Technische Universität München

Department of Mathematics

Bachelor's Thesis

On Multilevel Algorithms for the Estimation of Failure Probabilities and Rare Event Simulation

Konstantin Riedl

Supervisor: Prof. Dr. Elisabeth Ullmann
Advisor: Prof. Dr. Elisabeth Ullmann
Submission Date: August 28, 2018

I hereby declare that this thesis is my own work and that no other sources have been used except those clearly indicated and referenced.

Konstantin Riedl

Munich, August 28, 2018

Abstract

Multilevel algorithms play an important role in the estimation of rare event probabilities for computationally expensive systems.
This thesis introduces and investigates a multilevel estimator for cross-entropy based importance sampling. A hierarchy of approximations of different quality is used to efficiently derive a suitable biasing density for importance sampling. This involves solving optimization problems with respect to the Kullback-Leibler divergence. Furthermore, it is shown that a selective refinement strategy can be employed. These modifications lead to significantly reduced computational cost compared to the single level version.
Numerical experiments in one- and two-dimensional physical space demonstrate the applicability of the method to small failure probabilities.

Zusammenfassung

Multilevelalgorithmen kommt beim Schätzen von Wahrscheinlichkeiten seltener Ereignisse eine wichtige Rolle zu, insbesondere bei rechenintensiven Systemen.
In dieser Arbeit wird ein Multilevelschätzer vorgestellt und untersucht, welchem Cross-Entropy basiertes Importance Sampling zu Grunde liegt. Eine hierarchische Anordnung von Approximationen verschiedener Ausflösungen wird bei der effizienten Konstruktion einer problemspezifischen Dichte für Importance Sampling verwendet. Dabei werden Optimierungsprobleme bezüglich der Kullback-Leibler Divergenz betrachtet und gelöst. Des Weiteren stellt sich heraus, dass von einer selektiven Verfeinerungstechnik Gebrauch gemacht werden kann. Diese Veränderungen führen zu einem erheblich verringerten Rechenaufwand im Vergleich zu der Variante, bei der nur eine einzige Diskretisierung verwendet wird.
Numerische Experimente im Ein- und Zweidimensionalen demonstrieren, dass die Methode für kleine Fehlerwahrscheinlichkeiten geeignet ist.

Contents

Introduction

Several environmental, engineering and biological systems are based on the — in many situations certainly not too unfounded — hope or presumption that some event is very unlikely to happen. Such a, therefore called, rare event is mostly associated with a failure of the corresponding system.
Mathematical models usually use partial differential equations in combination with suitable boundary conditions to describe the underlying physical processes. Additionally to the well-known formulas of the physical laws also data, for example, a material parameter, is needed to characterize the particular situation. Mostly this information is affected by uncertainty due to a lack of accurate measurements or simply not being able to depict reality exactly. Incorporating this uncertainty of the input data into the partial differential equation yields a stochastic partial differential equation, whose solution will not be deterministic.
Generally, in an application-oriented context certain properties of the solution are the quantities of interest. Mostly they can be expressed as functionals.
In the risk and reliability analysis of such systems rare events and their probabilities, so called failure probabilities, are of high relevance. A failure occurs if the value of a particular quantity of interest falls below (or exceeds) a critical threshold.
Consider, for instance, groundwater flow, which can be described by Darcy's law, in combination with a radioactive waste disposal. Since it is impossible to have a full knowledge about the structure of the soil, a way out is to model it as a random porous medium. In this example it is hopefully very unlikely, in case of an unintentional incident, for contaminated water to leave some security zone around the repository within short amount of time.
Typically the procedure to achieve an accurate and reliable value for a rare event probability consists of three major steps. The first one is concerned with the challenge of how to quantitively describe the uncertainty in the model. It is followed by the problem of (numerically) solving the stochastic partial differential equation. The final task is the evaluation of the functional to obtain the quantity of interest and therefore its failure probability. Each of these steps can be a challenge in its own right.
In order to include the uncertainty into the model, the coefficients, the right hand side and the boundary conditions are represented as random fields, which are random variables over some appropriate probability space with values in some function space. A typical modeling assumption, known as the finite-dimensional noise assumption, is that the randomness in the input data can be represented or at least approximated sufficiently well by a finite number of stochastic degrees of freedom.
For tackling the solving procedure several different approaches have been developed in recent years. The one, which is taken in this thesis, is sampling based. A finite number of realizations of the random fields is generated and for each fixed realization the (then

deterministic) variational form of the partial differential equation is solved using standard finite element, finite difference or discontinuous Galerkin methods. This approach is referred to as Monte Carlo method.
From the range of these single realizations statistics of the solution can be obtained, like mean, variance or some moments. But also more evolved quantities of interest can be computed after further processing the solution.
Rare events are the type of quantity dealt with in this thesis. Characteristically they occur with a very small probability, in the range of 10^{-9} to 10^{-6}. This complicates their estimation, because in the case of the Monte Carlo method the desire for a precise result requires an enormous amount of samples in the order of the inverse of the occurrence probability. For an expensive to evaluate model this quickly becomes unaffordable, both computationally and in terms of time, although at least for the latter one it is worth mentioning that sampling based methods in general benefit from parallelization.
Hence, sophisticated remedies to crude Monte Carlo sampling are indispensable [RC04, RK17].
The high computational effort for a single sample motivates to employ a surrogate, i.e., an approximative model imitating the behavior of the true model but coming at lower cost. In [LX10] the widely used generalized polynomial chaos expansion [XK02] serves, after truncation, as surrogate and is combined with the original model resulting in a hybrid method. This strategy is enhanced in [LLX11] with the aim of addressing rare events. Both methods reduce the cost without admitting accuracy loss, but have not been applied to complex systems and are not suitable for a high stochastic dimension.
Variance reduction techniques, which have been designed particularly for high-dimensional stochastic sample spaces, include subset simulation [AB01, AW14] and line sampling [KPS04]. The former one, also known as splitting or importance splitting, introduces intermediate nested failure regions and expresses the rare event probability as a product of conditional probabilities corresponding to these nested failure regions. Sampling with respect to the regions employs Markov chain Monte Carlo [MRRT53]. An advantage of this strategy is that prior knowledge about the model in not necessary, meaning that it can be used as a black-box. The desired variance reduction is obtained by conditioning.
A further prominent variance reduction technique is importance sampling [RC04, Owe13, TK10], where samples are generated from a problem-specific biasing distribution. The choice of this distribution is significant for the success of the strategy and in general not evident, especially for high stochastic dimensions [AB03]. Even though the optimal importance distribution, which leads to a zero-variance estimator, is known analytically, it is not usable as it involves the rare event region. The cross-entropy method [dBKMR05, KRG13, RK04, HdMR02] provides an effective procedure to obtain an approximately optimal importance distribution among a prescribed family of distributions. It is iterative and generates nested failure regions, which approach the final rare event region. This involves minimization problems subject to this family with respect to the Kullback-Leibler divergence [Kul59]. Typical choices for such families are Gaussian distributions or exponential distributions; in [GPS19] the use of Gaussian mixtures is investigated and furthermore observed that neither is suitable for high-dimensional problems. [BK08] introduces a new adaptive method, which adopts aspects of the cross-entropy method but overcomes the likelihood ratio degeneracy issues importance sampling faces in high dimensions.
However, despite the achieved variance reduction, the computational cost of subset sim-

ulation and importance sampling utilizing the cross-entropy method is still significantly large if the model evaluation is costly. To further reduce this cost, improvements of both methods have been influenced by a multilevel idea introduced for Monte Carlo in [Hei01, Gil08]. The so-called multilevel Monte Carlo method leverages a hierarchy of numerical approximations of decreasing accuracy and computational effort in order to estimate certain statistics of the model [CGST11, Gil15]. The resulting low accuracy solutions are used as control variates for the solutions with high accuracy. Variance reduction is obtained since large parts of the uncertainty can be captured on the cheap models. This, as a consequence, saves cost as the number of necessary computations on the most expensive level decreases. The method has been extended in [PWG16] such that any kind of surrogate can be used. Unfortunately, the application to rare event estimation is not straightforward and suffers from the non-smoothness of the failure probability functional. Results in quantile estimation [EEHM14] have established a selective refinement strategy, which takes advantage of the special shape of the failure probability functional. The idea is that in many cases the information from a low accuracy model already suffices to decide with certainty whether failure occurs or not. This strategy promising less evaluations of the high accuracy model has been combined with multilevel Monte Carlo in [EHM16, FHMN16]. The resulting method convinces with asymptotic cost, which is as high as solving a single high accuracy model. However, it is not capable of addressing really small failure probabilities.
In contrast, a multilevel version of subset simulation, introduced in [UP15], concerns rare event probabilities and moreover copes with a high stochastic dimension.
This thesis investigates a novel multilevel approach to importance sampling utilizing the cross-entropy method, which has been proposed by Peherstorfer, Kramer and Willcox in a slightly more general manner in [PKW18]. Instead of deriving the importance distribution from realizations of the high accuracy model a hierarchy of models of lower accuracy and cost is exploited to efficiently construct a biasing distribution using a multilevel version of the cross-entropy method. This effects a significant speed-up without loosing accuracy. It is furthermore observed that a selective refinement strategy as in [EHM16] can be employed. Generalizations of the method presented here include the use of multifidelity hierarchies [PCMW16, PWG18] instead of a multilevel hierarchy.
The outline of this thesis is as follows. In the light of the groundwater flow example from the beginning Chapter 1 describes the problem of estimating failure probabilities. Chapter 2 revives crude Monte Carlo sampling, before importance sampling is introduced and the cross-entropy method described. The multilevel idea and the selective refinement strategy are subject of Chapter 3. In Chapter 4 the already mentioned multilevel cross-entropy importance sampling method is presented and investigated. Chapter 5 provides numerical experiments studying a heat transfer and a groundwater flow problem. The thesis concludes with a further discussion.

Chapter 1

Failure Probabilities

This first chapter serves the formal presentation of the problem setup and uses the groundwater flow example from the introduction for the purpose of illustration.
For the sake of completeness random fields and a representation possibility are introduced. Lastly the numerical approximation of the model's solution is discussed briefly.

1.1 Problem Setup

Model. The deterministic model of the physical relevant environment acts as the starting point for the formulation of the stochastic version. Let u denote the solution of

$$\mathcal{M}(u) = 0, \tag{1.1}$$

where $\mathcal{M}$ covers all the physical properties, which have to be satisfied. This, for example, can include one or several partial differential equations (PDEs) and appropriate boundary conditions (BCs). Implicitly equation (1.1) is meant to hold on a certain bounded d-dimensional physical domain $D \subset \mathbb{R}^d$, i.e., more formally $\mathcal{M}\big(u(x)\big) = 0$ for all $x \in D$.
To include uncertainty in the model, which might stem from a lack of knowledge or also from intrinsic variety, let $\big(\Omega, \mathcal{F}, \mathbb{P}\big)$ be an abstract probability space, where Ω denotes the sample space, $\mathcal{F}$ the σ-algebra defined on Ω, rich enough to support all randomness, and $\mathbb{P}$ the probability measure. In order to make the uncertainty more tangible let $\Theta \subset \mathbb{R}^k$ be a set containing k-dimensional parameters and $\big(\Theta, \mathcal{B}(\Theta)\big)$ the corresponding measurable space with respect to the Borel σ-algebra $\mathcal{B}$, versatile enough to parametrize all randomness via the random variable $Z : \Omega \to \Theta$. The pushforward measure $Z\#\mathbb{P}$ is assumed to be absolutely continuous with respect to the k-dimensional Lebesgue measure λ and therefore has, according to the Radon-Nikodym theorem, a density p, i.e., $\mathrm{d}(Z\#\mathbb{P}) = p\,\mathrm{d}\lambda$.
Consequently, and in contrast to the deterministic case, where the parameters have been intrinsically related to the model $\mathcal{M}$, due to their consideration as being uncertain they have to be explicitly taken into account now.
Thus the model depends on these random parameters, which themselves are sampled from the space Ω. The stochastic version of the model (1.1) then reads as

$$\mathcal{M}\big(u, Z(\omega)\big) = 0. \tag{1.2}$$

Accordingly, the dependency on ω also transfers to the solution u, yielding

$$u = u(\omega) = u\big(Z(\omega)\big). \tag{1.3}$$

For a fixed $\omega \in \Omega$, $u(\omega)$ is called a realization of the solution and implicitly stands for $u\big(Z(\omega)\big)$. Regarding the notation, it depends on the context whether $u(\omega)$ or $u(z)$ is written, denoting a realization of Z by $z \in \Theta$. In the sense of Hadamard the problem is assumed to be, at least $\mathbb{P}$-almost surely, well-posed, meaning that existence and uniqueness of the solution are presupposed as well as a continuous dependence on the input data.
Considering the example of radioactive water traveling from a damaged repository through the soil, there are two physical laws, which have to hold; namely for the Darcy velocity $\boldsymbol{u}$, describing the volume of the discharge into the vectors direction, the continuity equation

$$\operatorname{div}(\boldsymbol{u}) = g \quad \text{in } D$$

has to be fulfilled for some source term g. Additionally for $\boldsymbol{u}$ and the hydrostatic pressure p Darcy's law

$$a^{-1}\boldsymbol{u} + \nabla p = 0 \quad \text{in } D$$

has to be valid, where a is the permeability, a local property of the soil, i.e., $a = a(x)$. In the notation of model (1.1), which also includes the boundary conditions on ∂D, the solution u would be a vector containing $\boldsymbol{u}$ and p.
Since it is practically impossible to rely on a precise knowledge of this input quantity, the permeability a is modeled as a lognormal random field, i.e., $a = a(x, \omega)$. To make the random field manageable it is firstly decomposed in a way such that the stochastic and spatial parts appear separately and is secondly truncated with the aim that there is only the need to handle finitely many stochastic degrees of freedom. Thereby it gets possible to understand the stochastic degrees of freedom as the random parameter Z.

