamma$

standard deviation $\sigma_X = \frac{\pi}{\sqrt{6}} \cdot a_N$

median $= b_N - a_N \cdot \ln(\ln(2))$

mode $= b_N$

entropy $= \ln(a_N) + \gamma + 1$

where $\gamma = 0.577\ldots$ is the Euler-Mascheroni constant.

PDF

The PDF $f_X(x)$ of the *Gumbel* distribution is:

$$f_X(x) = \frac{1}{a_N} \exp\left(-\frac{x - b_N}{a_N} - \exp\left(-\frac{x - b_N}{a_N}\right)\right) \tag{B.45}$$

CDF

The CDF $F_X(x)$ of the *Gumbel* distribution is defined as:

$$F_X(x) = \exp\left(-\exp\left(-\frac{x - b_N}{a_N}\right)\right) \tag{B.46}$$

Quantile function

The *quantile function* of the *Gumbel* distribution is:

$$F_X^{-1}(p) = b_N - a_N \cdot \ln(-\ln(p)) \ , \quad p \in (0, 1) \tag{B.47}$$

Parametrization in terms of μ_X and σ_X

If the mean and standard deviation are given, the parameters a_N and b_N can be derived as:

$$a_N = \frac{\sigma_X \cdot \sqrt{6}}{\pi} \tag{B.48}$$

$$b_N = \mu_X - \frac{\sigma_X \cdot \gamma \cdot \sqrt{6}}{\pi} \tag{B.49}$$

Distribution of maximum of N iid Gumbel distributed random variables

Let $X_i, i \in \{1, \ldots, N\}$ be N independent and identically distributed Gumbel random variables with scale parameter a and location parameter b. The distribution of the maximum is also a Gumbel distributed random variable with parameters:

$$a_N = a \tag{B.50}$$

$$b_N = b + a \cdot \ln(N) \tag{B.51}$$

B.2.9 Weibull distribution (Type III extreme value distribution for minima)

The minimum of random variables with a lower bound converges to the *Weibull* distribution.

In the following, let X be a random variable that follows a *Weibull* distribution with scale parameter a_N and shape parameter k, and let x denote a particular outcome of X.

Properties

parameters scale: $a_N \in (0, \infty)$, shape: $k \in (0, \infty)$

support $x \in (0, \infty)$

mean $a_N \cdot \Gamma(1 + 1/k)$

standard deviation $\sigma_X = a_N \cdot \sqrt{\Gamma(1 + 2/k) - (\Gamma(1 + 1/k))^2}$

median $= a_N (\ln(2))^{1/k}$

entropy $= \gamma \cdot (1 - 1/k) + \ln(a_N/k) + 1$

where $\gamma = 0.577\ldots$ is the Euler-Mascheroni constant, and $\Gamma(\cdot)$ is the Gamma function.

PDF

The PDF $f_X(x)$ of the *Weibull* distribution is:

$$f_X(x) = \frac{k}{a_N} \left(\frac{x}{a_N}\right)^{k-1} \exp\left[-\left(\frac{x}{a_N}\right)^k\right] \tag{B.52}$$

CDF

The CDF $F_X(x)$ of the *Weibull* distribution is defined as:

$$F_X(x) = 1 - \exp\left[-\left(\frac{x}{a_N}\right)^k\right] \tag{B.53}$$

Quantile function

The *quantile function* of the *Weibull* distribution is:

$$F_X^{-1}(p) = a_N \cdot [-\ln(1-p)]^{\frac{1}{k}} \quad , \quad p \in (0,1) \tag{B.54}$$

Distribution of maximum of N iid Weibull distributed random variables

Let $X_i, i \in \{1, \ldots, N\}$ be N independent and identically distributed Weibull random variables with scale parameter a and shape parameter k. The distribution of the maximum is also a Weibull distributed random variable with parameters:

$$a_N = \frac{a}{N^{\frac{1}{k}}} \tag{B.55}$$

$$k_N = k \tag{B.56}$$

Appendix C

Maximum entropy probability distributions – continuous case

C.1 Introduction

In the following, probability distributions are introduce for which the differential entropy (Section 2.4.4) is maximized given specified constraints.

Let $X \in \mathbb{R}$ be a stochastic variable. Furthermore, let h_i be n functions of X, where the expectations of the h_i are given as auxiliary conditions, i.e., $\mathrm{E}_X[h_i] = k_i$ for all $i = 1, \ldots, n$. Furthermore, let Γ denote the support of random variable X, where Γ is a closed subset of $\mathbb{R}$. The *maximum entropy probability distribution* $g(x)$ for X that is positive everywhere in Γ can be found using variational calculus [Boltzmann, 1877]: The following Lagrangian function with $n + 1$ Lagrange multipliers is defined:

$$L = \int_\Gamma g(x) \ln g(x) \, \mathrm{d}x + \lambda_0 \left(\int_\Gamma g(x) \, \mathrm{d}x - 1 \right) + \sum_{i=1}^{n} \lambda_i \left(\mathrm{E}_X[h_i] - k_i \right) \tag{C.1}$$

The *maximum entropy probability distribution* $g(x)$ fulfills the Euler-Lagrange equation:

$$\frac{\partial H}{\partial g} - \frac{\mathrm{d}}{\mathrm{d}x} \frac{\partial H}{\partial g'} = 0 \tag{C.2}$$

where $H(x)$ can be formulated based on Eq. (C.1) as:

$$H(x) = g(x) \ln g(x) + \lambda_0 \left(g(x) - 1 \right) + \sum_{i=1}^{n} \lambda_i \left(h_i(x) - k_i \right) \tag{C.3}$$

This leads to

$$\frac{\partial H}{\partial g} = \ln g(x) + 1 + \lambda_0 + \sum_{i=1}^{n} \lambda_i h_i(x) \tag{C.4}$$

$$\frac{\partial H}{\partial g'} = 0 \tag{C.5}$$

Thus, to maintain Eq. (C.2), $g(x)$ must be:

$$g(x) = c \cdot \exp\left(-\sum_{i=1}^{n} \lambda_i h_i(x) \right) \tag{C.6}$$

where the constant $c = \exp(-1 - \lambda_0)$ ensures that $\int_\Gamma g(x)\,\mathrm{d}x = 1$. The value of the λ_i can be obtained from the specified auxiliary conditions.

C.2 Specified bounds

If only the support of the distribution is given, but no additional auxiliary conditions, it is easy to deduce from Eq. (C.6) that the maximum entropy probability distribution is the uniform distribution; i.e., $g(x) = c$, with $c = \frac{1}{b-a}$.

C.3 Specified mean and bounds

We search the maximum entropy probability distribution for a distribution $g(x)$ that has given mean μ and support $x \in [a, b]$, where $a < b$ and $a, b \in \mathbb{R}$.

The distribution $g(x)$ is subjected to the constraint $\mathrm{E}_X[X] = \mu$. Thus, $g(x)$ has shape:

$$g(x) = c \cdot \exp(-\lambda_1 x) \tag{C.7}$$

The constant c can be derived from the condition:

$$\int_a^b g(x)\,\mathrm{d}x = 1 \tag{C.8}$$

... which leads to ...:

$$c = \frac{\lambda_1}{\exp(-\lambda_1 a) - \exp(-\lambda_1 b)} \tag{C.9}$$

$$\mu = \frac{(\lambda_1 a + 1)\exp(-\lambda_1 a) - (\lambda_1 b + 1)\exp(-\lambda_1 b)}{\lambda_1 \left(\exp(-\lambda_1 a) - \exp(-\lambda_1 b)\right)} \tag{C.10}$$

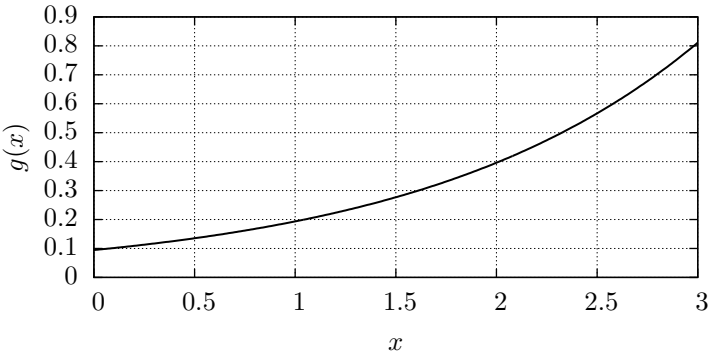


Figure C.1: Maximum entropy probability distribution for mean 2 and support $[0, 3]$. (Example C.2)

Example C.1. *Positive and given mean*:
For support $S = [0, \infty)$, the distribution becomes the *exponential distribution*:

$$g(x) = \frac{1}{\mu} \exp\left(-\frac{x}{\mu}\right) \tag{C.11}$$

Example C.2. *Support* $[0, 3]$ *and mean* 2:
If the mean of the distribution is fixed to 2, and the support is $[0, 3]$, λ_1 evaluates to approximately 0.72. The corresponding maximum entropy probability distribution $g(x)$ is shown in Fig. C.1.

C.4 Specified mean, standard deviation and bounds

We search the maximum entropy probability distribution for a distribution $g(x)$ that has given mean μ and standard deviation σ. The support of $g(x)$ is $[a, b]$, where $a < b$ and $a, b \in \mathbb{R}$.

The distribution $g(x)$ is subjected to the constraints $\mathrm{E}[X] = \mu$ and $\mathrm{E}[X^2] = \sigma^2 + \mu^2$. Consequently, $g(x)$ has shape:

$$g(x) = c \cdot \exp\left(-\lambda_1 x - \lambda_2 x^2\right) \tag{C.12}$$

The distribution $g(x)$ stated in Eq. (C.12) is a *truncated Normal* distribution that can be rewritten as:

$$g(x) = \frac{1}{sq} \cdot \varphi(\xi) \tag{C.13}$$

where $s = \frac{1}{\sqrt{2\lambda_2}}$, $\xi = \frac{x-m}{s}$, $m = -\frac{\lambda_1}{2\lambda_2}$, $q = \Phi(\beta) - \Phi(\alpha)$, $\alpha = \frac{a-m}{s}$ and $\beta = \frac{b-m}{s}$.

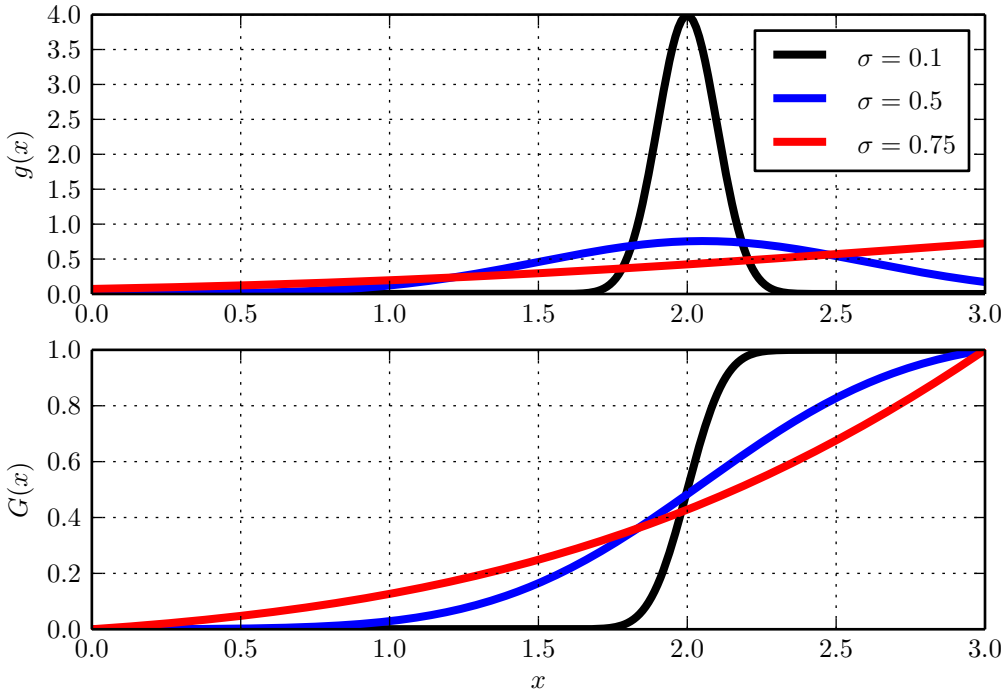


Figure C.2: Maximum entropy probability distribution for mean 2 and standard deviations 0.1, 0.5 and 0.75 on bounded support $[0, 3]$. (Example C.3)

Example C.3. *Support* $[0, 3]$*, mean* 2 *and varying standard deviation*:
We set the support to $[0, 3]$ and fix the mean to 2. The maximum entropy probability distributions for standard deviations 0.1, 0.5 and 0.75 are illustrated in Fig. C.2.

C.5 Positive and specified mean and standard deviation

The maximum entropy probability distribution for given mean and standard deviation on a positive support is a special case of Section C.4: It is a *truncated Normal* distribution.

Example C.4. *Positive support, mean* 1 *and varying standard deviation*:
We restrict the support to positive values and fix the mean to 1. The maximum entropy probability distributions for standard deviations 1, 0.75 and 0.5 are illustrated in Fig. C.3.

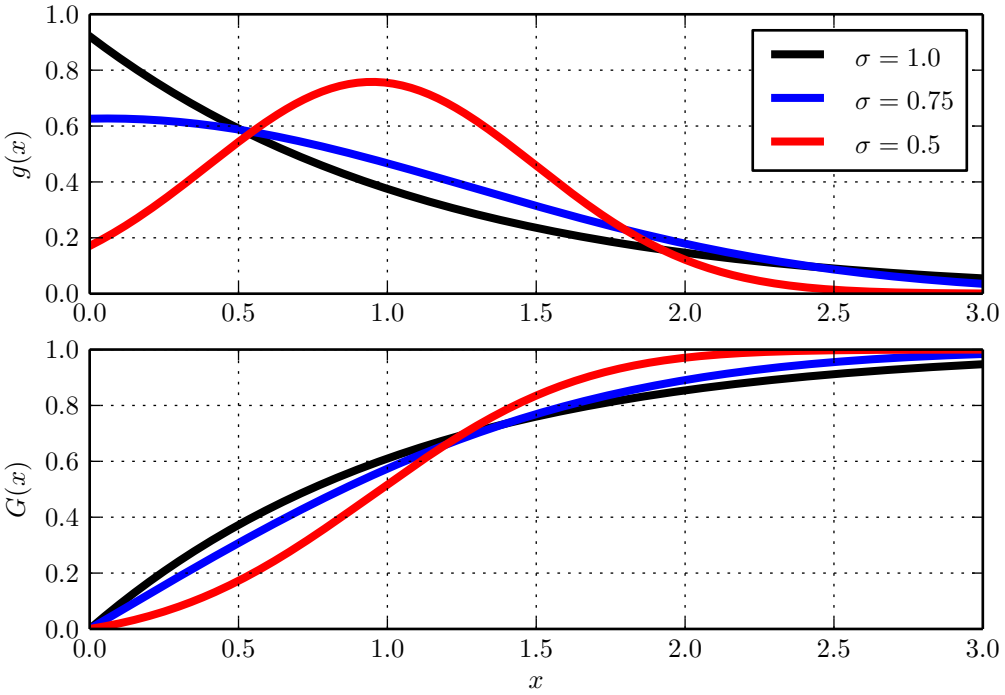


Figure C.3: Maximum entropy probability distribution for mean 1 and standard deviations 1, 0.75 and 0.5 on support $[0, -\infty)$. (Example C.4)

C.6 Specified mean, standard deviation

The maximum entropy probability distribution for given mean and standard deviation on support $(-\infty, \infty)$ can be easily deduced from Eq. (C.12): It is a *Normal* distribution.