Quantity of Interest (QoI). A specific characteristic of a realization of the solution u considered in the following can be expressed as a continuous but not necessarily linear functional

$$X : \mathcal{S} \to \mathbb{R}, \tag{1.4}$$

where $\mathcal{S}$ denotes the space of all possible solutions.
The point evaluation $X_y(u) = u(y)$ at some $y \in D$ or a boundary integral $X_\Gamma(u) = \int_\Gamma u \cdot \mathrm{d}S$ for some $\Gamma \subset \partial D$ would be simple examples. The subsurface flow problem requests a more demanding functional involving particle tracking. Starting from the location of the repository's leak, call it $\boldsymbol{x}_0$, the value to obtain is the time τ it takes a particle, being released in $\boldsymbol{x}_0$, to reach the boundary ∂S, assuming that $S \subset D$ has been chosen as the safety zone. Taking the knowledge of the Darcy velocity $\boldsymbol{u}$ for granted, the relation between the true fluid velocity $\boldsymbol{v}$ and $\boldsymbol{u}$ is given by $\boldsymbol{v} = \frac{\boldsymbol{u}}{\varphi}$, where φ is the porosity of the soil. In order to compute τ the streamline $\boldsymbol{x}(t)$ of the particle is necessary, which satisfies the ordinary differential equation

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{v}, \quad \boldsymbol{x}(0) = \boldsymbol{x}_0.$$

Then the functional's value is

$$X_S(u) = \tau \quad \text{with} \quad \tau = \operatorname*{arg\,min}_{t \in [0,\infty)} \big\{ \boldsymbol{x}(t) \in \partial S \big\}.$$

Having included uncertainty into the model also directly affects on the QoI, more precisely, $X(u(Z)) : \Omega \to \mathbb{R}$ gets a random variable. For convenience but by a slight abuse of

notation X instead of $X\big(u(Z)\big)$ is written from Section 1.3 on, if not explicitly stated otherwise. The cumulative distribution function (cdf) of $X\big(u(Z)\big)$ is denoted by F, on which the following regularity assumption is imposed.

Assumption 1.1
The cdf F is Lipschitz continuous with constant $C_p < \infty$, i.e., for any $x, y \in \mathbb{R}$ it holds $|F(x) - F(y)| \leq C_p|x - y|$.

Failure Probability. For a certain realization of the input data, determined by a sample $\omega \in \Omega$, the modeled system $\mathcal{M}$ fails, if the corresponding value of the QoI X falls below a certain threshold, referred to as ξ. Exceeding some threshold can be treated completely analogously by negating both QoI and threshold. Since in practice failure happens rather seldom, the union of all these samples is named rare event R_ξ. Its probability with respect to the measure $\mathbb{P}$ is called failure or rare event probability P_ξ, in formulas

$$P_\xi = \mathbb{P}(R_\xi) = \mathbb{P}(X \leq \xi). \tag{1.5}$$

Defining the failure probability functional formally as

$$Q_\xi : \mathcal{S} \to \{0, 1\} \quad \text{with} \quad u \mapsto Q_\xi(u) = \mathbb{1}_{\{X(u) \leq \xi\}}, \tag{1.6}$$

the probability in equation (1.5) can be expressed as the expectation value of $Q_\xi\big(u(Z)\big)$. Transferring the previous abbreviation also to Q_ξ, meaning that later on, again, if not explicitly mentioned differently, Q_ξ is written instead of $Q_\xi\big(u(Z)\big)$, one obtains that Q_ξ becomes a $\{0, 1\}$-valued random variable and that the above observation can be written as $P_\xi = \mathbb{E}\big[Q_\xi\big(u(Z)\big)\big] = \mathbb{E}Q_\xi$, where the last equality just takes advantage of the shorthand notation.
For the groundwater flow problem short travel times from the leak of the repository to the boundary of some safety zone are critical, since in these cases the nuclides are still noxious and there is no possibility to take countermeasures to prevent a disaster. Thus ξ could be selected in a way to ensure that most radionuclides have decayed or that there is enough time to inform the population.

Parametrized Functionals. Additionally to the two functionals defined in the previous paragraphs two further quantities are introduced in the following, strongly aligned to the preceded ones, but totally deterministic. Therefore they are also referred to as the parametric versions of the QoI and the failure probability functional, respectively.
The first one parametrizes the composition of the solving procedure with the consecutive evaluation of the QoI X and thus produces the map

$$f : \Theta \to \mathbb{R} \quad \text{with} \quad z \mapsto f(z) = X\big(u(z)\big). \tag{1.7}$$

Analogously it is proceeded with the failure probability functional Q_ξ resulting in a function, which returns to each parameter $z \in \Theta$ whether failure occurs or not. More precisely the map

$$I_\xi : \Theta \to \{0, 1\} \quad \text{with} \quad z \mapsto I_\xi(z) = \mathbb{1}_{\{f(z) \leq \xi\}} \tag{1.8}$$

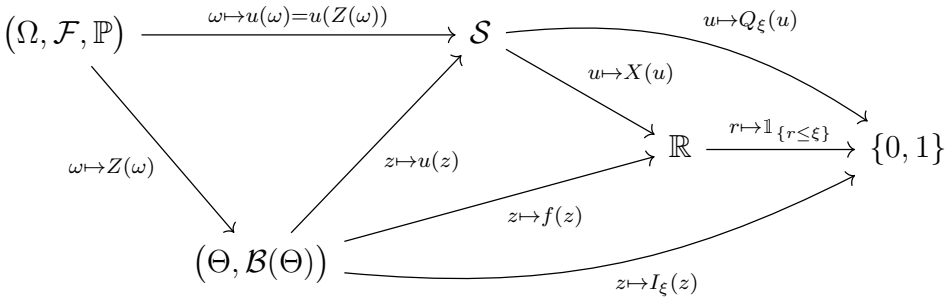
arises.

Clearly, there is an obvious relation between the parametrized versions and the standard functionals, namely, exploiting the shorthand notation: $X = f \circ Z$ and $Q_\xi = I_\xi \circ Z$. For the sake of clarity no abbreviated form is introduced for these quantities.
The last analogy to mention is that the failure probability can be written as

$$\begin{aligned} P_\xi &= \mathbb{E}\big[Q_\xi\big(u(Z)\big)\big] = \int_\Omega Q_\xi\big(u\big(Z(\omega)\big)\big)\,\mathrm{d}\mathbb{P}(\omega) = \int_\Theta Q_\xi\big(u(z)\big)\,\mathrm{d}\big(Z\#\mathbb{P}\big)(z) \\ &= \int_\Theta Q_\xi\big(u(z)\big)p(z)\,\mathrm{d}\lambda(z) = \int_\Theta \big(Q_\xi \circ u\big)(z)p(z)\,\mathrm{d}\lambda(z) = \int_\Theta I_\xi(z)p(z)\,\mathrm{d}\lambda(z), \end{aligned} \tag{1.9}$$

where, besides definitions, the first line applies the change-of-measure formula and from first to second line $Z\#\mathbb{P} \ll \lambda$ is exploited.

Overview. An overview of the connection of all relevant maps defined previously is given in the following commutative diagram.



1.2 Random Fields

Where PDEs use deterministic functions to describe some data, in the context of stochastic partial differential equations (SPDEs) it is desirable to consider function-valued random variables, called random fields, for modeling uncertainty in the coefficients or the solution. The following definition follows a different perspective using an infinite number of random variables taking less abstract values. But it can be shown that indeed both definitions are equivalent.

Definition 1.2 (Random Field)
Let $D \subset \mathbb{R}^d$ be a set, $\big(\Omega, \mathcal{F}, \mathbb{P}\big)$ a probability space and $\big(E, \mathcal{E}\big)$ a measurable space. An E-valued random field a is a mapping

$$a : D \times \Omega \to E$$

such that $a(x, \cdot)$ is $\mathcal{F}$-$\mathcal{E}$-measurable for each $x \in D$.
The function $a(\cdot, \omega) : D \to E$ for some fixed $\omega \in \Omega$ is called realization of a.

Gaussian random fields are an often used special case of random fields, amongst others, because their occurrence is natural and most nice properties of normally distributed random variables are transferred in some way, e.g., that they are uniquely determined by a mean and covariance function. Unfortunately they are unsuitable for the purpose of modeling some real world parameters like permeability, since negative values are attained with strictly positive probability. As a remedy often log-normal random fields are used.

In order to sample from a random field a convenient representation is necessary. To this end the Karhunen-Loève expansion is a helpful tool, providing a Fourier series type expansion. For a second order random field a one can obtain the form

$$a(x,\omega) = \mu(x) + \sum_{i\in\mathbb{N}} \sqrt{\nu_i}\phi_i(x)\theta_i(\omega),$$

where $\mu(x)$ is the mean function, $\big(\nu_i, \phi_i(x)\big)$ are the eigenpairs of the covariance operator

$$\big(\mathcal{C}f\big)(x) = \int_D c(x,y)f(y)\,\mathrm{d}\lambda(y)$$

with covariance function c and where $\theta_i(\omega) = 1/\sqrt{\nu_i}\big\langle a(x,\omega)-\mu(x), \phi_i(x)\big\rangle_{L^2(D)}$ are uncorrelated random variables with mean 0 and variance 1. If a is Gaussian, then $\theta_i \sim \mathcal{N}(0,1)$. In numerical praxis the latter series is truncated after finitely many terms. The resulting finite family of random variables θ_i can be summarized in a random vector Z, parametrizing the model's randomness.
For full definitions, derivations, theorems and proofs it is referred to [LPS14, Ch. 7].

1.3 Numerical Discretization of the Model

Besides the already addressed stochastic approximation of the SPDE there is also need for a spatial discretization. This procedure is considered only in an abstract manner here. For this purpose $\ell = 0,\dots,L$ denotes a hierarchy of levels, where $\ell = 0$ is the coarsest and $\ell = L$ the finest level; the larger the value of ℓ, the better the approximation quality. Correspondingly the discretized models are labeled with $\mathcal{M}_\ell$, yielding solutions u_ℓ and affecting for the shorthand notation, that the QoI X and the failure probability functional Q_ξ both admit a family of functionals

$$\{X_\ell\}_{\ell=0}^L \quad\text{and}\quad \{Q_{\xi,\ell}\}_{\ell=0}^L, \tag{1.10}$$

respectively. A discretization of the QoI X itself might be included in X_ℓ as well.
Analogously also for the parametric version of the QoI f and of the failure probability functional I_ξ families

$$\{f_\ell\}_{\ell=0}^L \quad\text{and}\quad \{I_{\xi,\ell}\}_{\ell=0}^L \tag{1.11}$$

result. Their definition follows the natural way. Lastly also for the rare event R_ξ as well as its probability P_ξ discretized versions $R_{\xi,\ell}$ and $P_{\xi,\ell}$ arise. On the finest level it is assumed that $P_{\xi,L}$ approximates P_ξ sufficiently well for the particular purpose.
A typical and well-established way to obtain such approximations is the finite element method, but also different numerical techniques might be necessary or possible.
One regularity assumption on the discretizations presupposed throughout the thesis is an extension of the already demanded Lipschitz continuity of the cdf F.

Assumption 1.3
Let $F_{q,\ell}$ denote the cdf of the random variable $f_\ell(Z)$, where $Z \sim q\lambda$. For all discretized versions and for any considered distribution of Z, $F_{q,\ell}$ is Lipschitz continuous in a uniform way, i.e., there exists one $C < \infty$ such that for any level ℓ and any considered distribution $q\lambda$, $F_{q,\ell}$ has a Lipschitz constant smaller than C.