Appendix D

Stochastic fields

This section contains material originally published in [Betz et al., 2014c]. Some passages and figures are directly taken from the mentioned reference.

D.1 General introduction

A *stochastic field* (or *random field*) is the extension of a stochastic process to dimensions larger *one*. A univariate continuous stochastic field $X_{\mathbf{t}}$ may be loosely defined as a stochastic function that describes a stochastic variable at each point $\mathbf{t} \in \Omega$ of a continuous domain $\Omega \subset \mathbb{R}^N$. The dimension $N \in \mathbb{N}_{>0}$ of a stochastic field is the dimension of its topological space Ω. If the stochastic field $\mathbf{X_t}$ describes a stochastic vector at each point $\mathbf{t}$ in Ω, the field is called multivariate.

The field is said to be Gaussian if the distribution of $(X_{\mathbf{t}_1}, \ldots, X_{\mathbf{t}_n})$ is jointly Gaussian for any $(\mathbf{t}_1, \ldots, \mathbf{t}_n) \in \Omega$ and any $n \in \mathbb{N}_{>0}$. A Gaussian field is completely defined by its mean function $\mu : \Omega \to \mathbb{R}$ and autocovariance function $\mathrm{Cov} : \Omega \times \Omega \to \mathbb{R}$. The autocovariance function can be expressed as $\mathrm{Cov}(\mathbf{t}_1, \mathbf{t}_2) = \sigma(\mathbf{t}_1) \cdot \sigma(\mathbf{t}_2) \cdot \rho(\mathbf{t}_1, \mathbf{t}_2)$, where $\sigma : \omega \to \mathbb{R}$ is the standard deviation function of the stochastic field and $\rho : \Omega \times \Omega \to \mathbb{R}$ is its autocorrelation coefficient function.

D.2 Random field discretization

A continuous stochastic field represents a stochastic quantity at each point of a continuous domain, and, thus, consists of an infinite number of stochastic variables. For computational purposes, the stochastic field has to be expressed using a finite number of stochastic variables. The approximation $\hat{H}(\cdot)$ of a continuous random field $H(\cdot)$ by a finite set of random variables

$\{\chi_i, i = 1, \ldots, M\}$ with $M \in \mathbb{N}_{>0}$ is referred to as *random field discretization*. Efficient methods to discretize stochastic fields on potentially complicated domains are discussed in [Betz et al., 2014c].

The approximation error $\varepsilon_H(\mathbf{t})$ of the random field discretization is defined as the difference between the original field and its approximation, i.e., $\varepsilon_H(\mathbf{t}) = H(\mathbf{t}) - \hat{H}(\mathbf{t})$. The expectation of the squared approximation error is called the *mean square error*. Integration of the mean square error over the domain Ω gives the *global mean square error* [Ghanem and Spanos, 1991]:

$$\overline{\varepsilon}_H^2 = \int_\Omega \mathrm{E}\left[(\varepsilon_H(\mathbf{t}))^2\right] \,\mathrm{d}\mathbf{t} \tag{D.1}$$

An alternative error measure for random field discretization is the normalized variance of the approximation error, denoted $\varepsilon_\sigma(\mathbf{t})$ [Li and Der Kiureghian, 1993]:

$$\varepsilon_\sigma(\mathbf{t}) = \frac{\mathrm{Var}\left[H(\mathbf{t}) - \hat{H}(\mathbf{t})\right]}{\mathrm{Var}\left[H(\mathbf{t})\right]} \tag{D.2}$$

$\varepsilon_\sigma(\mathbf{t})$ is called *error variance* in literature. The corresponding global error measure, namely the *mean error variance*, is defined as the weighted integral [Sudret and Der Kiureghian, 2000]:

$$\overline{\varepsilon}_\sigma = \frac{1}{|\Omega|} \int_\Omega \varepsilon_\sigma(\mathbf{t}) \,\mathrm{d}\mathbf{t} \tag{D.3}$$

where $|\Omega| = \int_\Omega \mathrm{d}\mathbf{t}$.

It is convenient to assume that the mean of the random field can be represented exactly. In this case, the expectation of the approximation error is zero, and the expectation of the squared approximation error is equivalent to the variance of the approximation error, i.e., $\mathrm{E}\left[(\varepsilon_H(\mathbf{t}))^2\right] = \mathrm{Var}\left[\varepsilon_H(\mathbf{t})\right]$. Consequently, the error variance is proportional to the mean square error. If the standard deviation of the field is constant on the domain Ω, i.e., $\sigma = \sigma(\mathbf{t}) \,\forall\, \mathbf{t} \in \Omega$, the global mean square error can be expressed in terms of the mean error variance as:

$$\overline{\varepsilon}_H^2 = |\Omega| \cdot \sigma^2 \cdot \overline{\varepsilon}_\sigma \tag{D.4}$$

D.3 Karhunen–Loève expansion of random fields

The Karhunen–Loève (KL) expansion is a series expansion method for the representation of the random field. The expansion is based on a spectral decomposition of the autocovariance function of the field. It states that a second-order random field $H(\mathbf{t})$ can be represented

exactly by the following expansion [Karhunen, 1947; Loève, 1948]:

$$H(\mathbf{t}) = \mu(\mathbf{t}) + \sum_{i=1}^{\infty} \sqrt{\lambda_i}\, \varphi_i(\mathbf{t})\, \xi_i \tag{D.5}$$

where $\mu(\mathbf{t})$ is the mean function of the field, ξ_i are standard uncorrelated random variables, and $\lambda_i \in [0, \infty)$, $\varphi_i : \Omega \to \mathbb{R}$ are the eigenvalues and eigenfunctions of the autocovariance kernel obtained from solving the homogeneous Fredholm integral equation of the second kind:

$$\int_{\Omega} \mathrm{Cov}(\mathbf{t}, \mathbf{t}')\, \varphi_i(\mathbf{t}')\, \mathrm{d}\mathbf{t}' = \lambda_i \varphi_i(\mathbf{t}) \tag{D.6}$$

In this context, the autocovariance function $\mathrm{Cov}(\mathbf{t}, \mathbf{t}')$ is also referred to as kernel function. Any valid covariance function is a bounded, symmetric and positive semi-definite kernel [Vanmarcke, 2010]. Moreover, a continuous kernel function is assumed. Note that the kernel does not have to be stationary. According to Mercer's theorem, the eigenvalues λ_i are nonnegative, the eigenfunctions corresponding to positive eigenvalues are continuous and orthogonal to each other, and the kernel function can be written as the uniformly convergent expansion $\mathrm{Cov}(\mathbf{t}, \mathbf{t}') = \sum_{i=1}^{\infty} \lambda_i\, \varphi_i(\mathbf{t})\, \varphi_i(\mathbf{t}')$, where the eigenfunctions in the expression are normalized. Consequently, the eigenfunctions must be orthonormal to each other, i.e., $\int_{\Omega} \varphi_i(\mathbf{t})\, \varphi_j(\mathbf{t})\, \mathrm{d}\mathbf{t} = \delta_{ij}$, where δ_{ij} is one if $i = j$ and zero otherwise. Moreover, they form a complete basis of the space $L^2(\Omega)$ of square integrable functions on Ω.

If the random field $H(\mathbf{t})$ is Gaussian, then ξ_i are independent standard normal random variables [Ghanem and Spanos, 1991]. In any other case, the joint distribution of ξ_i is almost impossible to obtain. Hence, the KL expansion is mainly applicable to the discretization of Gaussian fields.

The direct modeling of non-Gaussian random fields by means of the KL expansion was discussed by Phoon et al. [Phoon et al., 2002]. The authors proposed an iterative framework to simulate non-stationary non-Gaussian processes. The procedure was refined in [Phoon et al., 2005] for highly skewed non-Gaussian processes. Moreover, non-Gaussian fields are commonly modeled by combining the KL expansion with the polynomial chaos expansion. Ghanem [Ghanem, 1999] proposed a general framework in which the non-Gaussian field is projected onto an orthogonal polynomial basis with argument an underlying Gaussian field that is then discretized by the KL expansion. Matthies and Keese [Matthies and Keese, 2005] proposed to perform the KL expansion of the non-Gaussian field and project the random variables involved in the expansion to an underlying independent Gaussian random variable space. In Section D.8, we discuss the treatment of a special case of non-Gaussian random fields within the context of the KL expansion.

Analytical solutions of the IEVP can be obtained only for specific types of autocovariance functions defined on rectangular domains. For random fields with arbitrary autocovariance

functions defined on domains of complex geometrical shape, the solution of the IEVP needs to be approximated numerically. An overview of the numerical solution of Fredholm integral equations is given in [Atkinson, 1997].

D.4 Truncated KL expansion

The KL expansion can be approximated by sorting the eigenvalues λ_i and the corresponding eigenfunctions $\varphi_i(\mathbf{t})$ in a descending order and truncating the expansion after M terms:

$$\tilde{H}(\mathbf{t}) = \mu(\mathbf{t}) + \sum_{i=1}^{M} \sqrt{\lambda_i}\, \varphi_i(\mathbf{t})\, \xi_i \tag{D.7}$$

For fixed M, the resulting random field approximation $\tilde{H}(\mathbf{t})$ is optimal among series expansion methods with respect to the global mean square error (Eq. (D.1)) [Ghanem and Spanos, 1991]. The variance of $\tilde{H}(\mathbf{t})$ is given as

$$\mathrm{Var}\left[\tilde{H}(\mathbf{t})\right] = \sum_{i=1}^{M} \lambda_i\, \varphi_i^2(\mathbf{t}) \tag{D.8}$$

In case of the truncated KL expansion, the error variance introduced in Eq. (D.2) can be expressed as [Sudret and Der Kiureghian, 2000]:

$$\varepsilon_{\sigma,\mathrm{KL}}(\mathbf{t}) = 1 - \frac{\sum_{i=1}^{M} \lambda_i\, \varphi_i^2(\mathbf{t})}{\sigma^2(\mathbf{t})} \tag{D.9}$$

wherein the numerator in the fraction represents the variance of the truncated field. Eq. (D.9) can be derived by expressing $H(\mathbf{t})$ by its KL expansion and using the orthonormality of the random variables ξ_i. From Eq. (D.9) it can be deduced that the truncated KL expansion always underestimates the true variability of the original random field. This property of the KL expansion was also discussed in [Sudret and Der Kiureghian, 2000]. Moreover, as pointed out in [Stefanou and Papadrakakis, 2007], the truncated KL expansion of homogeneous random fields is only approximately homogeneous, since the standard deviation function of the truncated field will always vary in space. The mean error variance is given as:

$$\overline{\varepsilon}_{\sigma,\mathrm{KL}} = 1 - \frac{1}{|\Omega|} \sum_{i=1}^{M} \lambda_i \int_{\Omega} \frac{\varphi_i^2(\mathbf{t})}{\sigma^2(\mathbf{t})}\, \mathrm{d}\mathbf{t} \tag{D.10}$$

The equation of the mean square error can be transformed to $\mathrm{E}\left[\left(\varepsilon_H(\mathbf{t})\right)^2\right] = \sigma^2(\mathbf{t}) - \sum_{i=1}^{M} \lambda_i\, \varphi_i^2(\mathbf{t})$ using the orthonormality of the random variables ξ_i. The global mean square

error reads [Sudret and Der Kiureghian, 2000]:

$$\overline{\varepsilon}^2_{H,\mathrm{KL}} = \int_\Omega \sigma^2(\mathbf{t})\,\mathrm{d}\mathbf{t} - \sum_{i=1}^{M} \lambda_i \tag{D.11}$$

If the standard deviation of the field is constant, the equation for the mean error variance, Eq. (D.10), reduces to [Sudret and Der Kiureghian, 2000]:

$$\overline{\varepsilon}_{\sigma,\mathrm{KL}} = 1 - \frac{1}{|\Omega|}\frac{1}{\sigma^2}\sum_{i=1}^{M} \lambda_i \tag{D.12}$$

For this special case, the truncated KL expansion is also optimal with respect to the mean error variance.

D.5 Numerical solution of the KL expansion

Integral eigenvalue problems of the type given in Eq. (D.6) are difficult to solve analytically except for a few autocovariance functions defined on domains Ω of simple geometric shape. Analytical solutions for exponential and triangular kernels are discussed in [Ghanem and Spanos, 1991] for one-dimensional domains. Extensions to multidimensional rectangular domains can be derived assuming a separable covariance structure (e.g., see [Sudret and Der Kiureghian, 2000]). In general, the integral eigenvalue problem is solved numerically. The random field approximation of the truncated KL expansion given in Eq. (D.7) is approximated as:

$$\hat{H}(\mathbf{t}) = \mu(\mathbf{t}) + \sum_{i=1}^{M} \sqrt{\hat{\lambda}_i}\,\hat{\varphi}_i(\mathbf{t})\,\hat{\xi}_i \tag{D.13}$$

where $\hat{\lambda}_i$ and $\hat{\varphi}_i$ are approximations to the true eigenvalues λ_i and eigenfunctions φ_i. $\hat{\xi}_i$ are standard uncorrelated random variables, i.e., $\mathrm{E}\left[\hat{\xi}_i\,\hat{\xi}_j\right] = \delta_{ij}\ \forall\ i,j \leq M$. Note that due to the approximate character of the numerical solution, the random variables $\hat{\xi}_i$ are not necessarily orthogonal to the random variables ξ_i used in the representation of Eq. (D.5). This means that the expression for the error variance given in Eq. (D.9) cannot be derived from Eq. (D.2) for the numerical approximation of the KL expansion. Therefore, the error measures listed in section D.4 are not strictly valid for the approximated truncated KL expansion. Moreover, it is important to note that the random field approximation given in Eq. (D.13) does no longer possess the optimality property of the truncated analytical KL expansion.