Chapter 2

Standard Monte Carlo and Importance Sampling

Based on the assumption that one has a black-box solver for the evaluation or at least approximation of one realization of the failure probability functional Q_ξ at hand, this chapter starts with introducing a very simple and natural idea, statistically motivated, to achieve the value of the failure probability, the so called crude Monte Carlo method [RC04]. A quick analysis of a measure for the relative error directly exposes the weakness of this method in the case of rare event estimation and thus justifies the need of elaborated variance reduction techniques overcoming this problem.
The objective of the remedy presented in the following is to focus on samples with a high impact on the estimated quantity. That is where importance sampling has its name from, since such samples are considered more important than others [Owe13, Chapter 9]. By sampling from a suitable biasing density, which prefers relevant samples, the variance can be reduced. However, finding such a density is a challenging task and crucial for convergence. An iterative methodology to approximate the in some sense optimal biasing density among a family of densities is provided by the cross-entropy method [dBKMR05]. In this chapter the estimators are constructed in an abstract sense, meaning that their spatial discretization is not the main focus.

2.1 Crude Monte Carlo Sampling

Starting from the integral representation

$$P_\xi = \mathbb{E}Q_\xi = \int_\Omega Q_\xi(\omega)\,\mathrm{d}\mathbb{P}(\omega) \tag{2.1}$$

an application of a quadrature formula with N randomly distributed points according to the probability measure $\mathbb{P}$, equally weighted, directly results in the crude Monte Carlo (MC) estimate, whose estimator version is given below.

Definition 2.1 (Crude Monte Carlo Estimator)
Let $\{Q_\xi^i\}_{i=1}^N$ be independent and identically distributed (i.i.d.) copies of the random variable Q_ξ. Then

$$\widehat{P}_\xi^{\mathrm{MC}} = \frac{1}{N}\sum_{i=1}^{N} Q_\xi^i \tag{2.2}$$

defines the crude (or standard) Monte Carlo estimator for the failure probability P_ξ.

If $Q_\xi \in L^1$ the strong law of large numbers guarantees convergence $\mathbb{P}$-a.s. for $N \to \infty$. Furthermore, as $\mathbb{E}\widehat{P}_\xi^{\mathrm{MC}} = P_\xi$, unbiasedness follows immediately. Of course, this estimator is not feasible in practice, as the failure probability functional Q_ξ is not accessible. Replacing it by $Q_{\xi,\ell}$ for some level ℓ provides an unbiased estimator for $P_{\xi,\ell} := \mathbb{E}Q_{\xi,\ell}$, but in general not for P_ξ. This behavior, being incapable of influencing the approximation error $|P_{\xi,\ell} - P_\xi|$, as long as the discretization level is fixed, is typical for numerical estimators. Having assumed, that the latter error decreases as ℓ increases and that $P_{\xi,L} \approx P_\xi$ satisfactorily well, all estimators in this chapter can be thought of being executed on the finest level L in practice; i.e., in this case the estimator

$$\widehat{P}_{\xi,L}^{\mathrm{MC}} = \frac{1}{N}\sum_{i=1}^{N} Q_{\xi,L}^i \tag{2.3}$$

for i.i.d. copies $\{Q_{\xi,L}^i\}_{i=1}^N$ of $Q_{\xi,L}$ approximates P_ξ.
A common measure for the relative error of an estimator in the setting of rare events is the so-called coefficient of variation CV, the fraction between standard deviation and expectation of the estimator and therefore also known as the relative standard deviation. The subsequent proposition states the already mentioned weakness in more detail.

Proposition 2.2
To obtain a given precision $\epsilon > 0$ *for the coefficient of variation, i.e.,* $\mathrm{CV}\big(\widehat{P}_\xi^{\mathrm{MC}}\big) < \epsilon$, *the number of necessary copies* N *depends on the inverse of the failure probability* P_ξ.

Proof. Since $\widehat{P}_\xi^{\mathrm{MC}}$ is an unbiased estimator for P_ξ and Q_ξ is a Bernoulli random variable, one can directly calculate, using that the copies are i.i.d.:

$$\mathrm{CV}\big(\widehat{P}_\xi^{\mathrm{MC}}\big) = \frac{\sqrt{\mathrm{Var}\big(\widehat{P}_\xi^{\mathrm{MC}}\big)}}{\mathbb{E}\widehat{P}_\xi^{\mathrm{MC}}} = \frac{\sqrt{\frac{1}{N^2}\sum_{i=1}^{N} P_\xi(1-P_\xi)}}{P_\xi} = \frac{\sqrt{\frac{1}{N}P_\xi(1-P_\xi)}}{P_\xi} = \frac{1}{\sqrt{N}}\sqrt{\frac{1-P_\xi}{P_\xi}}$$

In order to achieve the given tolerance ϵ it has to hold $N > \frac{1-P_\xi}{\epsilon^2 P_\xi} = \epsilon^{-2}(P_\xi^{-1} - 1)$. □

To conclude this section and emphasize the previous proposition a simple numerical example, adopted from [EHM16], is presented and will also be revisited frequently.
An important note regarding this example series: For illustrative purposes the numerical calculations and the demonstrating figures are performed for different values of P_ξ, *namely* 10^{-4} *and* 10^{-2}, *respectively.*

Example 2.3 (Tail Estimate of the Normal Distribution – First Part)
The aim of this example is the point evaluation of the cdf of the standard normal distribution such that this value becomes $P_{\xi_{\mathrm{calc}}} = 10^{-4}$ for the numerical calculation and $P_{\xi_{\mathrm{plot}}} = 10^{-2}$ for the plot. The associated thresholds are approximately $\xi_{\mathrm{calc}} = -3.719$ and $\xi_{\mathrm{plot}} = -2.326$ for the two cases. Figure 2.1 depicts the situation; the area of the failure probability is shaded gray and referred to as failure region, the threshold ξ_{plot} is marked with a green star.
In order to model this particular situation, let $X \sim \mathcal{N}(0,1)$ be the QoI and define the discretized versions on level ℓ as perturbations $X_\ell = X + \gamma^{4+\ell}(U - \frac{1}{2})$ for $\ell \in \{0,\dots,L\}$ with $L = 3$, a refinement parameter $\gamma = \frac{1}{2}$ and $U \sim \mathcal{U}(0,1)$. Although the calculations

are equally expensive on all levels, the evaluations are considered more expensive the higher the level. Throughout this chapter just approximations on the finest level L are considered.

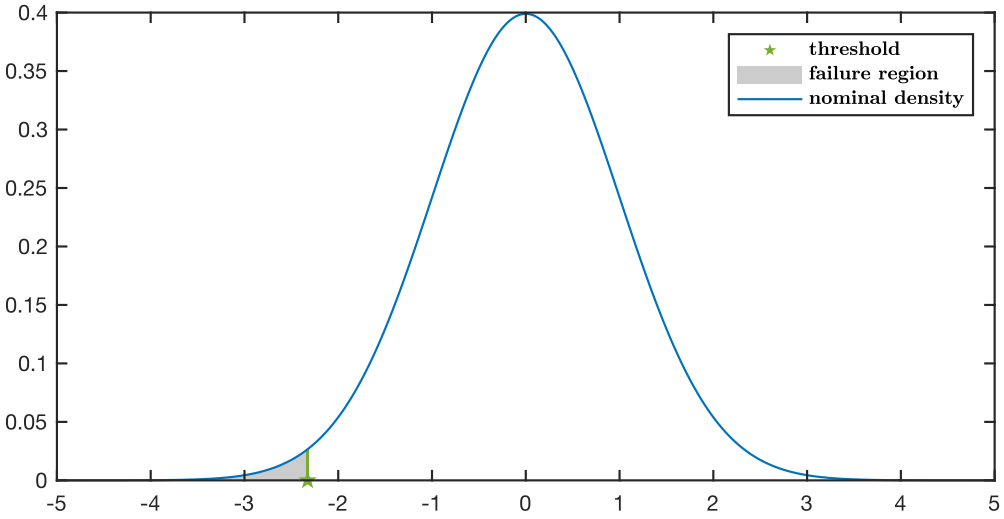


Figure 2.1: Tail estimate of the normal distribution (Example 2.3): Illustration of the initial situation.

Numerical testing and documentation are done as follows: For the four different sample sizes $N \in \{10^1, 10^2, 10^3, 10^4\}$ at first four results of the estimated failure probability are given, before the CV is evaluated as an error measure. Independently of N the estimation of the variance in the CV is based on 10^4 runs of the respective estimator of sample size N. This evaluation is repeated five times in total, the median is listed and the largest distance to all other values is given in brackets including its direction. The results are summarized in Table 2.1 below.

N	**Run 1**	**Run 2**	**Run 3**	**Run 4**	**estimated** CV based on 10^4 runs
10^1	0	0	0	0	29.9867 (-5.4951)
10^2	0	0	0	0	9.9025 (-0.4466)
10^3	0	0	$1 \cdot 10^{-3}$	0	3.1868 (-0.0493)
10^4	$1 \cdot 10^{-4}$	$2 \cdot 10^{-4}$	0	0	0.9965 ($+0.0080$)

Table 2.1: Tail estimate of the normal distribution (Example 2.3): Results of the MC estimation.

These rather bad results are not very surprising. For verification the coefficient of variation shall be checked theoretically for $N = 10^3$:

$$\mathrm{CV}\big(\widehat{P}_{\xi,L}^{\mathrm{MC}}\big) = \frac{1}{\sqrt{N}}\sqrt{\frac{1-P_\xi}{P_\xi}} = \frac{1}{\sqrt{10^3}}\sqrt{\frac{1-10^{-4}}{10^{-4}}} \approx 3.1621.$$

A brief résumé of crude Monte Carlo to conclude this section: For small values P_ξ and an expensive to evaluate functional the method gets quickly computationally infeasible.

2.2 Importance Sampling

Rare event estimation is a difficult task for crude Monte Carlo, but importance sampling (IS) provides a remedy to overcome the difficulties the straightforward MC method faces. Making use of the splitting

$$P_\xi = \mathbb{P}(R_\xi) = \int_{R_\xi} 1 \,\mathrm{d}\mathbb{P} + \int_{R_\xi^c} 0 \,\mathrm{d}\mathbb{P} \tag{2.4}$$

the problem gets visible. Outside of R_ξ the integrand is zero and thus evidently uninteresting. On the other hand, since R_ξ has a small volume with respect to the measure $\mathbb{P}$, it is likely that MC fails having at least one sample inside the much more interesting region R_ξ and thus returns a totally useless result. Intuitively there is a need for getting most or at least more samples from the relevant set R_ξ. Importance sampling does this by generating samples from a problem specific biasing density, which concentrates on R_ξ. In order to ensure unbiasedness of the resulting estimator the integrand is adjusted to compensate the sampling from the different distribution.
Of course, this idea can be also transferred to more complicated and more variable integrands, having rather small values on R_ξ^c and significant values on R_ξ.

2.2.1 Basic Idea

The parametrized integral representation

$$P_\xi = \int_\Theta I_\xi(z)p(z) \,\mathrm{d}\lambda(z) = \int_\Omega I_\xi\big(Z(\omega)\big) \,\mathrm{d}\mathbb{P}(\omega) = \mathbb{E}_p\big[I_\xi(Z)\big], \tag{2.5}$$

where $\mathbb{E}_p$ emphasizes that the expectation is taken with respect to the nominal density p, i.e., $Z \sim p\lambda$, is the initial point for the upcoming considerations. This notation carries over to Var and $\mathbb{P}$.
Now let q be a further probability density on $\mathbb{R}^k$ with $\mathrm{supp}(I_\xi \cdot p) \subset \mathrm{supp}(q)$; then the fundamental equality of importance sampling reads

$$\mathbb{E}_p\big[I_\xi(Z)\big] = \int_\Theta I_\xi(z)p(z) \,\mathrm{d}\lambda(z) = \int_\Theta I_\xi(z)\frac{p(z)}{q(z)}q(z) \,\mathrm{d}\lambda(z) = \mathbb{E}_q\left[I_\xi(Z)\frac{p(Z)}{q(Z)}\right], \tag{2.6}$$

where $\mathbb{E}_q$ indicates that the expectation is now taken with respect to the so-called importance density q, i.e., $Z \sim q\lambda$. As one can see, the multiplicative modification of the integrand in (2.6) with the likelihood ratio p/q counterbalances the sampling of Z from the biasing density q instead of p. That the expression is well-defined, can be verified by a short calculation, see, e.g., [Owe13, Chapter 9] and is due to the prerequisites on q.
Approximating the last expression in (2.6) with MC gives the importance sampling estimator, assuming that all evaluations are well-defined.

Definition 2.4 (Importance Sampling Estimator)
Let $\{Z^i\}_{i=1}^N$ be i.i.d. copies of the random variable Z distributed according to the probability measure $q\lambda$. Then

$$\widehat{P}_\xi^{\mathrm{IS}} = \frac{1}{N}\sum_{i=1}^N I_\xi(Z^i)\frac{p(Z^i)}{q(Z^i)} \tag{2.7}$$

defines the importance sampling estimator with respect to the biasing density q for the failure probability P_ξ.