Numerical algorithms for the solution of Fredholm integral eigenvalue problems approximate

the eigenfunctions by a set of functions $h_j : \Omega \to \mathbb{R}$ as:

$$\varphi_i(\mathbf{t}) \approx \hat{\varphi}_i(\mathbf{t}) = \sum_{j=1}^{N} d_j^i \, h_j(\mathbf{t}) \tag{D.14}$$

where the coefficients $d_j^i \in \mathbb{R}$ have to be determined. In general, all algorithms can be categorized into three main categories: degenerate kernel methods, Nyström methods, and projection methods. Projection methods can be further subdivided into collocation methods and Galerkin methods. An overview of this methods can be found in [Betz et al., 2014c].

D.6 Nyström method

The presentation of the Nyström method is given specifically in the following, as it has a close connection to the EOLE method (see Section D.7).

In the Nyström method [Atkinson, 1997], the integral in the eigenvalue problem of Eq. (D.6) is approximated by a numerical integration scheme. Applications to integral eigenvalue problems published in literature include [Hurtado, 2002; Wan and Karniadakis, 2006; Zhu et al., 2007]. Numerical algorithms are discussed in [Press et al., 1993; Atkinson and Shampine, 2008]. The problem is approximated as:

$$\sum_{j=1}^{N} w_j \, \mathrm{Cov}(\mathbf{t}, \mathbf{t}_j) \, \hat{\varphi}_i(\mathbf{t}_j) = \hat{\lambda}_i \hat{\varphi}_i(\mathbf{t}) \tag{D.15}$$

where $\mathbf{t}_j \in \Omega$ with $j \in \{1, \ldots, N\}$, $N \in \mathbb{N}$ represent a finite set of integration points, and w_j is the integration weight associated with each $\mathbf{t}_j$. For a given N, the distribution of the integration points $\mathbf{t}_j$ and the value of the integration weights w_j depend on the applied numerical integration scheme. Special integration techniques exist for kernels that are non-differentiable on the diagonal, see [Press et al., 1993; Atkinson and Shampine, 2008]. It is assumed that for the applied numerical integration scheme, the solution of Eq. (D.15) converges against the analytical solution with increasing N.

In the Nyström method, Eq. (D.15) is solved at the integration points, i.e.:

$$\sum_{j=1}^{N} w_j \, \mathrm{Cov}(\mathbf{t}_n, \mathbf{t}_j) \, \hat{\varphi}_i(\mathbf{t}_j) = \hat{\lambda}_i \hat{\varphi}_i(\mathbf{t}_n), \quad n = 1, \ldots, N \tag{D.16}$$

The above system of equations can be formulated in matrix notation as

$$\mathbf{C}\mathbf{W}\mathbf{y}_i = \hat{\lambda}_i \mathbf{y}_i \tag{D.17}$$

where $\mathbf{C}$ is a symmetric positive semi-definite $N \times N$ matrix with elements $c_{nj} = \mathrm{Cov}(\mathbf{t}_n, \mathbf{t}_j)$, $\mathbf{W}$ is a diagonal matrix of size N with nonnegative diagonal entries $\mathbf{W}_{jj} = w_j$, and $\mathbf{y}_i$ is a N-dimensional vector whose nth entry is $y_{i,n} = \hat{\varphi}_i(\mathbf{t}_n)$. Since the integration weights w_j are nonnegative, the matrix $\mathbf{W}$ is symmetric and positive semi-definite. The problem in Eq. (D.17) is a matrix eigenvalue problem. This matrix eigenvalue problem can be reformulated to an equivalent matrix eigenvalue problem $\mathbf{B}\mathbf{y}_i^* = \hat{\lambda}_i \mathbf{y}_i^*$, where the matrix $\mathbf{B}$ is defined as $\mathbf{B} = \mathbf{W}^{\frac{1}{2}}\mathbf{C}\mathbf{W}^{\frac{1}{2}}$, where $\mathbf{W}^{\frac{1}{2}}$ is a diagonal matrix with entries $\left(\mathbf{W}^{\frac{1}{2}}\right)_{jj} = \sqrt{w_j}$. The matrix $\mathbf{B}$ is a symmetric positive semi-definite matrix and, thus, the eigenvalues $\hat{\lambda}_i$ are nonnegative real numbers and the eigenvectors $\mathbf{y}_i^*$ are orthogonal to each other. The eigenvectors $\mathbf{y}_i$ can be obtain as $\mathbf{y}_i = \mathbf{W}^{-\frac{1}{2}}\mathbf{y}_i^*$, where $\mathbf{W}^{-\frac{1}{2}}$ denotes the inverse of the matrix $\mathbf{W}^{\frac{1}{2}}$.

Solving Eq. (D.15) for $\hat{\varphi}_i(\mathbf{t})$, we obtain the so-called Nyström interpolation formula of the eigenfunction $\hat{\varphi}_i(\mathbf{t})$. Taking into account that $\hat{\varphi}_i(\mathbf{t}_j) = \frac{1}{\sqrt{w_j}}\, y_{i,j}^*$, this results in:

$$\hat{\varphi}_i(\mathbf{t}) = \frac{1}{\hat{\lambda}_i} \sum_{j=1}^{N} \sqrt{w_j}\, y_{i,j}^*\, \mathrm{Cov}(\mathbf{t}, \mathbf{t}_j) \tag{D.18}$$

where $y_{i,j}^*$ is the jth element of the eigenvector $\mathbf{y}_i^*$.

The eigenfunctions have to be normalized such that $\int_\Omega \left(\hat{\varphi}_i(\mathbf{t})\right)^2 \mathrm{d}\mathbf{t} = 1$. Applying a numerical integration scheme, the inner product $\int_\Omega \hat{\varphi}_i(\mathbf{t})\, \hat{\varphi}_j(\mathbf{t})\, \mathrm{d}\mathbf{t}$ can be approximated as $\sum_{n=1}^{N} w_n\, \hat{\varphi}_i(\mathbf{t}_n)\, \hat{\varphi}_j(\mathbf{t}_n)$. Using the same integration points and integration weights as the ones used in Eq. (D.16), the approximation of the inner product can be simplified to $\int_\Omega \hat{\varphi}_i(\mathbf{t})\, \hat{\varphi}_j(\mathbf{t})\, \mathrm{d}\mathbf{t} \approx \left(\mathbf{y}_i^*\right)^T \mathbf{y}_j^*$. Therefore, the approximate eigenfunctions are orthonormal if and only if the eigenvectors are orthonormal.

D.7 Equivalence of the EOLE method with the Nyström method

The connection between the EOLE method and the Nyström method was first published in [Betz et al., 2014c], and is repeated in the following.

The expansion optimal linear estimation (EOLE) method is a series expansion method for discretization of random fields that was developed in [Li and Der Kiureghian, 1993] based on linear estimation theory. Here we show that the EOLE method with a uniform distribution of points over the domain can be considered a special case of the Nyström method.

Assume that points $\mathbf{t}_j, j = 1, \ldots, N$ uniformly distributed over the domain Ω are available. The points $\mathbf{t}_j$ can be chosen either at random by sampling the uniform distribution over Ω or by application of the rectangle quadrature using the nodes of an equispaced structured grid. If the domain Ω does not have a simple shape, the integration procedure can be performed on a geometrically simpler domain Ω^* that contain Ω, i.e., $\Omega \subseteq \Omega^*$. In this case, points outside

of Ω are not taken into account. If the points $\mathbf{t}_j, j = 1, \ldots, N$ are selected with one of the above procedures, then all the integration weights w_j in the integration scheme in Eq. (D.15) will be the same, i.e., $w_j = w \; \forall \, j = 1, \ldots, N$. Consequently, matrix $\mathbf{W}$ in Eq. (D.17) can be written as $\mathbf{W} = w\mathbf{I}$, where $\mathbf{I}$ is the identity matrix and $w = |\Omega|/N$. In this special case, the matrix eigenvalue problem of Eq. (D.17) can be reformulated as:

$$\mathbf{C}\mathbf{y}_i = \hat{\lambda}_i^* \mathbf{y}_i \tag{D.19}$$

where $\hat{\lambda}_i^*$ is related to $\hat{\lambda}_i$ in Eq. (D.17) as $\hat{\lambda}_i^* = \frac{N}{|\Omega|}\hat{\lambda}_i$. $\hat{\lambda}_i^*$ and $\mathbf{y}_i$ are the eigenvalues and eigenfunctions of the covariance matrix $\mathbf{C}$, respectively. Assuming normalized eigenvectors $\mathbf{y}_i$, i.e., $\|\mathbf{y}_i\| = 1$ for all i, gives after some algebra the following approximate truncated KL expansion:

$$\hat{H}(\mathbf{t}) = \mu(\mathbf{t}) + \sum_{i=1}^{M} \frac{\hat{\xi}_i}{\sqrt{\hat{\lambda}_i^*}} \sum_{j=1}^{N} y_{i,j} \, \mathrm{Cov}(\mathbf{t}, \mathbf{t}_j) \tag{D.20}$$

where $y_{i,j}$ is the jth element of $\mathbf{y}_i$.

The matrix eigenvalue problem of Eq. (D.19) is the problem that needs to be solved for the EOLE method, and the expansion in Eq. (D.20) is equivalent to the one obtained in the EOLE method. Consequently, the EOLE method is equivalent to an approximate KL expansion, whereby the integral eigenvalue problem is solved by the Nyström method with a uniform distribution of integration points.

D.8 Non-Gaussian translation random fields

General non-Gaussian random fields are not suitable to be expressed by means of Gaussian random fields. If a non-Gaussian random field belongs to the class of translation fields, it can be expressed in terms of a Gaussian random field through a nonlinear mapping of the form $H_{\text{transl.}}(\mathbf{t}) = \mathrm{g}(H(\mathbf{t}))$, where $H_{\text{transl.}}(\mathbf{t})$ represents the non-Gaussian random field defined in terms of the Gaussian field $H(\mathbf{t})$ and the strictly increasing nonlinear mapping $\mathrm{g} : \mathbb{R} \to \mathbb{R}$ [Grigoriu, 1984]. Discretization of the field $H_{\text{transl.}}(\mathbf{t})$ is achieved by replacing $H(\mathbf{t})$ by its KL expansion $\tilde{H}(\mathbf{t})$ and applying $\tilde{H}_{\text{transl.}}(\mathbf{t}) = \mathrm{g}(\tilde{H}(\mathbf{t}))$. However, it cannot be confirmed that the transformed field $\tilde{H}_{\text{transl.}}(\mathbf{t})$ inherits the optimality property that the Gaussian random field approximation $\tilde{H}(\mathbf{t})$ may possess [Li and Der Kiureghian, 1993]. All random field models used in reliability analysis and probabilistic mechanics belong essentially to the class of translation fields.

A subclass of translation random fields constitute fields where the Nataf multivariate distribution (see Section 3.2.2) is applied to perform the nonlinear mapping $\mathrm{g}(\cdot)$. For this class of translation random fields, the underlying Gaussian field has zero mean and unit variance. Its autocorrelation coefficient function $\rho(\mathbf{t}, \mathbf{t}')$ is linked to the target autocorrelation coeffi-

cient function $\rho_{\text{transl.}}(\mathbf{t}, \mathbf{t}')$ of the desired non-Gaussian field through an integral equation [Li and Der Kiureghian, 1993]. However, not for all $\rho_{\text{transl.}}(\mathbf{t}, \mathbf{t}')$ a corresponding $\rho(\mathbf{t}, \mathbf{t}')$ can be found [Grigoriu, 1998]. Moreover, it is computationally demanding to evaluate a $\rho(\mathbf{t}, \mathbf{t}')$ that is associated with a given $\rho_{\text{transl.}}(\mathbf{t}, \mathbf{t}')$, because of their implicit relationship in form of an integral equation. Therefore, it is often simpler to estimate the autocorrelation coefficient function of the underlying Gaussian random field $\rho(\mathbf{t}, \mathbf{t}')$ directly. This can be achieved by transforming available data to Gaussian data using the inverse mapping $\mathrm{g}^{-1} : \mathbb{R} \to \mathbb{R}$. It should be noted that a direct estimation of $\rho(\mathbf{t}, \mathbf{t}')$ will result in a different autocorrelation function of $H(\mathbf{t})$ than the one arising from the solution of the integral equation equation according to translation field theory. However, such a direct estimation will overcome the problem that often occurs when the solution of the integral equation does not result in an autocorrelation function that is nonnegative definite.

Appendix E

Proofs

E.1 Proofs from Section 2.2.2

In this section, the axioms (P4a) to (P7c) of Probability Theory stated in Appendix E.1 are presented.