The next proposition establishes the basis for variance reduction.

Proposition 2.5
The importance sampling estimator $\widehat{P}_\xi^{\mathrm{IS}}$ has the following properties:

i) It is unbiased for P_ξ with variance

$$\mathrm{Var}\big[\widehat{P}_\xi^{\mathrm{IS}}\big] = \frac{1}{N}\mathrm{Var}_q\left[I_\xi(Z)\frac{p(Z)}{q(Z)}\right]. \tag{2.8}$$

ii) If the variance in equation (2.8) is finite, then convergence $\mathbb{P}$-a.s. is guaranteed for $N \to \infty$.

iii) There is a density q_* minimizing the variance, namely

$$q_*(z) = I_\xi(z)\frac{p(z)}{P_\xi}. \tag{2.9}$$

iv) The estimator with importance density q_* from (2.9) has zero variance.

Proof. The proof can be found in Appendix A.3. □

Remark 2.6
The zero variance property of the special estimator addressed in Proposition 2.5 *iv)* ensures that sampling Z according to $q_*\lambda$ would result in an estimator who gives the correct failure probability with just one single realization.

Now that the optimal importance density is analytically known, it is possible to analyze its behavior. Clearly, sampling from $q_*\lambda$ is not possible, since the unknown quantity P_ξ is involved as well as the parametric version of the rare event R_ξ, i.e., $Z(R_\xi)$, but nonetheless properties of good and moreover usable importance densities can be derived. So, q should imitate the peaks of $I_\xi \cdot p$ and should be small, where $I_\xi \cdot p$ is small.
Choosing good importance densities is a severe task and requires sound guessing, mostly paired with numerical help and involving more or less consolidated knowledge of the failure region.
Before the next part provides the tools and intuition for introducing a suitable iterative numerical algorithm to this end, Example 2.3 is continued.

Example 2.7 (Tail estimate of the normal distribution – Second Part)
It is assumed, that an apt importance density is given, pictured orange in Figure 2.2, which approximates the optimal density, displayed green, and additionally permits sampling at low cost.
The numerical tests and the documentation are done in the same way as before and the results are aggregated in Table 2.2. The importance density is a Gaussian density with a mean of approximately -3.9578 and a variance of approximately $2.0420 \cdot 10^{-1}$, which has been derived using the upcoming cross-entropy method from Subsection 2.2.3. More precisely, 10^3 samples have been used during the algorithm to find an in some sense optimal biasing density over a family of Gaussian densities with free mean and a variance bounded from below by 10^{-1} for the numerical calculation and $5 \cdot 10^{-2}$ for the plot.

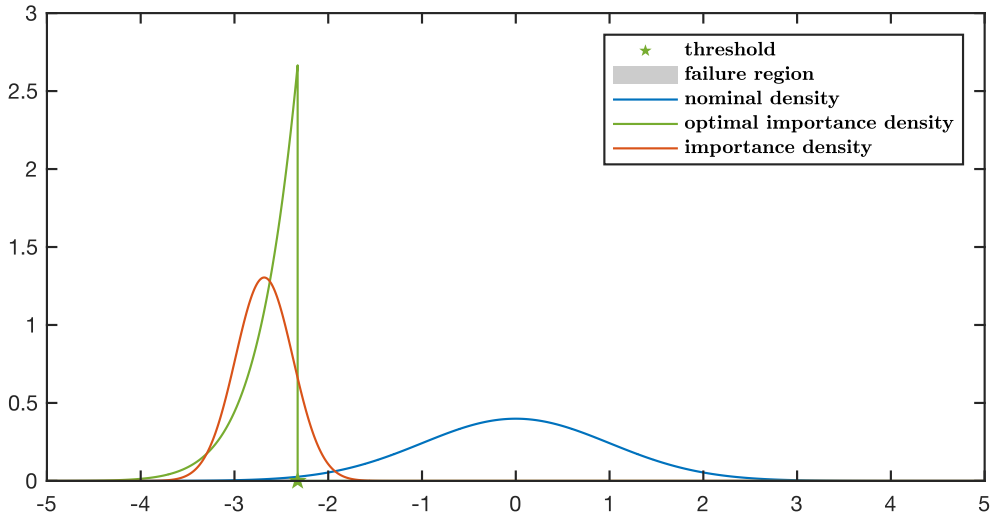


Figure 2.2: Tail estimate of the normal distribution (Example 2.7): Importance sampling using a Gaussian density with mean -2.6829 and variance $9.3532 \cdot 10^{-2}$.

N	**Run 1**	**Run 2**	**Run 3**	**Run 4**	**estimated** CV based on 10^4 runs
10^1	$8.2395 \cdot 10^{-5}$	$9.8767 \cdot 10^{-5}$	$1.2820 \cdot 10^{-4}$	$5.7214 \cdot 10^{-5}$	0.3869 (-0.0076)
10^2	$9.8124 \cdot 10^{-5}$	$8.9840 \cdot 10^{-5}$	$9.0429 \cdot 10^{-5}$	$8.7558 \cdot 10^{-5}$	0.1219 ($+0.0021$)
10^3	$1.0610 \cdot 10^{-4}$	$1.0111 \cdot 10^{-4}$	$1.0102 \cdot 10^{-4}$	$9.2658 \cdot 10^{-5}$	0.0385 (-0.0005)
10^4	$9.9692 \cdot 10^{-5}$	$1.0074 \cdot 10^{-4}$	$9.9548 \cdot 10^{-5}$	$9.9708 \cdot 10^{-5}$	0.0121 ($+0.0001$)

Table 2.2: Tail estimate of the normal distribution (Example 2.7): Results of the IS estimation using a Gaussian density with mean -3.9578 and variance $2.0420 \cdot 10^{-1}$.

As a first, somehow expectable observation it has to be mentioned that the results are far better than the respective ones with standard Monte Carlo, but, and this point has to be emphasized, a suitable importance density has been given. In general this is, of course, not the case and the construction of such a biasing density is not for free. Secondly, the decay of the CV within N follows neatly the theoretically expected one of $\sqrt{N}$, which is typical for Monte Carlo like approaches. The third point, which shall be discussed now tackles the severe restriction to the variance and explains its necessity as well as a methodology to weaken it. Therefore the variance of the estimators is analyzed; the value of the threshold $\xi \in \mathbb{R}$ plays no role in the following considerations. Using Steiner's translation theorem and $I^2_{\xi,L}(Z) = I_{\xi,L}(Z)$ yields

$$\operatorname{Var}\left[\widehat{P}^{\mathrm{IS}}_{\xi,L}\right] = \frac{1}{N}\left(\mathbb{E}_q\left[I^2_{\xi,L}(Z)\frac{p^2(Z)}{q^2(Z)}\right] - P^2_{\xi,L}\right) = \frac{1}{N}\left(\mathbb{E}_p\left[I_{\xi,L}(Z)\frac{p(Z)}{q(Z)}\right] - P^2_{\xi,L}\right).$$

For the remainder it suffices to consider $\mathbb{E}_p\left[I_{\xi,L}(Z)\frac{p(Z)}{q(Z)}\right]$. A short calculation shows

$$\mathbb{E}_p\left[I_{\xi,L}(Z)\frac{p(Z)}{q(Z)}\right] = \int_\Theta I_{\xi,L}(z)\frac{p^2(z)}{q(z)}\,\mathrm{d}\lambda(z) = \int_{-\infty}^{\xi}\frac{\sigma}{\sqrt{2\pi}}e^{-z^2+\frac{(z-\mu)^2}{2\sigma^2}}\,\mathrm{d}\lambda(z),$$

where in the latter equality the failure domain and the densities are inserted and the parameters for the importance density are abbreviated by μ and σ^2. Lemma A.5 in Ap-

pendix A.3 shows that this integral and therefore the variance is unbounded for $\sigma < 1/\sqrt{2}$, and thus also for the density used above. This indicates that small σ's have to be handled very carefully and implausible outcomes could result thereof and require a check, see, e.g., [Rob15] for a discussion on estimators with infinite variance.
A remedy motivated from a theoretical point of view, tackling the origin of the unboundedness, the light tails of the normal distribution q, is to use a corresponding Student's t distribution $\tilde{q}\lambda$ instead, because of its heavier tails. The selection is heuristic; a first freedom is to choose the Student's t degree of freedom ν, specifying how close it shall be to a normal distribution; a second freedom is to decide, whether the variances or covariance matrices in more dimensions of q and $\tilde{q}$ shall coincide or if the shape parameter or shape matrix of $\tilde{q}$ shall be the same as the one of q interpreted as Student's t distribution. It can indeed be shown (see Lemma A.7 in Appendix A.3), that the resulting variance is bounded. Similarly as before one analyzes

$$\mathbb{E}_p\left[I_{\xi,L}(Z)\frac{p(Z)}{\tilde{q}(Z)}\right] = \int_\Theta I_{\xi,L}(z)\frac{p^2(z)}{\tilde{q}(z)}\,\mathrm{d}\lambda(z) = \int_{-\infty}^{\xi} \alpha e^{-z^2}\left(1+\frac{(z-\mu)^2}{\nu\delta^2}\right)^{\frac{\nu+1}{2}}\mathrm{d}\lambda(z),$$

where $\nu > 1$ denotes the Student's t degree of freedom, μ the mean and δ^2 the shape parameter of the importance density. α is just a prefactor with $\alpha = \left(\Gamma(\frac{\nu}{2})\sqrt{\nu\delta^2}\right)/\left(2\sqrt{\pi}\Gamma(\frac{\nu+1}{2})\right)$. By dropping the strong restriction on the variance and substituting it by a weakened one, e.g., a lower bound of only 10^{-2}, this idea is numerically examined. Estimates of the CV are documented in the usual way in Table 2.3. The experiment is done one time for the original way using normal sampling and two times for the Student's t variant for two different degrees of freedom $\nu \in \{13, 3\}$, using the approach of keeping the shape parameter. The parameters of the normal distribution (beyond) are -3.8627 for the mean and $1.1485 \cdot 10^{-2}$ for the variance.

N	**estimated** CV **for normal sampling** based on 10^4 runs	**estimated** CV **for Student's t sampling** ν=13, based on 10^4 runs	**estimated** CV **for Student's t sampling** ν=3, based on 10^4 runs
10^1	1.4355 (+0.6478)	0.2070 (−0.0010)	0.2984 (+0.0044)
10^2	0.8618 (+0.5329)	0.1002 (+0.0011)	0.0951 (−0.0007)
10^3	0.6152 (+0.3395)	0.0828 (+0.0006)	0.0302 (+0.0004)
10^4	0.4309 (+0.2687)	0.0807 (−0.0001)	0.0095 (−0.0001)

Table 2.3: Tail estimate of the normal distribution (Example 2.7): Comparison of coefficient of variation for normal and Student's t sampling. The Gaussian (reference) density has mean -3.8627 and variance $1.1485 \cdot 10^{-2}$

Before the results of the Student's t variant are investigated, the contrast between the last column of Table 2.2 and the first of Table 2.3 emphasizes the need of the strict variance restriction when using normal sampling. Taking into account the remaining two columns shows the improvement of the CV when using the Student's t variant. The results for $\nu = 3$ show a correct decay with $\sqrt{N}$, are even a bit better than the ones from the previous table and seem to be quite stable. The middle column ($\nu = 13$) shows clearly better results than normal sampling, but rather bad values (except for small N) compared to the smaller degree of freedom, which corresponds to heavier tails. This is reasonable as the Student's t distribution converges to the corresponding normal distribution as $\nu \to \infty$.

Once again a brief résumé at the end of this subsection: Importance sampling makes the estimation of small values P_ξ cheaper or for really small values in the first place possible. But the knowledge of a suitable biasing density is inevitable and its quality is crucial for the methodology to work. In fact, an inappropriate biasing density can make the problem even worse. The big challenge consequently is to find an importance density which is firstly in some sense close to the optimal one and secondly arises from a feasible numerical algorithm. Both requirements are met by the method, which is presented in Subsection 2.2.3 later on.