Proof E.1. (P4a):
Starting from (P3), we get:

$$\Pr[b \wedge b|a] = \Pr[b|b \wedge a] \Pr[b|a]$$
$$\Pr[b|a] = \Pr[b|b \wedge a] \Pr[b|a]$$

which holds for general cases if and only if $\Pr[b|b \wedge a] = 1$. □

Proof E.2. (P4b):
(P4b) follows directly from combining (P4a) with (P2). □

Proof E.3. (P4c):
(P4c) follows directly from (P2) and (P1). □

Proof E.4. (P5a):
Defining $d = (c \Rightarrow b) \wedge a$, we use (P3) to state:

$$\Pr[b \wedge c|d] = \Pr[c|b \wedge d] \Pr[b|d]$$

As $(c \Rightarrow b)$, we have $\Pr[b \wedge c|d] = \Pr[c|d]$. Thus, we can write:

$$\Pr[c|d] = \Pr[c|b \wedge d] \Pr[b|d]$$

Since $\Pr[c|b \wedge d] \leq 1$ (see (P4c)):

$$\Pr[c|d] \leq \Pr[b|d]$$

□

Proof E.5. (P5b):
Using the results derived in Proof E.4, and noting that $\Pr[c|b \wedge d] = 1$ we get

$$\Pr[c|d] = \Pr[b|d]$$

□

Proof E.6. (P6):
Based on De Morgan's law[1], we can write:

$$\begin{aligned}
\Pr[b \vee c|a] &= \Pr\left[\overline{\overline{b} \wedge \overline{c}}|a\right] \\
&= 1 - \Pr\left[\overline{c}|\overline{b} \wedge a\right] \Pr\left[\overline{b}|a\right] \\
&= 1 - \left(1 - \Pr\left[c|\overline{b} \wedge a\right]\right)(1 - \Pr\left[b|a\right]) \\
&= \Pr\left[c|\overline{b} \wedge a\right] + \Pr\left[b|a\right] - \Pr\left[c|\overline{b} \wedge a\right] \Pr\left[b|a\right] \\
&= \Pr\left[c|\overline{b} \wedge a\right](1 - \Pr\left[b|a\right]) + \Pr\left[b|a\right] \\
&= \frac{\Pr\left[\overline{b} \wedge c|a\right]}{\Pr\left[\overline{b}|a\right]}(1 - \Pr\left[b|a\right]) + \Pr\left[b|a\right] \\
&= \Pr\left[\overline{b} \wedge c|a\right] + \Pr\left[b|a\right] \\
&= \Pr\left[\overline{b}|c \wedge a\right] \Pr\left[c|a\right] + \Pr\left[b|a\right] \\
&= (1 - \Pr\left[b|c \wedge a\right]) \Pr\left[c|a\right] + \Pr\left[b|a\right] \\
&= \Pr\left[b|a\right] + \Pr\left[c|a\right] - \Pr\left[b|c \wedge a\right] \Pr\left[c|a\right] \\
&= \Pr\left[b|a\right] + \Pr\left[c|a\right] - \Pr\left[b \wedge c|a\right]
\end{aligned}$$

□

Proof E.7. (P7a):
Let $b = b_1 \vee b_2 \vee \ldots \vee b_N$. Clearly, $b|a$ always occurs and $\Pr\left[b|a\right] = 1$, as a states that the propositions b_i are mutually exclusive and collectively exhaustive.

$$\begin{aligned}
\Pr[c|a] &= \Pr\left[c \wedge b|a\right] \\
&= \Pr\left[c \wedge (b_1 \vee b_2 \vee \ldots \vee b_N)\,|a\right] \\
&= \Pr\left[(c \wedge b_1) \vee \ldots \vee (c \wedge b_N)|a\right] \\
&= \sum_{i=1}^{N} \Pr\left[c \wedge b_i|a\right] \qquad \text{(taking into account that } \Pr[b_i \wedge b_j] = 0 \text{ for } i \neq j)
\end{aligned}$$

□

Proof E.8. (P7b):
Continuing Proof E.7, we can write:

$$\Pr[c|a] = \sum_{i=1}^{N} \Pr\left[c \wedge b_i|a\right]$$

[1] De Morgan's law: $\overline{b \vee c} = \overline{b} \wedge \overline{c}$

$$= \sum_{i=1}^{N} \Pr\left[c|b_i \wedge a\right] \cdot \Pr[b_i|a]$$

□

Proof E.9. (P7c):

$$\begin{aligned}\Pr[b_k|c \wedge a] &= \frac{\Pr[b_k \wedge c|a]}{\Pr[c|a]} \\ &= \frac{\Pr[c|b_k \wedge a] \cdot \Pr[b_k|a]}{\Pr[c|a]}\end{aligned}$$

□

E.2 Proofs from Section 2.3.4

Proof E.10. The distribution of $\theta_1|Z$ is Normal with mean μ_* and standard deviation σ_* according to Eqs. (2.37) and (2.38):
The joint distribution of Z and θ_1 is:

$$\begin{aligned}p_{\theta_1,Z}(\theta_1, Z) =& \frac{\exp\left[-\frac{\frac{(\theta_1-\mu_1)^2}{\sigma_1^2} + \frac{(Z-\theta_1-\mu_2)^2}{\sigma_2^2} - \frac{2\rho(\theta_1-\mu_1)(Z-\theta_1-\mu_2)}{\sigma_1\sigma_2}}{2(1-\rho^2)}\right]}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \\ =& \frac{\sqrt{\sigma_1^2+\sigma_2^2+2\rho\sigma_1\sigma_2}}{\sigma_1\sigma_2\sqrt{1-\rho^2}\sqrt{2\pi}} \\ & \cdot \exp\left[-\frac{\frac{(\theta_1-\mu_1)^2}{\sigma_1^2} + \frac{(Z-\theta_1-\mu_2)^2}{\sigma_2^2} - \frac{(1-\rho^2)(Z-\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2+2\rho\sigma_1\sigma_2} - \frac{2\rho(\theta_1-\mu_1)(Z-\theta_1-\mu_2)}{\sigma_1\sigma_2}}{2(1-\rho^2)}\right] \\ & \cdot p_Z(Z) \\ =& \frac{1}{\sigma_*\sqrt{2\pi}} \exp\left[-\frac{1}{2}\frac{(\theta_1-\mu_*)^2}{\sigma_*^2}\right] \cdot p_Z(Z) \\ =& p_{\theta_1,Z}(\theta_1, Z) \cdot p_Z(Z)\end{aligned}$$

Thus, $p_{\theta_1,Z}(\theta_1, Z)$ is:

$$p_{\theta_1,Z}(\theta_1, Z) = \frac{1}{\sigma_*\sqrt{2\pi}} \exp\left[-\frac{1}{2}\frac{(\theta_1-\mu_*)^2}{\sigma_*^2}\right]$$

with

$$\begin{aligned}\mu_* &= \frac{\mu_1\sigma_2^2 + (Z-\mu_2)\sigma_1^2 + \rho\sigma_1\sigma_2(Z-\mu_2+\mu_1)}{\sigma_1^2+\sigma_2^2+2\rho\sigma_1\sigma_2} \\ \sigma_* &= \frac{\sigma_1\sigma_2\sqrt{1-\rho^2}}{\sqrt{\sigma_1^2+\sigma_2^2+2\rho\sigma_1\sigma_2}}\end{aligned}$$

□

Proof E.11. Derivation of $p(\omega)$ given in Eq. (2.45) based on the beta-distributed $\cos^2\omega$:

$$\begin{aligned} p(\omega) &= \frac{1}{2}\left|\sin(\omega)\cos(\omega)\right| \cdot \frac{\left(\cos^2\omega\right)^{-\frac{1}{2}}\left(1-\cos^2\omega\right)^{\frac{M-3}{2}}}{B\left(\frac{1}{2},\frac{M-1}{2}\right)} \\ &= \frac{\left|\sin(\omega)\cos(\omega)\right|}{2\left|\cos\omega\right|} \cdot \frac{\left|\sin\omega\right|^{M-3}}{B\left(\frac{1}{2},\frac{M-1}{2}\right)} \\ &= \frac{\left|\sin\omega\right|^{M-2}}{2 \cdot B\left(\frac{1}{2},\frac{M-1}{2}\right)} \end{aligned}$$

□

E.3 Proofs from Section 2.3.5

Proof E.12. The mean, standard deviation and auto-correlation coefficient function of an AR(1) model is given by Eq. (2.56), Eq. (2.57) and Eq. (2.58), respectively.

$$\mathrm{E}[X_t] = c + a \cdot \mathrm{E}[X_{t-1}] + \mathrm{E}[b_t]\,,$$

where b_t is white noise and, consequently, $\mathrm{E}[b_t] = 0$. Assuming stationarity of the mean, we have $\mathrm{E}[X_t] = \mathrm{E}[X_{t-1}] = \mu_X$.

$$\begin{aligned} \mu_X &= c + a \cdot \mu_X \\ \mu_X &= \frac{c}{1-a} \end{aligned}$$

The variance of X_t is defined as:

$$\begin{aligned} \mathrm{E}\left[(X_t - \mu_X)^2\right] &= \mathrm{E}\left[\left(c + a \cdot X_{t-1} + b_t - c - a\mu_X\right)^2\right] \\ &= a^2\,\mathrm{E}\left[\left(X_{t-1} - \mu_X\right)^2\right] + \mathrm{E}\left[{b_t}^2\right] + 2a\,\mathrm{E}\left[X_{t-1} \cdot b_t\right] \\ &= a^2\,\mathrm{E}\left[\left(X_{t-1} - \mu_X\right)^2\right] + {\sigma_b}^2 \end{aligned}$$

For a weak-sense stationary stochastic process we have:

$$\begin{aligned} {\sigma_X}^2 &= a^2{\sigma_X}^2 + {\sigma_b}^2 \\ \sigma_X &= \frac{\sigma_b}{1-a^2} \end{aligned}$$

Let $U_t = X_t - \mu_X = a \cdot U_{t-1} + b_t$ be a stationary process that has zero mean. The covariance between X_{t_1} and X_{t_2} can be expressed as (assuming without loss of generality that $t_2 > t_1$):

$$\mathrm{Cov}\left[X_{t_1}, X_{t_2}\right] = E\left[U_{t_1} \cdot U_{t_2}\right] = a^{t_2 - t_1} \cdot {\sigma_X}^2$$

Thus, the auto-correlation coefficient function is:

$$\rho_X(\tau) = a^{|\tau|}$$

with $\tau = t_2 - t_1$. □

E.4 Proofs from Section 3.4.3

This section contains proofs related to the transition PDF of the Metropolis-Hastings algorithm.

Proof E.13. The transition PDF defined in Eq. (3.16) integrates to one:

$$\begin{aligned}\int p(\mathbf{v}|\mathbf{w})\,\mathrm{d}\mathbf{v} &= \int r^*_{\mathbf{w}}(\mathbf{v}) \cdot q(\mathbf{v}|\mathbf{w})\,\mathrm{d}\mathbf{v} + \int \delta_{\mathbf{w}}(\mathbf{v})\,\mathrm{d}\mathbf{v} \cdot \int (1 - r^*_{\mathbf{w}}(\boldsymbol{\xi}))\, q(\boldsymbol{\xi}|\mathbf{w})\,\mathrm{d}\boldsymbol{\xi} \\ &= \int r^*_{\mathbf{w}}(\mathbf{v}) \cdot q(\mathbf{v}|\mathbf{w})\,\mathrm{d}\mathbf{v} + \int (1 - r^*_{\mathbf{w}}(\boldsymbol{\xi}))\, q(\boldsymbol{\xi}|\mathbf{w})\,\mathrm{d}\boldsymbol{\xi} \\ &= \int q(\mathbf{v}|\mathbf{w})\,\mathrm{d}\mathbf{v} = 1\end{aligned}$$

□

Proof E.14. The Metropolis-Hastings algorithm fulfills detailed balance:

Let $\mathbf{w}$ be the current state of the chain (i.e., the seed), and $\mathbf{v}$ is the next state of the chain. We need to show that the algorithm maintains detailed balance Eq. (3.13). Let us distinguish the following two cases:

Case 1: $\mathbf{v} = \mathbf{w}$
If the next state of the chain equals the current state, it is obvious that the reversibility condition holds.

Case 2: $\mathbf{v} \neq \mathbf{w}$
We can express the transition PDF in this case as:

$$p(\mathbf{v}|\mathbf{w}) = q(\mathbf{v}|\mathbf{w}) \cdot \min\left(1, \frac{p_s(\mathbf{v})}{p_s(\mathbf{w})} \cdot \frac{q(\mathbf{w}|\mathbf{v})}{q(\mathbf{v}|\mathbf{w})}\right) \tag{E.1}$$

Correspondingly:

$$p(\mathbf{w}|\mathbf{v}) = q(\mathbf{w}|\mathbf{v}) \cdot \min\left(1, \frac{p_s(\mathbf{w})}{p_s(\mathbf{v})} \cdot \frac{q(\mathbf{v}|\mathbf{w})}{q(\mathbf{w}|\mathbf{v})}\right) \tag{E.2}$$

Using the relation $\min\left(1, \frac{a}{b}\right) = \min\left(1, \frac{b}{a}\right) \frac{a}{b}$ for all positive real numbers a and b, we can rewrite Eq. (E.2) as:

$$p(\mathbf{w}|\mathbf{v}) = q(\mathbf{v}|\mathbf{w}) \cdot \min\left(1, \frac{p_s(\mathbf{v})}{p_s(\mathbf{w})} \cdot \frac{q(\mathbf{w}|\mathbf{v})}{q(\mathbf{v}|\mathbf{w})}\right) \cdot \frac{p_s(\mathbf{w})}{p_s(\mathbf{v})} \tag{E.3}$$

Which is equivalent to:

$$\min\left(1, \frac{p_s(\mathbf{v})}{p_s(\mathbf{w})} \cdot \frac{q(\mathbf{w}|\mathbf{v})}{q(\mathbf{v}|\mathbf{w})}\right) = \frac{p(\mathbf{w}|\mathbf{v})}{q(\mathbf{v}|\mathbf{w})} \cdot \frac{p_s(\mathbf{v})}{p_s(\mathbf{w})} \tag{E.4}$$

Putting Eq. (E.4) into Eq. (E.1) we get:

$$p(\mathbf{v}|\mathbf{w}) \cdot p_s(\mathbf{w}) = p(\mathbf{w}|\mathbf{v}) \cdot p_s(\mathbf{v}) \tag{E.5}$$

□

Appendix F

Statistical data analysis – Descriptive statistics

F.1 Numerical descriptors of data

Numerical descriptors are used to quantify an entire data set by a few representative numbers.

F.1.1 Univariate analysis - independent samples

Let $x_i \in \mathbb{R}$, $i = 1, \ldots, N$ denote a set of samples. Furthermore, it is assumed that the x_i are independent realizations of a random variable X whose distribution is usually unknown.

F.1.1.1 Sample mean

The average over all samples x_i is called the *sample mean* $\overline{x}$ and is defined as:

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{F.1}$$

The sample mean is an unbiased estimator and, consequently, $\mathrm{E}[\overline{x}] = \mathrm{E}[X]$. The variance of the estimator is given as $\mathrm{Var}[\overline{x}] = \frac{1}{N} \mathrm{Var}[X]$. For large N, the distribution of $\overline{x}$ can often be regarded as approximately normal, due to the central limit theorem.

Proof F.1. Eq. (F.1) is an unbiased estimator for $\mathrm{E}[X]$:

$$\mathrm{E}[\overline{x}] = \frac{1}{N} \mathrm{E}\left[\sum_{i=1}^{N} x_i\right]$$

$$= \frac{1}{N} \sum_{i=1}^{N} \mathrm{E}[X]$$
$$= \mathrm{E}[X]$$

Note that the samples do not have to be independent for the estimator to be unbiased.