2.2.2 Information Theory

At first it is necessary to make the term of the distance between two probability densities more precise. Therefore the Kullback-Leibler divergence, also known as relative entropy, is introduced, a quantity, who has its origin in information theory [Kul59] and is omnipresent in several interdisciplinary topics like neuroscience, see, e.g., [Pan03], applied statistics, see, e.g., [AG07], artificial intelligence and machine learning, see, e.g., [Wol18].
Even though it would suffice to accept the upcoming Definition 2.8 of the Kullback-Leibler divergence for the future purposes and see it, loosely speaking, as some quantity describing how two measures differ from one another, a few more words shall be spent on the intuition behind by giving a glimpse into information theory [SW49].
The focus of attention in the field of information theory is the quantification and transmission of information. Therefore it is necessary to clarify the terminology of information and especially separate it from the resonating notions of meaning and semantics. Abstractly speaking, the information in some message is the freedom to choose the single characters at will, or in other words, citing Weaver, "this word information [...] relates not so much to what [one does] say, as to what [one] could say" [SW49, p. 8]. To make this more rigorous, Shannon, the founder of this field, deduced a measure to specify the information diversity [Sha48]. Tellingly for a message m it is called self-information or surprisal $I(m)$ and defined by

$$I(m) = \log\left(\frac{1}{\varrho(m)}\right), \tag{2.10}$$

where $\varrho(m)$ denotes the probability that m is the message of choice among all imaginable messages. As the event space M, the set of all possible messages, is typically finite, the power set can be chosen as σ-algebra and ϱ is a discrete probability measure. That the formula above is the natural way of quantifying information is outlined in Lemma A.8 in Appendix A.3. As this is quite formal, the intuitive way to think of $I(m)$ is to see this quantity as the (minimum) number of characters needed to encode the message, e.g., the amount of necessary bits, when using $\log_2$ in equation (2.10).
Based thereon let M denote the specific discrete random variable from some underlying probability space with values in M, such that the pushforward measure turns out to have density ϱ. Then the entropy of the random variable M, a quantity associated with the uncertainty of its values, is given by

$$\mathrm{H}(\mathsf{M}) = \mathbb{E}\big[I(\mathsf{M})\big] = \sum_{m_i \in M} I(m_i)\varrho(m_i) = \sum_{m_i \in M} \log\left(\frac{1}{\varrho(m_i)}\right)\varrho(m_i). \tag{2.11}$$

With this construction $\mathrm{H}(\mathsf{M})$ can be thought of as the expected (minimum) number of characters needed to encode the random variable M.
Consider at this stage the example of throwing a dice. It is easy to verify, using, e.g., the Lagrange function for the corresponding constrained optimization problem, that $\mathrm{H}(\mathsf{M})$ is maximized if ϱ is a uniform distribution. Conversely the entropy can be made arbitrarily small if one of the six numbers appears with probability tending to 1.
Finally, having quantified information the focus is turned towards the main concern of messages, namely the transmission and therefore the propagation of information. In the variety of theory, tackling, e.g., optimal encoding, error-correction, quantum information and many more, just one thought experiment shall be carried out: In order to send messages, it is desirable to compress them beforehand. Assume that, if the density ϱ is available, some message can be compressed in an optimal way such that this coincides with its information content. As the true ϱ is typically unknown, some other density $\tilde{\varrho}$ is assumed. This entails information loss. The Kullback-Leibler divergence now measures the expected number of additionally necessary characters for the compression procedure. Mathematically this can be formalized by

$$\mathrm{KL}(\varrho\|\tilde{\varrho}) = \mathbb{E}\left[\log\left(\frac{\varrho(\mathsf{M})}{\tilde{\varrho}(\mathsf{M})}\right)\right] = \sum_{m_i\in M}\log\left(\frac{\varrho(m_i)}{\tilde{\varrho}(m_i)}\right)\varrho(m_i). \tag{2.12}$$

This quantity is also known as relative entropy or information gain, where the second synonym just reverses the perspective from the derivation.
Before the introduced quantities are carried over to their continuous versions, the term cross entropy (CE) shall be defined as well, in order to convince that the name of the method presented in the following Subsection 2.2.3 is justified, although the cross entropy doesn't appear at first glance. As before two densities are present, namely the underlying one, ϱ, and the assumed one, $\tilde{\varrho}$. Similarly to the previous constructions the cross entropy $\mathrm{H}(\mathsf{M};\varrho,\tilde{\varrho})$ averages the number of characters necessary to encode M assuming $\tilde{\varrho}$, i.e.,

$$\mathrm{H}(\mathsf{M};\varrho,\tilde{\varrho}) = \sum_{m_i\in M}\log\left(\frac{1}{\tilde{\varrho}(m_i)}\right)\varrho(m_i). \tag{2.13}$$

Intuitively $\mathrm{H}(\mathsf{M};\varrho,\tilde{\varrho}) = \mathrm{H}(\mathsf{M}) + \mathrm{KL}(\varrho\|\tilde{\varrho})$ has to hold.
Two last comments to conclude this excursion to information theory and to lead over to the cross-entropy method: Firstly, the last equation showing the connection of the defined terms, encourages to interpret the Kullback-Leibler divergence as some kind of distance of probability measures and motivates the properties shown for the continuous version in Proposition 2.9. Secondly, it gets evident that the problem of optimizing the Kullback-Leibler divergence for some density $\tilde{\varrho}$ is equivalent to optimizing the cross-entropy for $\tilde{\varrho}$.
With this knowledge of information theory the introduced concept of the Kullback-Leibler divergence can be transferred to the setting of continuous random variables and densities. For convenience the logarithm in the previous definitions is considered to base e from now on.

Definition 2.8 (Kullback-Leibler Divergence, colloquially also termed Cross Entropy)
Let q and $\tilde{q}$ denote two probability densities on the parameter space Θ and let $Z \sim q\lambda$.

Then the Kullback-Leibler (KL) divergence is defined as

$$\mathrm{KL}(q\|\tilde{q}) = \mathbb{E}_q \ln \frac{q(Z)}{\tilde{q}(Z)} = \int_\Theta q(z) \ln \frac{q(z)}{\tilde{q}(z)} \,\mathrm{d}\lambda(z) \tag{2.14}$$

if the induced measure $q\lambda$ is absolutely continuous with respect to $\tilde{q}\lambda$ and as $\mathrm{KL}(q\|\tilde{q}) = \infty$ otherwise.

Strictly speaking the KL divergence is not a distance, since it is not symmetric and does not satisfy the triangle inequality. Although this deficit could be repaired by using a symmetrized and moreover smoothened version, e.g., the Jensen-Shannon divergence, given by $\mathrm{JS}(q\|\tilde{q}) = \frac{1}{2}\big(\mathrm{KL}(q\|\overline{q}) + \mathrm{KL}(\tilde{q}\|\overline{q})\big)$ with $\overline{q} = \frac{1}{2}(q+\tilde{q})$, it is not worthwhile as the successive considerations would complicate tremendously. At least two relevant properties of a metric are met by the KL divergence as the following proposition shows.

Proposition 2.9
The Kullback-Leibler divergence $\mathrm{KL}(q\|\tilde{q})$ between two densities q and $\tilde{q}$ is non-negative and takes the value 0 iff $q = \tilde{q}$ λ-a.e.

Proof. Since $-\ln$ is a convex function on $(0,\infty)$, by Jensen's inequality it holds

$$\begin{aligned}\mathrm{KL}(q\|\tilde{q}) = \mathbb{E}_q \ln \frac{q(Z)}{\tilde{q}(Z)} = \mathbb{E}_q\left[-\ln \frac{\tilde{q}(Z)}{q(Z)}\right] \geq -\ln \mathbb{E}_q \frac{\tilde{q}(Z)}{q(Z)} = -\ln \int_\Theta \frac{\tilde{q}(z)}{q(z)} q(z) \,\mathrm{d}\lambda(z) \\ = -\ln \int_\Theta \tilde{q}(z) \,\mathrm{d}\lambda(z) = -\ln 1 = 0.\end{aligned}$$

Since equality in Jensen's inequality holds for a strictly convex function, as $-\ln$ is, only if the inner integrand, here $\tilde{q}/q$, is constant λ-a.e., it follows $\tilde{q} = c \cdot q$ λ-a.e. for a $c \in \mathbb{R}$. Since q and $\tilde{q}$ are both densities it immediately follows $c = 1$. This completes the proof. □

2.2.3 Cross-Entropy Method

In a nutshell, the aim of this method is to find a density, which is close to the optimal importance density q_* with respect to the Kullback-Leibler divergence (2.14) and hence may be an appropriate biasing density for importance sampling.
To avoid working with an unmanageable number of probability densities a family of pdfs $\mathcal{Q}$ is specified and assumed, that each $q \in \mathcal{Q}$ can be uniquely represented by a finite-dimensional parameter v, leading to the notation q_v. For convenience let $p = q_u \in \mathcal{Q}$. As said, the goal is to construct a density $q_{v_*} \in \mathcal{Q}$, referred to as optimal CE density, such that

$$q_{v_*} = \operatorname*{arg\,min}_{q_v \in \mathcal{Q}} \mathrm{KL}(q_*\|q_v). \tag{2.15}$$

In general, q_* and q_{v_*} describe different densities. The minimization (2.15) can be rewritten in the parameter form

$$\begin{aligned} v_* &= \operatorname*{arg\,min}_{v \text{ s.t. } q_v \in \mathcal{Q}} \mathbb{E}_{q_*} \ln \frac{q_*(Z)}{q_v(Z)} = \operatorname*{arg\,min}_{v \text{ s.t. } q_v \in \mathcal{Q}} \mathbb{E}_p\left[\frac{q_*(Z)}{p(Z)} \ln \frac{q_*(Z)}{q_v(Z)}\right] \\ &= \operatorname*{arg\,max}_{v \text{ s.t. } q_v \in \mathcal{Q}} \mathbb{E}_p\left[\frac{q_*(Z)}{p(Z)} \ln q_v(Z)\right] \overset{(2.9)}{=} \operatorname*{arg\,max}_{v \text{ s.t. } q_v \in \mathcal{Q}} \mathbb{E}_p\big[I_\xi(Z) \ln q_v(Z)\big]. \end{aligned} \tag{2.16}$$

Starting from this continuous version, the straightforward way to estimate v_* by a Monte Carlo approach most probably fails due to the rareness of the event $R_\xi = \{f(Z) \leq \xi\}$. Because of this, the cross-entropy Algorithm 1 is iterative and uses repeated importance sampling with biasing densities q_{v_m} constructed along a sequence of nested failure events $R_{\xi_m} = \{f(Z) \leq \xi_m\}$, specified by their intermediate thresholds ξ_m for $m = 1, 2, \dots$ such that

$$\infty > \xi_1 > \xi_2 > \cdots \geq \xi \quad \text{and therefore} \quad \Omega \supset R_{\xi_1} \supset R_{\xi_2} \supset \cdots \supset R_\xi.$$

This idea, admitting a lack of notation when waiving the hat for the estimated density parameters and the estimated thresholds, is outlined more precisely before the algorithm is stated: Instead of finding directly a parameter v_* belonging to the rare event threshold ξ, the point of view is changed and on the basis of the nominal density $p = q_u$ an initial threshold ξ_1 is determined in a way that $\mathbb{P}(R_{\xi_1}) = \mathbb{P}_{q_u}(R_{\xi_1}) \approx \rho$ for a $\rho \in (0, 1)$, which is typically around 10^{-1}. This lifted threshold can be estimated using a fixed number of i.i.d. copies Z^i, $i \in \{1, \dots, N_{\mathrm{CE}}\}$, of Z, distributed according to $Z\#\mathbb{P} = p\lambda$, then calculating the related performances of the parametrized QoI, i.e., $f(Z^i)$ and finding the ρ sample quantile. Afterwards the parameter v_1 of the first biasing density is estimated by solving the so-called stochastic counterpart of

$$v_1 = \operatorname*{arg\,max}_{v \text{ s.t. } q_v \in \mathcal{Q}} \mathbb{E}_p\big[I_{\xi_1}(Z) \ln q_v(Z)\big], \tag{2.17}$$

which is no longer affected by the original rareness of R_ξ and reuses the already performed calculations. It is given by

$$v_1 = \operatorname*{arg\,max}_{v \text{ s.t. } q_v \in \mathcal{Q}} \frac{1}{N_{\mathrm{CE}}} \sum_{i=1}^{N_{\mathrm{CE}}} I_{\xi_1}(Z^i) \ln q_v(Z^i). \tag{2.18}$$

How to practically obtain the estimate v_1 is addressed in connection with the general step: Therefore assume that the intermediate threshold ξ_{m-1} and the previously estimated density parameter v_{m-1} are given. Since $q_{v_{m-1}}$ naturally is a better biasing density than the preceding ones the consecutive threshold ξ_m is chosen such that $\mathbb{P}_{q_{v_{m-1}}}(R_{\xi_m}) \approx \rho$. Similarly to the first step, ξ_m is estimated by evaluating the performances of the parametrized QoI f for i.i.d. copies Z^i of Z, distributed according to $q_{v_{m-1}}\lambda$, and taking the ρ sample quantile. Then the parameter v_m of the m-th density is obtained by solving

$$\begin{aligned} v_m &= \operatorname*{arg\,max}_{v \text{ s.t. } q_v \in \mathcal{Q}} \mathbb{E}_p\big[I_{\xi_m}(Z) \ln q_v(Z)\big] \\ &= \operatorname*{arg\,max}_{v \text{ s.t. } q_v \in \mathcal{Q}} \mathbb{E}_{q_{v_{m-1}}}\left[\big(I_{\xi_m}(Z) \ln q_v(Z)\big) \frac{p(Z)}{q_{v_{m-1}}(Z)}\right], \end{aligned} \tag{2.19}$$

where the importance sampling idea enters. As before the stochastic counterpart reads

$$v_m = \operatorname*{arg\,max}_{v \text{ s.t. } q_v \in \mathcal{Q}} \frac{1}{N_{\mathrm{CE}}} \sum_{i=1}^{N_{\mathrm{CE}}} \big(I_{\xi_m}(Z^i) \ln q_v(Z^i)\big) \frac{p(Z^i)}{q_{v_{m-1}}(Z^i)}. \tag{2.20}$$

As long as the intermediate thresholds have not reached ξ, the algorithm is continued. To avoid an endless loop a minimal stepwidth δ is utilized as shown in the pseudocode in the following.