□

F.1.1.2 Sample variance

The *sample variance* s^2 is an estimator for the variance of X. The unbiased estimator (i.e., $\mathrm{E}[s^2] = \mathrm{Var}[X]$) is defined as:

$$s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})^2 \tag{F.2}$$
$$= \frac{1}{N-1} \left(\left(\sum_{i=1}^{N} x_i^2 \right) - N \cdot \overline{x}^2 \right) \tag{F.3}$$

The advantage of Eq. (F.3) over Eq. (F.2) is that for Eq. (F.3) the sums $\sum_{i=1}^{N} x_i$ and $\sum_{i=1}^{N} x_i^2$ can be evaluated simultaneously within a single loop, whereas in Eq. (F.2) the sample mean $\overline{x}$ must be evaluated before the sum $\sum_{i=1}^{N} (x_i - \overline{x})^2$. However, the sum $\sum_{i=1}^{N} x_i^2$ is critical in terms of floating-point arithmetic, as is the difference $\left(\sum_{i=1}^{N} x_i^2\right) - N \cdot \overline{x}^2$, especially if $|\mathrm{E}[X]|$ is considerably larger than zero and $\mathrm{Var}[X]$ is also large.

Proof F.2. Eqs. (F.2) and (F.3) give an unbiased estimator for $\mathrm{Var}[X]$:

$$\begin{aligned}
\mathrm{E}[s^2] &= \frac{1}{N-1} \mathrm{E}\left[\sum_{i=1}^{N} x_i^2 - N \cdot \overline{x}^2 \right] \\
&= \frac{1}{N-1} \left(\sum_{i=1}^{N} \mathrm{E}\left[x_i^2\right] - N \cdot \mathrm{E}\left[\overline{x}^2\right] \right) \\
&= \frac{1}{N-1} \left[N \cdot \left(\mathrm{E}\left[X\right]^2 + \mathrm{Var}[X] \right) - \left(N \cdot \mathrm{E}[X]^2 + \mathrm{Var}[X] \right) \right] \\
&= \mathrm{Var}[X]
\end{aligned}$$

□

Proof F.3. In Proof F.2 we used $\mathrm{E}\left[\overline{x}^2\right] = \mathrm{E}[X]^2 + \frac{1}{N}\mathrm{Var}[X]$, this relation can be derived as:

$$\begin{aligned}
\mathrm{E}\left[\overline{x}^2\right] &= \frac{1}{N^2} \mathrm{E}\left[\sum_{k=1}^{N} \sum_{i=1}^{N} x_i x_k \right] \\
&= \frac{1}{N^2} \sum_{k=1}^{N} \sum_{i=1}^{N} \mathrm{E}\left[x_i x_k\right]
\end{aligned}$$

$$= \frac{1}{N}\left(\mathrm{E}\left[X\right]^2 + \mathrm{Var}\left[X\right]\right) + \frac{N-1}{N}\left(\mathrm{E}\left[X\right]^2\right)$$
$$= \mathrm{E}[X]^2 + \frac{1}{N}\,\mathrm{Var}[X]$$

Note that this relation only holds for independent samples. □

F.1.1.3 Distribution of the mean (for Normal population)

N statistically independent outcomes $X_1, \ldots, X_N$ of random quantity X have been observed. It is assumed that the mean μ_X and standard deviation σ_X of X is not known. Furthermore, it is implicitly assumed that the distribution of X is Normal. However, for large N, the derived results will hold for general types of distribution of X; this is a consequence of the *central limit theorem*, which states that for large N the distribution of $\overline{x}$ approaches a Normal distribution.

The sample mean $\overline{x}$ and standard deviation s of X can be estimated by means of Eqs. (F.1) and (F.3), respectively. The statistical uncertainty about the unknown *true* mean μ_X can be expressed as:

$$\mu_X = \overline{x} + \frac{s}{\sqrt{N}} \cdot T_{N-1} \tag{F.4}$$

where T_{N-1} is a random variable that has the Student's t-distribution with $N-1$ degrees of freedom. Thus, the probability that $\mu_x \leq b$ is:

$$\Pr\left(\mu_x \leq b\right) = \Pr\left(\overline{x} + \frac{s}{\sqrt{N}} \cdot T_{N-1} \leq b\right) \tag{F.5}$$
$$= \Pr\left[T_{N-1} \leq (b - \overline{x}) \cdot \frac{\sqrt{N}}{s}\right] \tag{F.6}$$

where Eq. (F.6) is essentially the CDF of a $N-1$ degree of freedom Student's t-distribution evaluated at $(b-\overline{x}) \cdot \frac{\sqrt{N}}{s}$.

F.1.2 Univariate analysis - chain-dependent samples

Let $x_{j,i} \in \mathbb{R}$, $j = 1, \ldots, M$, $i = 1, \ldots, N$ denote M sets of samples, where each set contains N samples. It is assumed that all samples $x_{j,i}$ are realizations of random variable X. Samples of different sets are not dependend. However, samples within any set $j = 1, \ldots, M$, $x_{j,1}, \ldots, x_{j,N}$, are dependent, where the specific dependence structure is unknown. An example for such a setting is given by M Markov chains of length N whose seeds are independent and distributed according to the target distribution.

F.1.2.1 Sample mean

The sample mean of the jth set can be computed as:

$$\overline{x}_j = \frac{1}{N} \sum_{i=1}^{N} x_{j,i} \tag{F.7}$$

The sample mean of all samples is:

$$\overline{x} = \frac{1}{M} \sum_{j=1}^{M} \overline{x}_j \tag{F.8}$$

Both $\overline{x}_j$ and $\overline{x}$ are unbiased estimators for the mean of X (compare Proof F.1).

The estimates $\overline{x}_1, \ldots, \overline{x}_j, \ldots, \overline{x}_M$ are independent, since samples of different sets are independent. Consequently, the variance of the estimates $\overline{x}_j$, denoted $\mathrm{Var}\left[\overline{X}_j\right]$, can be estimated by means of Eq. (F.3):

$$r = \frac{1}{M-1} \left(\left(\sum_{j=1}^{M} (\overline{x}_j)^2 \right) - M \cdot (\overline{x})^2 \right) \tag{F.9}$$

where r is an unbiased estimator for $\mathrm{Var}\left[\overline{X}_j\right]$, i.e., $\mathrm{E}[r] = \mathrm{Var}\left[\overline{X}_j\right]$. (compare Proof F.2). Here we implicitly assume that the dependence structure of samples $x_{j,1} \ldots, x_{j,N}$ is the same independent of j, and, thus, $x_{j,1} \ldots, x_{j,N}$ are independent and identical distributed realizations of a random variable denoted $\overline{X}_j$. Note that contrary to the case of independent samples described in Section F.1.1, we have $\mathrm{Var}\left[\overline{X}_j\right] \neq \frac{1}{N} \mathrm{Var}[X]$ in general. However, for the variance of the estimate $\overline{x}$ we can write $\mathrm{Var}\left[\overline{X}\right] = \frac{1}{M} \mathrm{Var}\left[\overline{X}_j\right]$.

F.1.2.2 Sample variance (using samples from a single set)

The sample variance s_j^2 of the samples in the jth set can be estimated as:

$$s_j^2 = \frac{1}{N-1} \left(\left(\sum_{i=1}^{N} (x_{j,i})^2 \right) - N \cdot (\overline{x}_j)^2 \right) \tag{F.10}$$

with $\overline{x}_j$ according to Eq. (F.7). However, the estimate s_j^2 is biased (i.e., $\mathrm{E}\left[s_j^2\right] \neq \mathrm{Var}[X]$), because the samples $x_{j,1}, \ldots, x_{j,N}$ are dependent:

Proof F.4. Eq. (F.10) gives a biased estimate:
First we investigate $\mathrm{E}\left[(\overline{x}_j)^2\right]$ (compare Proof F.3 for independent samples):

$$\mathrm{E}\left[(\overline{x}_j)^2\right] = \frac{1}{N^2} \sum_{k=1}^{N} \sum_{i=1}^{N} \mathrm{E}\left[x_{j,i} \cdot x_{j,k}\right]$$

$$= \frac{1}{N}\left(\mathrm{E}\left[X\right]^2 + \mathrm{Var}\left[X\right]\right) + \frac{N-1}{N}\left(\mathrm{E}\left[X\right]^2\right) + \frac{1}{N^2}\sum_{k=1}^{N}\sum_{\forall i \neq k}\mathrm{Cov}\left[x_{j,i}, x_{j,k}\right]$$

$$= \mathrm{E}[X]^2 + \frac{1}{N}\mathrm{Var}[X] + \frac{1}{N^2}\sum_{k=1}^{N}\sum_{\forall i \neq k}\mathrm{Cov}\left[x_{j,i}, x_{j,k}\right]$$

Using this result, we can express $\mathrm{E}[s_j^2]$ as follows:

$$\begin{aligned}
\mathrm{E}[s_j^2] &= \frac{1}{N-1}\left(\sum_{i=1}^{N}\mathrm{E}\left[(x_{j,i})^2\right] - N \cdot \mathrm{E}\left[(\overline{x}_j)^2\right]\right) \\
&= \frac{1}{N-1}\left[N \cdot \left(\mathrm{E}\left[X\right]^2 + \mathrm{Var}[X]\right) - \left(N \cdot \mathrm{E}[X]^2 + \mathrm{Var}[X] + \frac{1}{N}\sum_{k=1}^{N}\sum_{\forall i \neq k}\mathrm{Cov}\left[x_{j,i}, x_{j,k}\right]\right)\right] \\
&= \mathrm{Var}[X] - \frac{1}{N(N-1)}\sum_{k=1}^{N}\sum_{\forall i \neq k}\mathrm{Cov}\left[x_{j,i}, x_{j,k}\right]
\end{aligned}$$

Consequently, $\mathrm{E}[s_j^2] \neq \mathrm{Var}[X]$. □

From Proof F.4 it also follows that s_j^2 underestimates the true $\mathrm{Var}[X]$ on average for samples $x_{j,1} \ldots, x_{j,N}$ that exhibit a positive dependence structure; because in this case $\sum_{k=1}^{N}\sum_{\forall i \neq k}\mathrm{Cov}\left[x_{j,i}, x_{j,k}\right] > 0$.

The estimate s_j^2 defined in Eq. (F.10) is asymptotically unbiased as $N \to \infty$ if two samples $x_{j,i}$ and $x_{j,k}$ can be considered independent in case i and k are far apart.

F.1.2.3 Sample variance (using all samples)

An unbiased estimator for the variance of X is [Gelman and Rubin, 1992; Gelman et al., 2004a]:

$$s^2 = \frac{N-1}{N} \cdot q + r \tag{F.11}$$

where r is the sample variance of the sample means defined in Eq. (F.9), and q is the average of the M sample variances s_j^2. The quantity q is defined as:

$$q = \frac{1}{M}\sum_{j=1}^{M} s_j^2 \tag{F.12}$$

with s_j^2 according to Eq. (F.10).

Proof F.5. Eq. (F.11) is an unbiased estimate for the variance of X:

$$\begin{aligned}
\mathrm{E}[s^2] &= \frac{N-1}{N} \cdot \mathrm{E}[q] + \mathrm{E}[r] \\
&= \frac{1}{NM}\sum_{j=1}^{M}\left(\sum_{i=1}^{N}\mathrm{E}\left[(x_{j,i})^2\right] - N \cdot \mathrm{E}\left[(\overline{x}_j)^2\right]\right) + \mathrm{Var}\left[\overline{X}_j\right]
\end{aligned}$$

$$
\begin{aligned}
&= \mathrm{E}\left[X^2\right] - \mathrm{E}\left[(\overline{X}_j)^2\right] + \mathrm{Var}\left[\overline{X}_j\right] \\
&= \mathrm{Var}\left[X\right] + \mathrm{E}\left[X\right]^2 - \mathrm{E}\left[(\overline{X}_j)\right]^2 \\
&= \mathrm{Var}\left[X\right]
\end{aligned}
$$

In the proof we use the relations:

- $\mathrm{E}\left[r\right] = \mathrm{Var}\left[\overline{X}_j\right]$,
- $\mathrm{E}\left[X^2\right] = \mathrm{Var}\left[X\right] + \mathrm{E}\left[X\right]^2$,
- $\mathrm{E}\left[(\overline{X}_j)^2\right] - \mathrm{Var}\left[\overline{X}_j\right] = \mathrm{E}\left[(\overline{X}_j)\right]^2$
- and $\mathrm{E}\left[(\overline{X}_j)\right] = \mathrm{E}\left[X\right]$ (see Proof F.1).

Note that the proof holds only if samples of different sets are actually independent. □

F.1.2.4 Effective number of samples

In the above discussion of the estimators for the sample mean, we have noted that in general $\mathrm{Var}\left[\overline{X}_j\right] \neq \frac{1}{N}\mathrm{Var}[X]$. Whereas for independent samples $x_{j,1}\ldots,x_{j,N}$ we have $\mathrm{Var}\left[\overline{X}_j\right] = \frac{1}{N}\mathrm{Var}[X]$. If the samples $x_{j,1}\ldots,x_{j,N}$ exhibit a positive dependence structure, we can write $\mathrm{Var}\left[\overline{X}_j\right] > \frac{1}{N}\mathrm{Var}[X]$. Let us now introduce a quantity $n_{\mathrm{eff},X,j}$ for which $\mathrm{Var}\left[\overline{X}_j\right] = \frac{1}{n_{\mathrm{eff},X,j}}\mathrm{Var}[X]$. The quantity $n_{\mathrm{eff},X,j}$ is referred to as effective number of independent samples in the set $x_{j,1}\ldots,x_{j,N}$. The value of $n_{\mathrm{eff},X,j}$ can be estimated as [Gelman et al., 2004a]:

$$n_{\mathrm{eff},X,j} = \frac{\mathrm{Var}[X]}{\mathrm{Var}\left[\overline{X}_j\right]} \approx \frac{s^2}{r} \tag{F.13}$$

with s^2 and r according to Eq. (F.11) and Eq. (F.9), respectively. The effective number of independent samples in the set consisting of all $M \cdot N$ samples $x_{j,i}$ is:

$$n_{\mathrm{eff},X} = M \cdot n_{\mathrm{eff},X,j} \tag{F.14}$$

The efficiency of the sampling procedure can be expressed as:

$$\mathrm{eff}_X = \frac{n_{\mathrm{eff},X}}{N \cdot M} = \frac{n_{\mathrm{eff},X,j}}{N} \tag{F.15}$$

The meaning of this quantity can be loosely interpreted as follows: $K \in [1, N]$ samples in the set consisting of N dependend samples contribute as much to the analysis as $\mathrm{eff}_X \cdot K$ truly independent samples would contribute.

Example F.1. *Efficiency of correlated samples that follow the Normal distribution*:
We investigate the influence of correlation on the efficiency of the sampling procedure for an

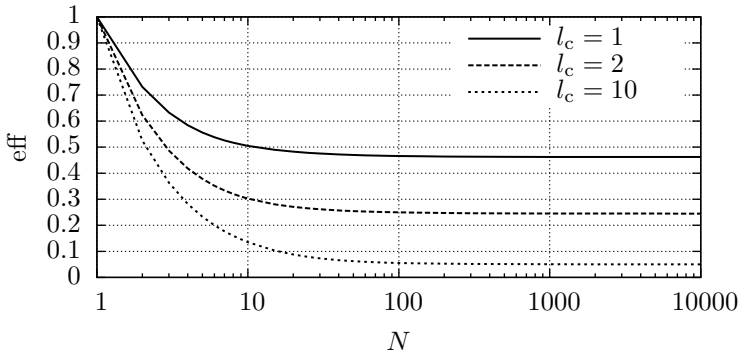


Figure F.1: Efficiency of the sampling procedure for an increasing number of samples N and different correlation length l_c. (Example F.1)

increasing N (the size of the sets of correlated samples). The sample used follow the standard Normal distribution (i.e., $\mathrm{Var}[X] = 1$). The correlation of the ith sample with the kth sample in a set is:

$$\rho(i,k) = \exp\left(-\frac{|k-i|}{l_c}\right)$$

where l_c is a factor that controls how fast the correlation decreases with an increasing distance $|k-i|$. The variance of the sample average $\mathrm{Var}\left[\overline{X}_j\right]$ can be derived analytically as:

$$\mathrm{Var}\left[\overline{X}_j\right] = \frac{\mathrm{Var}\left[X\right]}{N}\left(1 + \frac{2}{N}\cdot\sum_{i=1}^{N}\sum_{k=i+1}^{N}\rho(i,k)\right)$$

With the analytical expressions for $\mathrm{Var}\left[X\right]$ and $\mathrm{Var}\left[\overline{X}_j\right]$ available, the efficiency of the sampling procedure can be expressed as $\mathrm{eff} = \mathrm{Var}\left[X\right] / \left(N \cdot \mathrm{Var}\left[\overline{X}_j\right]\right)$. Consequently, eff does not depend on the number of sets M.

The relation between eff and N is plotted in Fig. F.1 for different l_c. An increase of l_c, i.e. an increase of the correlatetness of the samples, decreases the efficient number of independent samples. Moreover, as the number N of correlated samples is increased, the fraction of the equivalent number of independent samples also decreases and (for large N) converges to a threshold that depends on l_c.

F.1.2.5 Special case: chain-dependent Bernoulli trials

Sample variance

Let the $x_{j,i} \in \{0,1\}$ be Bernoulli trials, with $\mathrm{E}\left[x_{j,i}\right] = p$. Moreover, let $\tilde{p} = \overline{x}$ denote the sample mean defined in Eq. (F.8). The variance of $\tilde{p}$ can be computed as [Au and Beck, 2001]:

$$\mathrm{Var}\left[\tilde{p}\right] = \frac{p\cdot(1-p)}{M\cdot N}\cdot(1+\gamma) \tag{F.16}$$

where γ is defined as:

$$\gamma = 2 \cdot \sum_{k=1}^{N-1} \left(1 - \frac{k}{N}\right) \cdot \rho(k) \tag{F.17}$$

with $\rho(k) = R(k)/R(0)$ and $R(k)$ as

$$R(k) = \mathrm{E}\left[x_{j,1} \cdot x_{j,1+k}\right] - p^2 \quad \text{with } k \in \{0, \ldots, N-1\} \tag{F.18}$$

The expectation in Eq. (F.18) can be estimated as:

$$\mathrm{E}\left[x_{j,1} \cdot x_{j,1+k}\right] \approx \frac{1}{M \cdot (N-k)} \sum_{j=1}^{M} \sum_{i=1}^{N-k} x_{j,i} \cdot x_{j,i+k} \tag{F.19}$$

The value of p can be approximated using the sample mean $\tilde{p}$.

Efficiency

The factor $(1 + \gamma)$ determines by how much the variance of the sample mean of $M \cdot N$ independent Bernoulli trials needs to be increased to get $\mathrm{Var}\left[\tilde{p}\right]$. The quantity $n_{\mathrm{eff},\gamma,X} = M \cdot N/(1 + \gamma)$ can be interpreted as an effective number of independent samples [Au and Beck, 2001]. A measure for the efficiency of the sampling procedure is:

$$\mathrm{eff}_{\gamma,X} = \frac{1}{1+\gamma} \tag{F.20}$$

where $\mathrm{eff}_{\gamma,X}$ equals *one* if the Bernoulli trials are independent.

Glossary

binomial coefficient The binomial coefficient is a positive integer that can be computed as:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \qquad \text{for } 0 \leq k \leq n \tag{F.21}$$

where k and n are both nonnegative integers, and $n!$ denotes the factorial of n. For $k > n$, the binomial coefficient is *zero*. The following properties hold for the binomial coefficient:

$$\binom{n}{k} = \binom{n}{n-k} \qquad \text{for } 0 \leq k \leq n \tag{F.22}$$

$$\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1} \qquad \text{for } 1 \leq k \leq n \tag{F.23}$$

The binomial coefficient is an important quantity in combinatorics. For example, the following questions can be answered by means of the binomial coefficient:

- How many ways exist to choose k elements from a set of n elements (if the already picked elements are not replaced)?

  $$\binom{n}{k} \tag{F.24}$$

  This is the reason why the binomial coefficient is often read as: "n choose k".

- How many ways exist to choose k elements from a set of n elements if the picked elements are replaced?

  $$\binom{n+k-1}{k} \tag{F.25}$$

- How many different ways are there to represent a sequence of length n that consists of k *ones* and n *zeros*?

  $$\binom{n+k}{k} \tag{F.26}$$

- Same question as the previous one; with the additional restriction that at least

one *zero* must separate two *ones*?

$$\binom{n+1}{k} \tag{F.27}$$

bounded set Loosely speaking, a subset of a metric space is bounded if it is of finite size.

For example, the real line $\mathbb{R}$ equipped with the Euclidean topology:

- $\mathbb{R}$ is *unbounded*.
- $(0, 1)$ as well as $[0, 1]$ and $(0, 1]$ is *bounded*.
- $(-\infty, 0]$ is *unbounded*.

closed and open sets Loosely speaking, a set is called *open* if it does not contain any of its boundary points. A *closed set* is a set whose complement is an *open set*. In a topological space, a set is closed if and only if it coincides with its closure.

For example, the real line $\mathbb{R}$ equipped with the Euclidean topology:

- $\mathbb{R}$ is *open* and *closed*: The entire space is by definition *open*. Its complement is the empty set, which is by definition also *open*.
- $(0, 1)$ is *open* but not closed.
- $(-\infty, 0] \cup [1, \infty)$ is *closed* but not open.
- $[0, 1]$ is *closed* but not open.
- $(0, 1]$ is neither open nor closed.

closure Let S be a subset in a topological space. The *closure* of S consists of all points in S plus the limit points.

Cholesky decomposition Let $\mathbf{R}$ be a real symmetric matrix that is positive definite. The Cholesky decomposition of $\mathbf{R}$ has the following form:

$$\mathbf{R} = \mathbf{A}\mathbf{A}^{\mathrm{T}} \tag{F.28}$$

where $\mathbf{A}$ is a lower triangular matrix that has positive diagonal entries.

An example application is the generation of correlated standard Normal random variables: Let $\mathbf{u}$ be a vector of independent standard Normal random variables. A vector $\mathbf{y}$ that has correlation matrix $\mathbf{R}$ can be obtained by the transformation:

$$\mathbf{y} = \mathbf{A}\mathbf{u} \tag{F.29}$$

where $\mathbf{A}$ is the Cholesky decomposition of $\mathbf{R}$.

Euler-Mascheroni constant The *Euler-Mascheroni constant* is typically denoted by γ. The value of this mathematical constant is $0.57721566490153286060\ldots$.

finite and infinite sets A set is called *finite* if the number of elements in the set is a *natural number*. The number of elements in the set is referred to as the *cardinality* of the set.

A set is said to be *infinite* if it is not *finite*.

Matrix eigenvalue problem

eigenvector: An eigenvector $\mathbf{v}$ of matrix a $N \times N$ $\mathbf{A}$ must satisfy

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v} \tag{F.30}$$

where $\mathbf{v}$ is a *non-zero* vector and the scalar λ is referred to as eigenvalue corresponding to eigenvector $\mathbf{v}$ of matrix $\mathbf{A}$.

characteristic polynomial: Eigenvalues of matrix $\mathbf{A}$ are the roots of the following equation:

$$p(\lambda) := \det(\mathbf{A} - \lambda\mathbf{I}) = 0 \tag{F.31}$$

where $p(\lambda)$ is the characteristic polynomial of $\mathbf{A}$ which can be factorized as:

$$p(\lambda) = (\lambda - \lambda_1)^{n_1} \cdot (\lambda - \lambda_2)^{n_2} \cdot \ldots \cdot (\lambda - \lambda_K)^{n_K} = 0 \tag{F.32}$$

where $\lambda_1, \ldots, \lambda_K$ are the K eigenvalues of matrix $\mathbf{A}$. The integers n_k with $k = 1, \ldots, K$ are called the *algebraic multiplicity* of eigenvalue λ_k. We have:

$$\sum_{k=1}^{K} n_k = N \tag{F.33}$$

Each eigenvalue λ_k has an associated eigenvalue equation:

$$(\mathbf{A} - \lambda_k\mathbf{I})\,\mathbf{v} = 0 \tag{F.34}$$

which has $m_k \in 1, \ldots, n_k$ linearly independent solutions; where the integer m_k is called the *geometric multiplicity* of eigenvalue λ_k.

eigendecomposition: For a $N \times N$ matrix $\mathbf{A}$ with N linearly independent eigenvectors $\mathbf{v}_i$, matrix $\mathbf{A}$ can be expressed as:

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1} \tag{F.35}$$

where $\mathbf{V}$ is a $N \times N$ matrix whose ith column is eigenvector $\mathbf{v}_i$. $\mathbf{\Lambda}$ is a diagonal matrix whose diagonal elements $\mathbf{\Lambda}_{ii}$ are eigenvalues λ_i.

eigendecomposition of the inverse: If matrix $\mathbf{A}$ has an eigendecomposition and if all eigenvalues are *non-zero*, the inverse of $\mathbf{A}$ is:

$$\mathbf{A}^{-1} = \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^{-1} \tag{F.36}$$

Consequently, the eigenvalues of $\mathbf{A}^{-1}$ are $\frac{1}{\lambda_i}$. Note that all eigenvectors are *real*.

real symmetric matrices: For a $N \times N$ real symmetric matrix $\mathbf{A}$, the eigenvectors are *orthogonal*. Assuming that they are additionally scaled such that they become *orthonormal*, we can write:

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\intercal \tag{F.37}$$

Thus, $\mathbf{V}$ is an orthogonal matrix.

generalized matrix eigenvalue problem:

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{B}\mathbf{v} \tag{F.38}$$

$\mathbf{v}$, λ is called the generalized eigenvector, eigenvalue of $\mathbf{A}$ and $\mathbf{B}$, respectively. The eigenvalue λ satisfy

$$\det(\mathbf{A} - \lambda\mathbf{B}) = 0 \tag{F.39}$$

If $\mathbf{A}$ and $\mathbf{B}$ are $N \times N$ matrices and the problem has N linearly independent eigenvectors, matrix $\mathbf{A}$ can be written as:

$$\mathbf{A} = \mathbf{B}\mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1} \tag{F.40}$$

Furthermore, if both $\mathbf{A}$ and $\mathbf{B}$ are symmetric and if $\mathbf{B}$ is additionally positive-definite, then the eigenvalues are all real, and the eigenvectors obey the following equation:

$$\mathbf{V}^\intercal\mathbf{B}\mathbf{V} = \mathbf{F} \tag{F.41}$$

where $\mathbf{F}$ is a diagonal matrix with positive entries.

support of a function Let $f : X \to \mathbb{R}$ be a real-valued function, where we restrict ourselves to X being a topological space. Furthermore, let S be a subset of X such that f is non-zero for points in S and zero for points not in S. The *support* of function f is the closure of the subset S. On topological spaces this is also referred to as *closed support*. Function f is said to have a *compact support* if the support is a compact subset of X.

totally ordered If a set X is *totally ordered*, then the following statements hold for any a, b and c in X:

1. *antisymmetry:* if $a \leq b$ and $b \leq a$ then $a = b$;
2. *transitivity:* if $a \leq b$ and $b \leq c$ then $a \leq c$;
3. *totality:* $a \leq b$ or $b \leq a$.