Algorithm 1 Cross-Entropy Algorithm

Input: Nominal density parameter u, family of pdfs $\mathcal{Q}$, parametric version of QoI f, sample size N_{CE}, threshold ξ, quantile parameter ρ and minimal stepwidth δ.

Output: Parameter v_* determining optimal CE density q_{v_*}.

1: Initialize density q_{v_0} with $v_0 = u$, i.e., $q_{v_0} = p$ and set CE step counter to $m = 1$.
2: **while** 1
3: Generate i.i.d. realizations $\{z^i\}_{i=1}^{N_{\mathrm{CE}}}$ of Z distributed according to $q_{v_{m-1}}\lambda$.
4: Evaluate the QoIs, i.e., compute $\{f(z^i)\}_{i=1}^{N_{\mathrm{CE}}}$.
5: Set the intermediate threshold ξ_m to the minimum of the ρ-quantile of the previous results and, if existent, $\xi_{m-1} - \delta$.
6: **if** $\xi_m > \xi$ **then**
7: Solve for v_m using $\{f(z^i)\}_{i=1}^{N_{\mathrm{CE}}}$ and set $m = m + 1$.
8: **else**
9: Set $\xi_m = \xi$, solve for v_m using $\{f(z^i)\}_{i=1}^{N_{\mathrm{CE}}}$ and **break** with $v_* = v_m$.
10: **end if**
11: **end while**

It remains to discuss the estimation of the density parameter v_m in line 7 and 9 of the algorithm, i.e., solving the stochastic counterpart of the optimization problem. Typically the objective function is differentiable and convex in the parameter v, yielding that v_m is the root of

$$\frac{1}{N_{\mathrm{CE}}} \sum_{i=1}^{N_{\mathrm{CE}}} \left(I_{\xi_m}(Z^i) \nabla_v \ln q_v(Z^i) \right) \frac{p(Z^i)}{q_{v_{m-1}}(Z^i)}. \tag{2.21}$$

Surprisingly, but very advantageously numerical differentiation in $\nabla_v \ln q_v(z)$ can be avoided for a large and relevant class of densities, as it can be carried out analytically there, e.g., for the so called natural exponential family [RK04]. The following example provides the gradient for another class, namely for the family of multivariate normal distributions, (uniquely) parametrized with their mean vector μ and covariance matrix Σ.

Example 2.10
Let $v = (\mu, \ \Sigma)$. A rather lengthy calculation shows that the derivative is given by

$$\nabla_v \ln q_v(z) = \left(\Sigma^{-1}(z-\mu), \ \ -\tfrac{1}{2}\left(\Sigma^{-1} - \Sigma^{-1}(z-\mu)(z-\mu)^T \Sigma^{-1} \right) \right).$$

Having derived the optimal cross-entropy density q_{v_*} via an application of Algorithm 1, it can be used as a biasing density in an importance sampling setting. Typically the theoretical conditions on the importance density, namely that $\mathrm{supp}(I_\xi \cdot p) = \mathrm{supp}(q_*) \subset \mathrm{supp}(q_{v_*})$ holds and sampling from q_{v_*} is possible, are met. This allows to define the importance sampling estimator, which uses the cross-entropy method to construct the biasing density. Preferably, but not necessarily, the number of samples N in the estimator itself is set to the number of samples already used during the derivation of the importance density, i.e., $N = N_{\mathrm{CE}}$.

Definition 2.11 (Cross-Entropy Importance Sampling Estimator)
Let $\{Z^i\}_{i=1}^N$ be i.i.d. copies of the random variable Z distributed according to the optimal

CE density $q_{v_*}\lambda$. Then

$$\widehat{P}_\xi^{\mathrm{CEIS}} = \frac{1}{N}\sum_{i=1}^{N} I_\xi(Z^i)\frac{p(Z^i)}{q_{v_*}(Z^i)} \tag{2.22}$$

defines the cross-entropy importance sampling (CEIS) estimator for the failure probability P_ξ.

As already mentioned for the MC estimator, $\widehat{P}_\xi^{\mathrm{CEIS}}$ is not feasible either, as I_ξ is not available. And completely analogously to before the parametrized failure probability functional has to be substituted by its discretization $I_{\xi,\ell}$ on some level ℓ in order to achieve a realizable estimator, which is then unbiased for $P_{\xi,\ell} = \mathbb{E}\big[I_{\xi,\ell}(Z)\big]$. But unlike MC or even IS with a prespecified density this adjustment has to be extended to the deduction of the density as well. Consequently the derived optimal CE density has to be equipped with a level index, i.e., $q_{v_{\ell,*}}$.
At the end of this subsection once again Example 2.3 is revived.

Example 2.12 (Tail Estimate of the Normal Distribution – Third Part)
In contrast to Example 2.7 the importance density has to be constructed first by means of the cross-entropy algorithm, searching for the optimal CE density in a family of normal densities with free mean and a variance bounded from below by 10^{-1} for the numerical calculation and $5 \cdot 10^{-2}$ for the plot. These restrictions help to bypass degeneracy issues, which are discussed a bit more at the end of Chapter 4. The minimal stepwidth δ is chosen to be 10^{-2} in both cases and ρ is 0.1 for the calculations and 0.15 for the generation of the plot. This procedure is illustrated in Figure 2.3. Each intermediate threshold is marked with an orange star. The final threshold coincides with the green marked ξ. Additionally the sequence of biasing densities arising during the algorithm is depicted, where the chronological order is visualized by increasing opacity for an increasing CE step counter m.

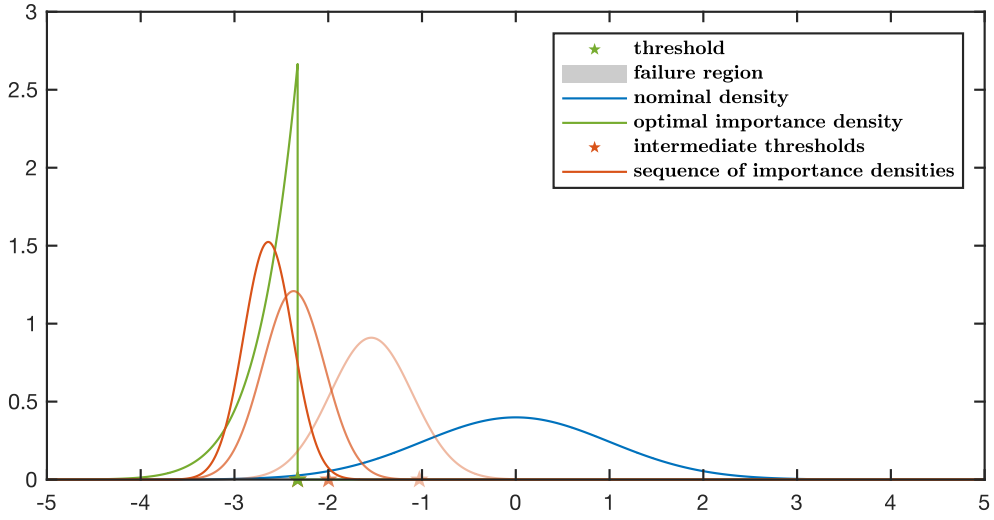


Figure 2.3: Tail estimate of the normal distribution (Example 2.12): Importance sampling utilizing the CE method to obtain an importance density. The parameters used in the CE method are $\rho = 0.15$, $\delta = 10^{-2}$ and $N_{\mathrm{CE}} = 10^3$. $\mathcal{Q}$ is the family of Gaussian densities with variance larger than $5 \cdot 10^{-2}$.

As usual, Table 2.4 presents the results for $N_{\mathrm{CE}} = N$.

N	**Run 1**	**Run 2**	**Run 3**	**Run 4**	**estimated** CV based on 10^4 runs
10^1	$6.7871 \cdot 10^{-5}$	$3.5760 \cdot 10^{-5}$	$1.7094 \cdot 10^{-4}$	$1.1655 \cdot 10^{-4}$	0.6922 (+0.0212)
10^2	$7.1560 \cdot 10^{-5}$	$1.1273 \cdot 10^{-4}$	$9.7037 \cdot 10^{-5}$	$1.1044 \cdot 10^{-4}$	0.1478 (+0.0044)
10^3	$1.0322 \cdot 10^{-4}$	$1.0126 \cdot 10^{-4}$	$9.4426 \cdot 10^{-5}$	$1.0458 \cdot 10^{-4}$	0.0362 (−0.0005)
10^4	$9.9023 \cdot 10^{-5}$	$1.0093 \cdot 10^{-4}$	$9.9554 \cdot 10^{-5}$	$9.9429 \cdot 10^{-5}$	0.0113 (−0.0001)

Table 2.4: Tail estimate of the normal distribution (Example 2.12): Results of the CEIS estimation. The parameters used in the CE method are $\rho = 0.1$ and $\delta = 10^{-2}$. $\mathcal{Q}$ is the family of Gaussian densities with variance larger than 10^{-1}.

Apart from slightly worse results for a relatively small number of samples compared to importance sampling with a predetermined density they are fairly similar to the ones in Table 2.2. The reason for this is quite evident after having a closer look at the difference of how the importance densities used for the failure probability estimate in each case are achieved. First think back to importance sampling with a given importance density. It has been mentioned there, that in fact the cross-entropy method has already been used once in order to obtain the specific density. Since this was a one-time calculation with a fixed number of samples in each step of the cross-entropy algorithm, namely $N = 10^3$, this is clearly different from the setting in this part of the example. Here all intermediate densities and therefore also the optimal CE density are derived based on the same number of samples as the final estimate is. The poorer results in Table 2.4 coincide with the cases, where this number is smaller than in the second part of this example.

In this subsection an iterative methodology has been developed to make importance sampling usable. A suitable importance density can be found among a prespecified parametric family via an optimization problem, which is analytically solvable in some relevant cases and is therefore cheap in terms of computational cost.

Chapter 3

The Multilevel Idea and Selective Refinement

Resuming the aim of the preceding section, this chapter follows a different path to gain variance reduction by exploiting a multilevel idea. Unlike before the spatial discretization is directly taken into account and becomes the focal point. Instead of estimating the failure probability functional on the finest level, it is made use of the linearity of the expectation, making room for writing $Q_{\xi,L}$ in terms of the coarse level approximation $Q_{\xi,0}$ and several correction terms up to the finest level. This technique is known as the multilevel Monte Carlo method and was introduced in the setting of SPDEs by Giles et al. in [CGST11].
Due to a lack of regularity in the failure probability functional Q_ξ the convergence result doesn't benefit from the full potential of the method in that case. In order to counterbalance this Elfverson, Hellmann and Malqvist have developed a method taking advantage of the failure probability functional's special shape [EHM16]. Namely, if from a coarser level approximation of the QoI the fine level value of the failure probability functional is predictable, computations can be saved. This strategy is called selective refinement. Combining multilevel Monte Carlo with selective refinement results in an estimator with better cost asymptotics, which is, loosely speaking, in the end as expensive as solving one realization on the finest level.

3.1 Multilevel Monte Carlo

The method's key idea is to utilize the telescopic sum representation

$$P_{\xi,L} = \mathbb{E}Q_{\xi,L} = \mathbb{E}Q_{\xi,0} + \sum_{\ell=1}^{L} \mathbb{E}[Q_{\xi,\ell} - Q_{\xi,\ell-1}] = \sum_{\ell=0}^{L} \mathbb{E}[Q_{\xi,\ell} - Q_{\xi,\ell-1}] \tag{3.1}$$

with $Q_{\xi,-1} \equiv 0$ and estimate each occurring expectation with MC using judiciously chosen sample sizes N_ℓ for the different correction terms $Q_{\xi,\ell} - Q_{\xi,\ell-1}$. The multilevel Monte Carlo (MLMC) estimator is then defined as follows.