Bibliography

Adminaite, D., Allsop, R., and Jost, G.: 9th Road Safety Performance Index Report. Technical report, European Transport Safety Council, 2015. `http://etsc.eu/wp-content/uploads/ETSC-9th-PIN-Report_Final.pdf`. (cited on page 13)

Ale, B.: Risk analysis and risk policy in the Netherlands and the EEC. *Journal of Loss Prevention in the Process Industries*, 4(1):58–64, 1991. (cited on page 14)

Andrieu, C. and Thoms, J.: A tutorial on adaptive MCMC. *Statistics and Computing*, 18(4):343–373, 2008. (cited on page 199)

Angelikopoulos, P., Papadimitriou, C., and Koumoutsakos, P.: X-TMCMC: Adaptive kriging for Bayesian inverse modeling. *Computer Methods in Applied Mechanics and Engineering*, 289:409–428, 2015. (cited on page 194)

Arias, E.: United States Life Tables. *National Vital Statistics Reports*, 64(11), 2011. (cited on page 14)

Atkinson, K.: *The numerical solution of integral equations of the second kind*, volume 4. Cambridge university press, 1997. (cited on page 250, 252)

Atkinson, K. and Shampine, L.: Algorithm 876: Solving Fredholm integral equations of the second kind in Matlab. *ACM Transactions on Mathematical Software (TOMS)*, 34(4):21, 2008. (cited on page 252)

Au, S.-K.: On mcmc algorithm for subset simulation. *Probabilistic Engineering Mechanics*, 43:117–120, 2016. (cited on page 64)

Au, S.-K. and Beck, J. L.: Estimation of small failure probabilities in high dimensions by Subset Simulation. *Probabilistic Engineering Mechanics*, 16(4):263–277, 2001. (cited on page 60, 61, 102, 109, 110, 113, 114, 165, 205, 208, 269, 270)

Au, S.-K. and Patelli, E.: Rare event simulation in finite-infinite dimensional space. *Reliability Engineering & System Safety*, 148:67–77, 2016. (cited on page 64)

Au, S.-K. and Wang, Y.: *Engineering risk assessment with subset simulation*. John Wiley & Sons, 2014. (cited on page 60)

Aven, T.: *Risk, surprises and black swans: fundamental ideas and concepts in risk assessment and risk management*. Routledge, 2014. (cited on page 92)

Bayes, T. and Price, R.: An Essay towards solving a Problem in the Doctrine of Chances. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, M. A. and F. R. S. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1763. (cited on page 21)

Beaumont, M. A., Cornuet, J.-M., Marin, J.-M., and Robert, C. P.: Adaptive approximate Bayesian computation. *Biometrika*, 2009. (cited on page 3, 154)

Beck, J.: Probability logic, information quantification and robust predictive system analysis. *EERL Report*, (2008-05), 2008. (cited on page 17, 22)

Beck, J. and Katafygiotis, L.: Updating of a model and its uncertainties utilizing dynamic test data. In *Computational Stochastic Mechanics*, pages 125–136. Springer, 1991. (cited on page 155)

Beck, J. L.: Bayesian system identification based on probability logic. *Structural Control and Health Monitoring*, 17(7):825–847, 2010. (cited on page 16, 17, 18, 19, 20, 21, 22, 27, 89, 90, 130, 142)

Beck, J. L.: Bayesian system identification and the Bayesian Ockham razor. In *Proceedings of the 9th International Conference on Structural Dynamics, EURODYN 2014*, pages 185–192, 2014. (cited on page 21, 22, 23, 26, 90, 130)

Beck, J. L. and Au, S.-K.: Bayesian updating of structural models and reliability using Markov chain Monte Carlo simulation. *Journal of Engineering Mechanics*, 128(4):380–391, 2002. (cited on page 157)

Beck, J. L. and Katafygiotis, L. S.: Updating models and their uncertainties. I: Bayesian statistical framework. *Journal of Engineering Mechanics*, 124(4):455–461, 1998. (cited on page 155)

Beck, J. L. and Yuen, K.-V.: Model selection using response measurements: Bayesian probabilistic approach. *Journal of Engineering Mechanics*, 130(2):192–203, 2004. (cited on page 22, 129, 155)

Bellhouse, D. R.: Bayes's portrait. *IMS Bulletin*, 17(3):279–278, 1988. (cited on page 21)

Bellhouse, D. R.: The reverend Thomas Bayes, FRS: A biography to celebrate the tercentenary of his birth. *Quality control and applied statistics*, 50(3):327, 2005. (cited on page 21)

Berger, J. et al.: The case for objective Bayesian analysis. *Bayesian Analysis*, 1(3):385–402, 2006. (cited on page 141)

Berger, J. O. and Bernardo, J. M.: On the development of reference priors. *Bayesian statistics*, 4(4):35–60, 1992. (cited on page 142)

Berger, J. O., Bernardo, J. M., and Sun, D.: The formal definition of reference priors. *The Annals of Statistics*, pages 905–938, 2009. (cited on page 142)

Bernardo, J. M.: Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 113–147, 1979. (cited on page 142)

Beskos, A., Roberts, G., Stuart, A., et al.: Optimal scalings for local metropolis–hastings chains on nonproduct targets in high dimensions. *The Annals of Applied Probability*, 19(3):863–898, 2009. (cited on page 52)

Betz, W., Beck, J. L., Papaioannou, I., and Straub, D.: Bayesian inference with Subset Simulation: strategies and improvements. *submitted to Computer Methods in Applied Mechanics and Engineering*, XX:XX, 2017. (cited on page 159, 178)

Betz, W., Mok, C. M., Papaioannou, I., and Straub, D.: Bayesian model calibration using structural reliability methods: application to the hydrological abc model. In M. Beer, S.-K. Au, and J. W. Hall, editors, *Vulnerability, Uncertainty, and Risk: Quantification, Mitigation, and Management*, pages 2734–2743. ASCE, 2014a. (cited on page 160)

Betz, W., Papaioannou, I., and Straub, D.: Adaptive variant of the BUS approach to Bayesian updating. In *Proceedings of 9th European Conference on Structural Dynamics*, 2014b. (cited on page 160, 178)

Betz, W., Papaioannou, I., and Straub, D.: Numerical methods for the discretization of random fields by means of the Karhunen–Loève expansion. *Computer Methods in Applied Mechanics and Engineering*, 271:109–129, 2014c. (cited on page 206, 247, 248, 252, 253)

Betz, W., Papaioannou, I., and Straub, D.: Closure to discussion on Transitional Markov Chain Monte Carlo: Observations and Improvements. *Journal of Engineering Mechanics*, submitted for review, 2016a. (cited on page 194, 202)

Betz, W., Papaioannou, I., and Straub, D.: Transitional Markov Chain Monte Carlo: Observations and Improvements. *Journal of Engineering Mechanics*, 142(5):04016016, 2016b. (cited on page 194, 202)

Bingham, N.: Finite additivity versus countable additivity, 2010. (cited on page 20, 22)

Bjerager, P.: Probability integration by directional simulation. *Journal of Engineering Mechanics*, 114(8):1285–1302, 1988. (cited on page 102)

Boltzmann, L.: *Paper number 39 in Gesammelte Werke*. Vienna Academy, 1877. (cited on page 241)

Breitung, K.: Asymptotic approximations for multinormal integrals. *Journal of Engineering Mechanics*, 110(3):357–366, 1984. (cited on page 102)

Brown, L. D., Cai, T. T., and DasGupta, A.: Interval estimation for a binomial proportion. *Statistical Science*, pages 101–117, 2001. (cited on page 106, 222)

Cheung, S. H. and Beck, J. L.: Bayesian model updating using hybrid Monte Carlo simulation with application to structural dynamic models with many uncertain parameters. *Journal of engineering mechanics*, 135(4):243–255, 2009. (cited on page 3, 22, 155, 198)

Cheung, S. H. and Beck, J. L.: Calculation of posterior probabilities for Bayesian model class assessment and averaging from posterior samples based on dynamic system data. *Computer-Aided Civil and Infrastructure Engineering*, 25(5):304–321, 2010. (cited on page 22, 141, 156)

Chiachio, M., Beck, J. L., Chiachio, J., and Rus, G.: Approximate bayesian computation by subset simulation. *SIAM Journal on Scientific Computing*, 36(3):A1339–A1358, 2014. (cited on page 3)

Ching, J. and Chen, Y.-C.: Transitional Markov chain Monte Carlo method for Bayesian model updating, model class selection, and model averaging. *Journal of engineering mechanics*, 133(7):816–832, 2007. (cited on page 3, 154, 194, 195, 196, 199, 202)

Ching, J. and Wang, J.-S.: Discussion on Transitional Markov Chain Monte Carlo: Observations and Improvements. *Journal of Engineering Mechanics*, submitted for review, 2016. (cited on page 202)

Chopin, N.: A sequential particle filter method for static models. *Biometrika*, 89(3):539–552, 2002. (cited on page 3, 194)

Cover, T. M. and Thomas, J. A.: Elements of information theory, 2006. (cited on page 37, 130)

Cowles, M. K. and Carlin, B. P.: Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91(434):883–904, 1996. (cited on page 50)

Cox, R. T.: Probability, frequency and reasonable expectation. *American journal of physics*, 14(1):1–13, 1946. (cited on page 5, 16, 17, 18, 20)

Csilléry, K., Blum, M. G., Gaggiotti, O. E., and François, O.: Approximate Bayesian computation (ABC) in practice. *Trends in ecology & evolution*, 25(7):410–418, 2010. (cited on page 3)

DAV: DAV-Unfallstatistik 2013: Übersicht. Technical report, Deutscher Alpenverein, 2014. `http://www.alpenverein.`

de/chameleon/public/8a5a1748-32ef-6de4-9b4a-9461b78b9922/140805-Unfallstatistik-Uebersicht-2013_24144.pdf. (cited on page 14)

De Finetti, B.: Foresight: its logical laws in subjective sources. *Studies in Subjective Probability*, pages 93–158, 1964. (cited on page 22)

De Finetti, B.: *Philosophical Lectures on Probability: collected, edited, and annotated by Alberto Mura*, volume 340. Springer Science & Business Media, 2008. (cited on page 20, 22)

Der Kiureghian, A. and Ditlevsen, O.: Aleatory or epistemic? Does it matter? *Structural Safety*, 31(2):105–112, 2009. (cited on page 10)

Der Kiureghian, A. and Liu, P.-L.: Structural reliability under incomplete probability information. *Journal of Engineering Mechanics*, 112(1):85–104, 1986. (cited on page 43, 44, 199)

DESTATIS: Allgemeine Sterbetafeln für Deutschland, 2010/12. Technical report, Statistisches Bundesamt, Wiesbaden, Germany, 2015. https://www.destatis.de/DE/Publikationen/Thematisch/Bevoelkerung/Bevoelkerungsbewegung/PeriodensterbetafelnBundeslaender5126204127004.pdf?__blob=publicationFile. (cited on page 14)

DiazDelaO, F., Garbuno-Inigo, A., Au, S., and Yoshida, I.: Bayesian updating and model class selection with Subset Simulation. *Computer Methods in Applied Mechanics and Engineering*, 2017. (cited on page 157, 160, 166, 188, 189, 204)

Ditlevsen, O.: Generalized second moment reliability index. *Journal of Structural Mechanics*, 7(4):435–451, 1979. (cited on page 96)

Ditlevsen, O. and Madsen, H. O.: *Structural reliability methods*. Technical University of Denmark, 2007. (cited on page 43, 44, 45, 95, 96, 100, 102, 159)

Ditlevsen, O., Melchers, R. E., and Gluver, H.: General multi-dimensional probability integration by directional simulation. *Computers & Structures*, 36(2):355–368, 1990. (cited on page 102)

Easwaran, K.: Why countable additivity? *Thought: A Journal of Philosophy*, 2(1):53–61, 2013. (cited on page 20)

Edwards, W., Lindman, H., and Savage, L. J.: Bayesian statistical inference for psychological research. *Psychological review*, 70(3):193–242, 1963. (cited on page 21)

ETSC: Transport Safety Performance in the EU: A Statistical Overview. Technical report, European Transport Safety Council, 2003. http://etsc.eu/wp-content/uploads/2003_transport_safety_stats_eu_overview.pdf. (cited on page 13, 14)

Eurocode 0: Basis of structural design. *European Committee for Standardization*, EN 1990:2010-12, 2015. (cited on page 99, 100, 211)

Finney, D.: On a method of estimating frequencies. *Biometrika*, 36(1/2):233–234, 1949. (cited on page 162)

Franck, I. M. and Koutsourelakis, P.: Sparse variational bayesian approximations for nonlinear inverse problems: Applications in nonlinear elastography. *Computer Methods in Applied Mechanics and Engineering*, 299:215–244, 2016. (cited on page 3)

Gelman, A., Barlin, J. B., Stern, H. S., and Rubin, D. B.: *Bayesian Data Analysis*. Chapman & Hall/CRC, 2nd edition, 2004a. (cited on page 2, 50, 154, 155, 267, 268)

Gelman, A., Carlin, J. B., Rubin, D. B., and Stern, H. S.: Bayesian data analysis, 2004b. (cited on page 199)

Gelman, A., Roberts, G. O., and Gilks: Efficient Metropolis jumping rules. In *Bayesian Statistics 5: Proceedings of the Fifth Valencia International Meeting*, pages 599–607. Oxford University Press, 1996. (cited on page 52, 68, 202)

Gelman, A. and Rubin, D. B.: Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472, 1992. (cited on page 267)

Ghanem, R.: Ingredients for a general purpose stochastic finite elements implementation. *Computer Methods in Applied Mechanics and Engineering*, 168(1):19–34, 1999. (cited on page 249)

Ghanem, R.-G. and Spanos, P.-D.: *Stochastic Finite Elements - A Spectral Approach*. Springer, New York, 1991. (cited on page 248, 249, 250, 251)

Gigerenzer, G.: *Risiko: Wie man die richtigen Entscheidungen trifft*. C. Bertelsmann Verlag, 2013. (cited on page 11)

Gilks, W., Richardson, S., and Spiegelhalter, D., editors: *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC, 1st edition, 1996. (cited on page 2, 50, 51, 154, 155)

Grigoriu, M.: Crossing of non-Gaussian translation processes. *Journal of Engineering Mechanics, ASCE*, 110(41):610–620, 1984. (cited on page 254)

Grigoriu, M.: Simulation of stationary non-Gaussian translation processes. *Journal of Engineering Mechanics, ASCE*, 124(2):121–126, 1998. (cited on page 255)

Haario, H., Saksman, E., and Tamminen, J.: Componentwise adaptation for high dimensional MCMC. *Computational Statistics*, 20(2):265–273, 2005. (cited on page 3, 155)

Hadjidoukas, P., Angelikopoulos, P., Papadimitriou, C., and Koumoutsakos, P.: Π4U: A high performance computing framework for Bayesian uncertainty quantification of complex models. *Journal of Computational Physics*, 284:1–21, 2015. (cited on page 194)

Haldane, J.: A note on inverse probability. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 28, pages 55–61. Cambridge Univ Press, 1932. (cited on page 106)

Haldane, J.: On a method of estimating frequencies. *Biometrika*, 33(3):222–225, 1945. (cited on page 162)

Hartigan, J. A.: *Bayes theory*. Springer Science & Business Media, 2012. (cited on page 21)

Hasofer, A. and Lind, N.: An exact and invariant first-order reliability format. *Journal of the Engineering Mechanics Division ASCE*, 100:111–121, 1974. (cited on page 102, 160)

Hastings, W. K.: Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. (cited on page 52, 171)

Hesterberg, T.: Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37(2):185–194, 1995. (cited on page 86)

Hohenbichler, M. and Rackwitz, R.: Non-normal dependent vectors in structural safety. *Journal of the Engineering Mechanics Division ASCE*, 107(6):1227–1238, 1981. (cited on page 42, 199)

Hohenbichler, M. and Rackwitz, R.: Improvement of second-order reliability estimates by importance sampling. *Journal of Engineering Mechanics*, 114(12):2195–2199, 1988. (cited on page 102, 160)

Hurtado, J.: Analysis of one-dimensional stochastic finite elements using neural networks. *Probabilistic Engineering Mechanics*, 17(1):35–44, 2002. (cited on page 252)

Hájek, A.: Interpretations of probability. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2012 edition, 2012. (cited on page 23)

Jaynes, E. T.: Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957. (cited on page 22, 106)

Jaynes, E. T.: *Papers on probability, statistics and statistical physics*. D. Reidel Publishing, Dordrecht, Holland, 1983. (cited on page 142, 163)

Jaynes, E. T.: *Probability theory: the logic of science*. Cambridge university press, 2003. (cited on page 12, 16, 20, 22, 24, 26, 106, 142, 163)

Jaynes, E. T. and Kempthorne, O.: Confidence intervals vs bayesian intervals. In *Foundations of probability theory, statistical inference, and statistical theories of science*, pages 175–257. Springer, 1976. (cited on page 94)

JCSS: JCSS Probabilistic Model Code. Technical report, Joint Committee on Structural Safety (JCSS), 2001–2015. (cited on page 211)

Jeffreys, H.: An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461, 1946. (cited on page 21, 106, 142)

Jeffreys, H.: *The theory of probability*. Oxford University Press, 1998. (cited on page 21, 106, 142)

Jensen, H., Millas, E., Kusanovic, D., and Papadimitriou, C.: Model-reduction techniques for Bayesian finite element model updating using dynamic response data. *Computer Methods in Applied Mechanics and Engineering*, 279:301–324, 2014. (cited on page 194)

Kadane, J. B.: *Principles of uncertainty*. CRC Press, 2011. (cited on page 22)

Karhunen, K.: Über lineare Methoden in der Wahrscheinlichkeitsrechnung. *Annales Academia Scientiarum Fennica*, 37:3–79, 1947. (cited on page 249)

Kass, R. E. and Wasserman, L.: The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435):1343–1370, 1996. (cited on page 141, 142)

Katafygiotis, L. S. and Zuev, K. M.: Geometric insight into the challenges of solving high-dimensional reliability problems. *Probabilistic Engineering Mechanics*, 23(2):208–218, 2008. (cited on page 66)

Keeton, C. R.: On statistical uncertainty in nested sampling. *Monthly Notices of the Royal Astronomical Society*, 414(2):1418–1426, 2011. (cited on page 193)

Kiureghian, A. D.: Measures of structural safety under imperfect states of knowledge. *Journal of Structural Engineering*, 115(5):1119–1140, 1989. (cited on page 96)

Kolmogorov, A. N.: Grundbegriffe der Wahrscheinlichkeitrechnung. *Ergebnisse Der Mathematik*, 1933. (cited on page 19)

Koutsourelakis, P., Pradlwarter, H., and Schuëller, G.: Reliability of structures in high dimensions, Part I: algorithms and applications. *Probabilistic Engineering Mechanics*, 19(4):409–417, 2004. (cited on page 102, 160)

Koutsourelakis, P.-S.: A multi-resolution, non-parametric, Bayesian framework for identification of spatially-varying model parameters. *Journal of computational physics*, 228(17):6184–6211, 2009a. (cited on page 3, 207)

Koutsourelakis, P.-S.: Accurate uncertainty quantification using inaccurate computational models. *SIAM Journal on Scientific Computing*, 31(5):3274–3300, 2009b. (cited on page 3, 207)

Kulhawy, F. H., Trautmann, C., Beech, J., O'Rourke, T., and W, M.: Transmission line structure foundations for uplift-compression loading. Technical report, Electric Power Research Institute, 1983. (cited on page 100)

Laplace, P. S.: Memoir on the probability of the causes of events. *Statistical Science*, 1(3):364–378, 1986. (cited on page 3)

Lebrun, R. and Dutfoy, A.: An innovating analysis of the Nataf transformation from the copula viewpoint. *Probabilistic Engineering Mechanics*, 24(3):312–320, 2009. (cited on page 44, 45)

Li, C.-C. and Der Kiureghian, A.: Optimal discretization of random fields. *Journal of Engineering Mechanics*, 119(6):1136–1154, 1993. (cited on page 206, 248, 253, 254, 255)

Lindley, D.: The future of statistics: a Bayesian 21st century. *Advances in Applied Probability*, 7:106–115, 1975. (cited on page 9)

Lindley, D. V.: *Understanding Uncertainty*. John Wiley & Sons, 2006. (cited on page 10, 22)

Liu, J. S.: *Monte Carlo strategies in scientific computing*. Springer, 2001. (cited on page 110)

Loève, M.: *Fonctions aleatoire du second ordre, supplement to P. Levy, Processus Stochastic et Mouvement Brownien*. Gauthier-Villars, Paris, 1948. (cited on page 249)

MacKay, D. J.: Bayesian interpolation. *Neural computation*, 4(3):415–447, 1992a. (cited on page 133)

MacKay, D. J.: *Bayesian methods for adaptive models*. Ph.D. thesis, California Institute of Technology, 1992b. (cited on page 155)

MacKay, D. J.: Probable networks and plausible predictions - a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, 6(3):469–505, 1995. (cited on page 3)

MacKay, D. J.: *Information theory, inference and learning algorithms*. Cambridge university press, 2003. (cited on page 37, 106)

Matthies, H. G. and Keese, A.: Galerkin methods for linear and nonlinear elliptic stochastic partial differential equations. *Computer Methods in Applied Mechanics and Engineering*, 194(12):1295–1331, 2005. (cited on page 249)

Melchers, R. E.: *Structural reliability analysis and prediction*. John Wiley & Son Ltd, 1999. (cited on page 13, 14, 95, 96, 97, 100, 102, 159)

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E.: Equation of state calculations by fast computing machines. *J. Chem. Phys*, 21:1087, 1953. (cited on page 52, 53, 171)

Muto, M. and Beck, J. L.: Bayesian updating and model class selection for hysteretic structural models using stochastic simulation. *Journal of Vibration and Control*, 14(1-2):7–34, 2008. (cited on page 22, 130)

Neal, R. M.: Probabilistic inference using Markov chain Monte Carlo methods, 1993. (cited on page 110)

Neal, R. M.: MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, editors, *Handbook of Markov Chain Monte Carlo*, chapter 5, pages 113–162. Chapman & Hall, 2011. (cited on page 3, 155)

Neal, R. M. and Hinton, G. E.: A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer, 1998. (cited on page 3)

Ortiz, G. A., Alvarez, D. A., and Bedoya-Ruíz, D.: Identification of Bouc–Wen type models using the Transitional Markov Chain Monte Carlo method. *Computers & Structures*, 146:252–269, 2015. (cited on page 194)

Papaioannou, I., Betz, W., and Straub, D.: Bayesian model updating of a tunnel in soft soil with settlement measurements. *Geotechnical Safety and Risk IV*, page 351, 2013. (cited on page 160)

Papaioannou, I., Betz, W., Zwirglmaier, K., and Straub, D.: MCMC algorithms for subset simulation. *Probabilistic Engineering Mechanics*, 41:89–103, 2015. (cited on page 6, 58, 64, 65, 69, 83, 84, 85, 110, 112, 188, 191, 199, 204, 206, 209)

Pasarica, C. and Gelman, A.: Adaptively scaling the Metropolis algorithm using expected squared jumped distance. *Statistica Sinica*, pages 343–364, 2010. (cited on page 52, 85, 86)

Paté-Cornell, M. E.: Quantitative safety goals for risk management of industrial facilities. *Structural Safety*, 13(3):145–157, 1994. (cited on page 14)

Phoon, K., Huang, H., and Quek, S.: Simulation of strongly non-Gaussian processes using Karhunen–Loève expansion. *Probabilistic Engineering Mechanics*, 20(2):188–198, 2005. (cited on page 249)

Phoon, K., Huang, S., and Quek, S.: Simulation of second-order processes using Karhunen–Loève expansion. *Computers & Structures*, 80(12):1049–1060, 2002. (cited on page 249)

Plummer, M., Best, N., Cowles, K., and Vines, K.: CODA: Convergence diagnosis and output analysis for MCMC. *R news*, 6(1):7–11, 2006. (cited on page 2, 155)

Popper, K. R.: The propensity interpretation of the calculus of probability, and the quantum theory. *Observation and Interpretation*, pages 65–70, 1957. (cited on page 23)

Press, W., VETTERLING, W., Flannery, B., and TEUKOLSKY, S.: Numerical recipes in C: the art of scientific computing. 2. *Cambridge university press*, 1:2–3, 1993. (cited on page 252)

Rackwitz, R.: Reliability analysis – a review and some perspectives. *Structural safety*, 23(4):365–395, 2001. (cited on page 102, 160)

Rackwitz, R. and Flessler, B.: Structural reliability under combined random load sequences. *Computers & Structures*, 9(5):489–494, 1978. (cited on page 102, 160)

Ramsey, F. P.: Truth and probability (1926). *The foundations of mathematics and other logical essays*, pages 156–198, 1931. (cited on page 22)

Robert, C. P. and Casella, G.: *Monte Carlo statistical methods*. Springer, 2nd edition, 2004. (cited on page 50, 110, 198)

Robert, G. and Tweedie, R.: Exponential convergence of Langevin diffusions and their discrete approximation. *Bernoulli*, 2:341–363, 1996. (cited on page 3, 155)

Roberts, G. O., Gelman, A., Gilks, W. R., et al.: Weak convergence and optimal scaling of random walk metropolis algorithms. *The annals of applied probability*, 7(1):110–120, 1997. (cited on page 52, 68, 200, 202)

Roberts, G. O. and Rosenthal, J. S.: Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18(2):349–367, 2009. (cited on page 52, 69)

Roberts, G. O., Rosenthal, J. S., et al.: Optimal scaling for various Metropolis-Hastings algorithms. *Statistical science*, 16(4):351–367, 2001. (cited on page 52, 68, 200)

Rosenblatt, M.: Remarks on a multivariate transformation. *The annals of mathematical statistics*, pages 470–472, 1952. (cited on page 42)

Simoen, E., Papadimitriou, C., and Lombaert, G.: On prediction error correlation in Bayesian model updating. *Journal of Sound and Vibration*, 2013. (cited on page 150)

Skilling, J.: Bayesian computation in big spaces-nested sampling and Galilean Monte Carlo. In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 31st International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, volume 1443, pages 145–156. AIP Publishing, 2012. (cited on page 191, 193)

Skilling, J. et al.: Nested sampling for general Bayesian computation. *Bayesian analysis*, 1(4):833–859, 2006. (cited on page 3, 190, 191, 192, 193)

Smith, A. F. and Gelfand, A. E.: Bayesian statistics without tears: a sampling–resampling perspective. *The American Statistician*, 46(2):84–88, 1992. (cited on page 162)

Stefanou, G. and Papadrakakis, M.: Assessment of spectral representation and Karhunen–Loève expansion methods for the simulation of Gaussian stochastic fields. *Computer Methods in Applied Mechanics and Engineering*, 196(21):2465–2477, 2007. (cited on page 250)

Stigler, S. M.: *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press, 1986. (cited on page 21)

Straub, D.: Reliability updating with equality information. *Probabilistic Engineering Mechanics*, 26(2):254–258, 2011. (cited on page 161)

Straub, D.: Engineering risk assessment. In *Risk – A Multidisciplinary Introduction*, pages 333–362. Springer, 2014. (cited on page 159)

Straub, D.: *Lecture Notes in Engineering Risk Analysis*. Engineering Risk Analysis Group, Technische Universität München, Germany, 2016. (cited on page 237)

Straub, D. and Papaioannou, I.: Bayesian updating with structural reliability methods. *Journal of Engineering Mechanics*, 2015. (cited on page 3, 154, 157, 159, 160, 161, 162, 166, 203)

Straub, D., Papaioannou, I., and Betz, W.: Bayesian analysis of rare events. *Journal of Computational Physics*, 314:538–556, 2016. (cited on page 160, 161)

Sudret, B. and Der Kiureghian, A.: Stochastic Finite Element Methods and Reliability - A State-of-the-Art Report. Technical Report UCB/SEMM-2000/08, Department of Civil & Environmental Engineering, Univ. of California, Berkeley, 2000. (cited on page 248, 250, 251)

Taleb, N. N.: *The Black Swan: The Impact of the Highly Improbable*, volume 2. Random House, 2007. (cited on page 92)

Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P.: Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–518, 1997. (cited on page 154)

Thiéry, A.: Optimal proposals for MCMC methods. First year phd report, University of Warwick, 2010. (cited on page 52, 85)

Tierney, L.: Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):1701–1728, 1994. (cited on page 50, 51, 171, 172)

Turner, B. M. and Van Zandt, T.: A tutorial on approximate Bayesian computation. *Journal of Mathematical Psychology*, 56(2):69–85, 2012. (cited on page 3)

Ullmann, E. and Papaioannou, I.: Multilevel estimation of rare events. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):922–953, 2015. (cited on page 207)

Vanik, M. W., Beck, J., and Au, S.: Bayesian probabilistic approach to structural health monitoring. *Journal of Engineering Mechanics*, 126(7):738–745, 2000. (cited on page 22)

Vanmarcke, E.: *Random Fields: Analysis and Synthesis*. World Scientific Publishing, Singapore, 2 edition, 2010. (cited on page 249)

Wan, X. and Karniadakis, G.: A sharp error estimate for the fast Gauss transform. *Journal of computational physics*, 219(1):7–12, 2006. (cited on page 252)

Wasserman, L.: Bayesian model selection and model averaging. *Journal of mathematical psychology*, 44(1):92–107, 2000. (cited on page 133)

Zadeh, L. A.: Fuzzy sets. *Information and control*, 8(3):338–353, 1965. (cited on page 24)

Zheng, W. and Chen, Y.-T.: Novel probabilistic approach to assessing barge–bridge collision damage based on vibration measurements through transitional Markov chain Monte Carlo sampling. *Journal of Civil Structural Health Monitoring*, 4(2):119–131, 2014. (cited on page 194)

Zhu, H., Zeng, X., Cai, W., Xue, J., and Zhou, D.: A sparse grid based spectral stochastic collocation method for variations-aware capacitance extraction of interconnects under nanometer process technology. In *Design, Automation & Test in Europe Conference & Exhibition, 2007. DATE'07*, pages 1–6. IEEE, 2007. (cited on page 252)

Zuev, K. M., Beck, J. L., Au, S.-K., and Katafygiotis, L. S.: Bayesian post-processor and other enhancements of Subset Simulation for estimating failure probabilities in high dimensions. *Computers & Structures*, 92:283–296, 2012. (cited on page 22, 84, 85, 105, 106, 114, 208)