Definition 3.1 (Multilevel Monte Carlo Estimator)
For any level $\ell \in \{0, \dots, L\}$ let $\left\{Q^i_{\xi,\ell} - Q^i_{\xi,\ell-1}\right\}_{i=1}^{N_\ell}$ be i.i.d. copies of the random variable

$Q_{\xi,\ell} - Q_{\xi,\ell-1}$. *Then*

$$\widehat{P}_{\xi,L}^{\mathrm{MLMC}} = \sum_{\ell=0}^{L} \frac{1}{N_\ell} \sum_{i=1}^{N_\ell} \left(Q_{\xi,\ell}^i - Q_{\xi,\ell-1}^i\right) \tag{3.2}$$

defines the multilevel Monte Carlo estimator up to level L for the failure probability $P_{\xi,L}$.

As long as all $Q_{\xi,\ell}$ are in L^1 strong convergence is guaranteed if $N_\ell \to \infty$ for any ℓ. Unbiasedness with respect to $P_{\xi,L}$ follows from the telescopic sum construction.
Before going deeper into the advantages of the MLMC estimator, a standard absolute error measure, the root-mean-square error (RMSE) of the estimator $\widehat{P}_{\xi,L}^{\mathrm{MLMC}}$ shall be defined as follows,

$$\mathrm{RMSE}\big(\widehat{P}_{\xi,L}^{\mathrm{MLMC}}\big) := \sqrt{\mathbb{E}\Big[\big(\widehat{P}_{\xi,L}^{\mathrm{MLMC}} - P_\xi\big)^2\Big]}. \tag{3.3}$$

This measure can be put into relation with the coefficient of variation. Therefore let $\widehat{P}$ denote an unbiased estimator of P. Then it holds

$$\mathrm{RMSE}\big(\widehat{P}\big) = \sqrt{\mathbb{E}\Big[\big(\widehat{P} - P\big)^2\Big]} = \sqrt{\mathrm{Var}\big(\widehat{P}\big)} = \mathbb{E}\widehat{P} \cdot \mathrm{CV}\big(\widehat{P}\big) = P \cdot \mathrm{CV}\big(\widehat{P}\big), \tag{3.4}$$

showing that, unbiasedness of the estimator provided, the CV can be interpreted as the RMSE's relative version.
In order to understand why MLMC is cheaper than MC the mean-square error (MSE) of both shall be calculated and compared. Straightforward computations show

$$\mathrm{MSE}\big(\widehat{P}_{\xi,L}^{\mathrm{MC}}\big) = \frac{1}{N}\mathrm{Var}\big(Q_{\xi,L}\big) + \big(\mathbb{E}[Q_{\xi,L} - Q_\xi]\big)^2 \tag{3.5}$$

and

$$\mathrm{MSE}\big(\widehat{P}_{\xi,L}^{\mathrm{MLMC}}\big) = \sum_{\ell=0}^{L} \frac{1}{N_\ell}\mathrm{Var}\big(Q_{\xi,\ell} - Q_{\xi,\ell-1}\big) + \big(\mathbb{E}[Q_{\xi,L} - Q_\xi]\big)^2, \tag{3.6}$$

where the latter terms, known as the squared expectable approximation error or also called the squared numerical bias contribution, cannot be improved by the estimator as long as the maximal level is fixed and besides only appear if $P_{\xi,L} \neq P_\xi$. Focusing on the remaining stochastic error contribution the following proposition suggests a decreasing number of necessary samples the finer the level gets, i.e., $N_\ell \to 0$ as ℓ approaches infinity.

Proposition 3.2
If $Q_{\xi,\ell}$ converges to Q_ξ in L^2, then $\mathrm{Var}\big(Q_{\xi,\ell} - Q_{\xi,\ell-1}\big) \to 0$ as $\ell \to \infty$.

Proof. Using L^2 convergence, i.e., $\mathbb{E}\big[|Q_{\xi,\ell} - Q_\xi|^2\big] \to 0$ as $\ell \to \infty$, one obtains

$$\begin{aligned} 0 \leq \mathrm{Var}\big(Q_{\xi,\ell} - Q_{\xi,\ell-1}\big) &= \mathbb{E}\big[(Q_{\xi,\ell} - Q_{\xi,\ell-1})^2\big] - \big(\mathbb{E}(Q_{\xi,\ell} - Q_{\xi,\ell-1})\big)^2 \leq \mathbb{E}\big[|Q_{\xi,\ell} - Q_{\xi,\ell-1}|^2\big] \\ &\leq 2\Big(\mathbb{E}\big[|Q_{\xi,\ell} - Q_\xi|^2\big] + \mathbb{E}\big[|Q_\xi - Q_{\xi,\ell-1}|^2\big]\Big) \to 0 \text{ as } \ell \to \infty. \end{aligned}$$

□

Loosely speaking the ability of capturing a large part of the uncertainty on the coarser grids and consequently saving computations on finer levels has been stated.
A standard way of controlling the error of MLMC is to require that RMSE $\leq \epsilon$ or equivalently MSE $\leq \epsilon^2$. To this end a customary proceeding, see, e.g., [Gil15], is to determine the finest level L in a way that the numerical bias contribution is bounded

by $\epsilon/\sqrt{2}$. Then it remains to treat the contribution of the sampling error, where the subsequent outline follows [EHM16, Appendix A]. In order to obtain the optimal sample sizes on the different levels, the total cost of the MLMC method is minimized under the constraint that the variance of the estimator is bounded by $\epsilon^2/2$. This can be done analytically by treating the numbers of samples as continuous variables, which avoids that the optimization problem becomes an NP-hard knapsack problem, and yields

$$N_\ell = 2\epsilon^{-2}\sqrt{\frac{\mathrm{Var}\big(Q_{\xi,\ell}-Q_{\xi,\ell-1}\big)}{\mathcal{C}\big(Q_{\xi,\ell}\big)+\mathcal{C}\big(Q_{\xi,\ell-1}\big)}}\sum_{\tilde{\ell}=0}^{L}\sqrt{\mathrm{Var}\big(Q_{\xi,\tilde{\ell}}-Q_{\xi,\tilde{\ell}-1}\big)\big(\mathcal{C}(Q_{\xi,\tilde{\ell}})+\mathcal{C}(Q_{\xi,\tilde{\ell}-1})\big)}, \tag{3.7}$$

where $\mathcal{C}$ denotes the nonlinear cost operator. It shall be pointed out that it is sufficient for $\mathcal{C}$ to be some relative quantity. An alternative derivation of the same result can be found in [Gil15]. In the last part of this section a theorem on convergence and computational cost of MLMC is given and applied in the failure probability setting.

Theorem 3.3 (Convergence and Computational Cost of MLMC)
Let $\gamma \in (0,1)$ denote a refinement parameter and furthermore let α, β, q denote the positive constants, describing the convergence rates and the cost rate, respectively, such that

$$i)\ \big|\mathbb{E}\big[Q_{\xi,\ell}-Q_\xi\big]\big| \lesssim \gamma^{\alpha\ell}, \qquad ii)\ \mathrm{Var}\big(Q_{\xi,\ell}-Q_{\xi,\ell-1}\big) \lesssim \gamma^{\beta\ell}, \qquad iii)\ \mathcal{C}\big(Q_{\xi,\ell}\big) \lesssim \gamma^{-q\ell}$$

are fulfilled. Then, if $\alpha \geq \frac{1}{2}\min(\beta, q)$, for any $\epsilon < 1/e$, there exists a fine level L and a sequence $\{N_\ell\}_{\ell=0}^{L}$ such that

$$\mathrm{RMSE}\big(\widehat{P}_{\xi,L}^{\mathrm{MLMC}}\big) \leq \epsilon \tag{3.8}$$

and

$$\mathcal{C}\big(\widehat{P}_{\xi,L}^{\mathrm{MLMC}}\big) \lesssim \begin{cases} \epsilon^{-2} & \text{if } q < \beta \\ \epsilon^{-2}(\log\epsilon)^2 & \text{if } q = \beta \\ \epsilon^{-2-\frac{q-\beta}{\alpha}} & \text{if } q > \beta \end{cases}. \tag{3.9}$$

Proof. The proof can be found in and slightly adapted from [CGST11, Theorem 1]. □

Everything stated so far is applicable to any random variable, meaning that Q_ξ and $\{Q_{\xi,\ell}\}_{\ell=1}^{L}$ could be replaced by an arbitrary random variable and its discretizations, respectively. Smoothness provided, one typically obtains $\beta \approx 2\alpha$ for the convergence rates in Theorem 3.3.
Unfortunately, in the case of the clearly non-smooth failure probability functional Q_ξ for the convergence rates it holds $\alpha = \beta = 1$, as the upcoming proposition establishes. To arrive at this it is necessary to specify the approximation qualities of the discretizations of the QoI X. A straightforward way, motivated by a priori error bounds from customary PDE settings, is to require a uniform bound of the error with respect to the realizations.

Assumption 3.4 (Full or Uniform Refinement)
For a refinement parameter $\gamma \in (0,1)$ accuracy on level ℓ in the sense of full refinement is defined by requiring, that

$$|X - X_\ell| \leq \gamma^\ell \tag{3.10}$$

holds $\mathbb{P}$-a.s. in Ω.

This situation can be illustrated as follows. The gray shaded area in Figure 3.1 highlights all admissible values for the approximation error.

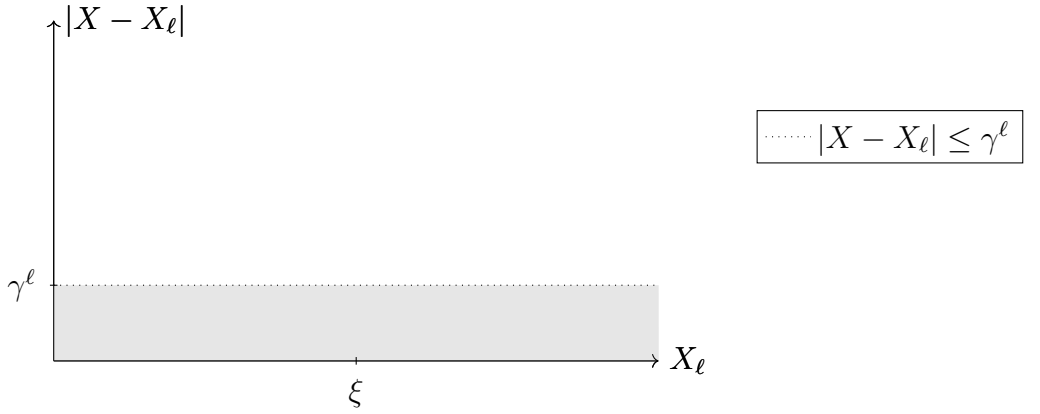


Figure 3.1: Illustration of full refinement (Assumption 3.4): The numerical error $|X - X_\ell|$ is bounded uniformly by γ^ℓ.

Proposition 3.5 (Convergence Rates of $Q_{\xi,\ell}$)
Let the full refinement property (3.10) be fulfilled.

i) Then it holds for the convergence rates i) and ii) in Theorem 3.3 that $\alpha = \beta = 1$, i.e., $\big|\mathbb{E}\big[Q_{\xi,\ell} - Q_\xi\big]\big| \lesssim \gamma^\ell$ and $\mathrm{Var}\big(Q_{\xi,\ell} - Q_{\xi,\ell-1}\big) \lesssim \gamma^\ell$.

ii) If the expectation of the error additionally follows $\big|\mathbb{E}\big[Q_{\xi,\ell} - Q_\xi\big]\big| \gtrsim \gamma^\ell$, then the convergence rate of the variance in part i) of the proposition cannot be improved.

Proof. Similarly to the proof of Lemmas 3.3–3.5 in [EHM16] for assertion *i)* at first the set $B = \{\omega \in \Omega : |X_\ell(\omega) - \xi| \leq \gamma^\ell\} \subset \Omega$ is defined. On B it holds due to the full refinement property (3.10) and the triangle inequality $|X - \xi| \leq |X - X_\ell| + |X_\ell - \xi| \leq 2\gamma^\ell$. Considering its complement $B^c = \{\omega \in \Omega : |X_\ell(\omega) - \xi| > \gamma^\ell\}$ the full refinement assumption yields $|X - X_\ell| \leq \gamma^\ell < |X_\ell - \xi|$. As the following lines show, this ensures $Q_{\xi,\ell} = Q_\xi$ on B^c, or equivalently $X_\ell \leq \xi \Leftrightarrow X \leq \xi$ on B^c.
"$\Rightarrow$": Since $|X_\ell - \xi| = \xi - X_\ell \leq \xi - X + |X - X_\ell| \leq \xi - X + |X_\ell - \xi|$ it follows $X \leq \xi$.
"$\Leftarrow$" by contradiction: Assuming $X \leq \xi$ as well as $X_\ell > \xi$ delivers a contradiction in the inequality chain $0 \leq |X - \xi| = \xi - X \leq \xi - X_\ell + |X_\ell - X| < \xi - X_\ell + |X_\ell - \xi| = 0$.
Making use of these observations, utilizing that $Q_{\xi,\ell} - Q_\xi$ only takes values in $\{0, \pm 1\}$ and exploiting the Lipschitz continuity of X's cdf F from Assumption 1.1 one can calculate

$$\begin{aligned}\big|\mathbb{E}[Q_{\xi,\ell} - Q_\xi]\big| &= \Big|\int_B Q_{\xi,\ell} - Q_\xi \,\mathrm{d}\mathbb{P}\Big| \leq \int_B 1 \,\mathrm{d}\mathbb{P} = \mathbb{P}(B) \leq \mathbb{P}\big(|X - \xi| \leq 2\gamma^\ell\big) \\ &= F(\xi + 2\gamma^\ell) - F(\xi - 2\gamma^\ell) \leq 4C_p\gamma^\ell.\end{aligned}$$

This proves the first part of the assertion. The second part can be traced back to the previous by an application of Steiner's translation theorem and the fact that $Q_{\xi,\ell}^2 = Q_{\xi,\ell}$.

$$\begin{aligned}\mathrm{Var}(Q_{\xi,\ell} - Q_{\xi,\ell-1}) &= \mathbb{E}\big[(Q_{\xi,\ell} - Q_{\xi,\ell-1})^2\big] - \big(\mathbb{E}[Q_{\xi,\ell} - Q_{\xi,\ell-1}]\big)^2 \leq \mathbb{E}\big[(Q_{\xi,\ell} - Q_{\xi,\ell-1})^2\big] \\ &= \mathbb{E}[Q_{\xi,\ell}^2 - 2Q_{\xi,\ell}Q_{\xi,\ell-1} + Q_{\xi,\ell-1}^2] \\ &= \mathbb{E}[Q_{\xi,\ell} - Q_\xi - 2Q_{\xi,\ell}Q_{\xi,\ell-1} + 2Q_\xi + Q_{\xi,\ell-1} - Q_\xi] \\ &\leq \big|\mathbb{E}[Q_{\xi,\ell} - Q_\xi]\big| + 2\big|\mathbb{E}[Q_{\xi,\ell}Q_{\xi,\ell-1} - Q_\xi]\big| + \big|\mathbb{E}[Q_{\xi,\ell-1} - Q_\xi\big| \\ &\leq 5\big(\big|\mathbb{E}[Q_{\xi,\ell} - Q_\xi]\big| + \big|\mathbb{E}[Q_{\xi,\ell-1} - Q_\xi]\big|\big) \\ &\leq 20C_p(1 + \gamma^{-1})\gamma^\ell,\end{aligned}$$

where $\big|\mathbb{E}[Q_{\xi,\ell}Q_{\xi,\ell-1} - Q_\xi]\big| \leq 2\big(\big|\mathbb{E}[Q_{\xi,\ell} - Q_\xi]\big| + \big|\mathbb{E}[Q_{\xi,\ell-1} - Q_\xi]\big|\big)$ has been used in the next-to-last inequality, which is shown in Lemma A.9 in Appendix A.3.
For assertion *ii)* one assumes that there is a better convergence rate, i.e., $\exists \beta > 1$ s.t. $\mathbb{V}[Q_{\xi,\ell} - Q_{\xi,\ell-1}] \leq C\gamma^{\beta\ell}$. Since $c\gamma^\ell \leq \big|\mathbb{E}[Q_{\xi,\ell} - Q_\xi]\big| \leq C\gamma^\ell$ holds, one can choose two levels k, ℓ with $k < \ell$ and $c\gamma^k < C\gamma^\ell$. Then

$$\big|\mathbb{E}[Q_{\xi,\ell} - Q_{\xi,k}]\big| \geq \Big|\big|\mathbb{E}[Q_{\xi,k} - Q_\xi]\big| - \big|\mathbb{E}[Q_{\xi,\ell} - Q_\xi]\big|\Big| \geq \tilde{c}\gamma^k,$$

where $\tilde{c}$ depends on the difference $\ell - k$ and is positive. Furthermore

$$\begin{aligned}
\tilde{c}\gamma^k \leq \big|\mathbb{E}[Q_{\xi,\ell} - Q_{\xi,k}]\big| &\leq \sum_{j=k}^{\ell-1} \big|\mathbb{E}[Q_{\xi,j+1} - Q_{\xi,j}]\big| \leq \sum_{j=k}^{\ell-1} \mathbb{E}\big[(Q_{\xi,j+1} - Q_{\xi,j})^2\big] \\
&= \sum_{j=k}^{\ell-1} \Big(\mathbb{V}[Q_{\xi,j+1} - Q_{\xi,j}] + \big(\mathbb{E}[Q_{\xi,j+1} - Q_{\xi,j}]\big)^2\Big) \\
&\lesssim \sum_{j=k}^{\ell-1} \big(\gamma^{\beta(j+1)} + \mathcal{O}(\gamma^{2(j+1)})\big) \lesssim \gamma^{\beta(k+1)} \lesssim \gamma^{\beta k}.
\end{aligned}$$

Keeping $\ell - k$ constant and $\ell, k \to \infty$ the latter calculation yields a contradiction if $\beta > 1$, since the rates don't match. □

Remark 3.6
A closer look at the first part of the proof of Proposition 3.5 reveals that the full refinement property (3.10) can be relaxed without weakening the validity of the statement. Namely, on the set B^c it suffices to require $|X - X_\ell| < |X_\ell - \xi|$ instead of the uniform bound $|X - X_\ell| \leq \gamma^\ell$. By definition of B^c the first one is indeed the milder one. As on B the assumption $|X - X_\ell| \leq \gamma^\ell$ is already the weaker one compared to $|X - X_\ell| < |X_\ell - \xi|$, it is kept. This encourages the modified refinement assumption, called selective refinement, which is introduced in the upcoming Section 3.2.

Before that, Theorem 3.3 is applied in the setting of failure probabilities. To this end it remains to quantify the computational cost for evaluating the functional $Q_{\xi,\ell}$. This is done by the following abstract assumption.

Assumption 3.7 (Cost for $Q_{\xi,\ell}$ and X_ℓ)
Let $q \in [0, \infty)$ denote the cost model parameter. The (expected) cost for computing one realization of $Q_{\xi,\ell}$ or X_ℓ is

$$\mathcal{C}[Q_{\xi,\ell}] = \mathcal{C}[X_\ell] = \gamma^{-q\ell}. \tag{3.11}$$

For customary numerical methods, like, e.g., finite elements or discontinuous Galerkin methods, properties of the spatial domain D and the used solver determine, amongst others, the cost model parameter q.

Corollary 3.8 (Convergence and Computational Cost of MLMC)
Let $\gamma \in (0, 1)$ denote a refinement parameter and assume that Assumption 3.4 and Assumption 3.7 hold. Then, for any $\epsilon < 1/e$, there exists a fine level L and a sequence $\{N_\ell\}_{\ell=0}^L$ such that

$$\mathrm{RMSE}\big(\widehat{P}_{\xi,L}^{\mathrm{MLMC}}\big) \leq \epsilon \tag{3.12}$$

and

$$\mathcal{C}\big(\widehat{P}_{\xi,L}^{\mathrm{MLMC}}\big) \lesssim \begin{cases} \epsilon^{-2} & \text{if } q < 1 \\ \epsilon^{-2}(\log \epsilon)^2 & \text{if } q = 1 \\ \epsilon^{-1-q} & \text{if } q > 1 \end{cases}. \tag{3.13}$$

Proof. This follows immediately from Theorem 3.3 and Proposition 3.5. □

3.2 Selective Refinement

Arising out of the special structure of the failure probability functional

$$Q_\xi = \mathbb{1}_{\{X \le \xi\}} \tag{3.14}$$

it is reasonable to expect, that this functional is sensitive to perturbations close to the critical value ξ, but rather insensitive to perturbations far from ξ. Thus it is enough to refine uniformly just around ξ, whereas for values with a bigger distance to ξ larger errors can be accepted. This idea is met by the selective refinement assumption. In order to distinguish the two strategies, the one proposed in this section is equipped with a prime, resulting in the notations X'_ℓ and $Q'_{\xi,\ell}$, respectively. $Q'_{\xi,\ell}$ arises from X'_ℓ in the natural way.

Assumption 3.9 (Selective Refinement)
For a refinement parameter $\gamma \in (0,1)$ accuracy on level ℓ in the sense of selective refinement is defined by requiring, that

$$|X - X'_\ell| \le \gamma^\ell \quad \text{or} \quad |X - X'_\ell| < |X'_\ell - \xi| \tag{3.15}$$

holds $\mathbb{P}$-a.s. in Ω.

As before an illustration of this strategy is given in Figure 3.2, which can be also found in [EHM16, p. 315].

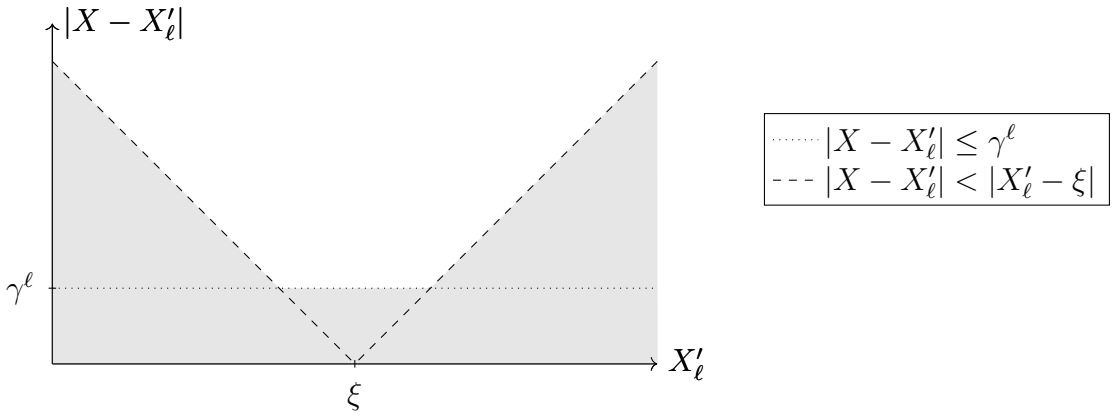


Figure 3.2: Illustration of selective refinement (Assumption 3.9): Far away from the threshold ξ the numerical error $|X - X'_\ell|$ may be larger than γ^ℓ.

A natural consistency assumption, which relates selective and full refinement, is to use the same numerical procedure for X_ℓ and X'_ℓ, respectively. This is made more rigorous by the implication in the subsequent assumption.

Assumption 3.10 (Coinciding Computational Procedure for Evaluating X_ℓ and X'_ℓ)
If (additionally to $|X - X_\ell| \leq \gamma^\ell$) also $|X - X'_\ell| \leq \gamma^\ell$ holds, then $X'_\ell = X_\ell$.

Having defined selective refinement its behavior can be analyzed.

3.2.1 Properties

The crucial observation working in the background of selective refinement is phrased in Proposition 3.11 below. It states that if X'_ℓ is further away from the threshold ξ than from the exact QoI X, i.e., lies in the gray shaded area and beneath the dashed line in Figure 3.2, this already ensures exactness for the approximated failure probability functional $Q'_{\xi,\ell}$.

Proposition 3.11 (Foundation of Selective Refinement)
If $|X - X'_\ell| < |X'_\ell - \xi|$ holds, where X'_ℓ fulfills the selective refinement property (3.15), then it follows $Q'_{\xi,\ell} = Q_\xi$, or equivalently $X'_\ell \leq \xi \Leftrightarrow X \leq \xi$.

Proof. Analogously to the part of the proof of Proposition 3.5, where the set B^c has been considered, one can show this statement by replacing X_ℓ with X'_ℓ and $Q_{\xi,\ell}$ with $Q'_{\xi,\ell}$, respectively. □

Remark 3.12
The same reasoning will be reused later on in the thesis, namely in the proof of Proposition 4.5, in order to verify the parametrized version of the slightly modified statement

$$|X_\ell - X_{\ell-1}| < |X_{\ell-1} - \xi_\ell| \Rightarrow Q_{\xi_\ell,\ell-1} = Q_{\xi_\ell,\ell}. \tag{3.16}$$

A surprising consequence of Proposition 3.11 is that the probability for $Q'_{\xi,\ell}$ to be exact is at least as high as for $Q_{\xi,\ell}$ to be exact, meaning that $Q'_{\xi,\ell}$ is not less accurate than $Q_{\xi,\ell}$. This is clarified in the following proposition.

Proposition 3.13 (Accuracy of $Q'_{\xi,\ell}$)
If X_ℓ and X'_ℓ are obtained by full and selective refinement, respectively, then it holds

$$\mathbb{P}(Q'_{\xi,\ell} = Q_\xi) \geq \mathbb{P}(Q_{\xi,\ell} = Q_\xi). \tag{3.17}$$

Proof. Analogously to the proof of Lemma 3.3 in [EHM16] the set $A = \{\omega \in \Omega : |X(\omega) - X'_\ell(\omega)| \leq \gamma^\ell\} \subset \Omega$ is defined. On A Assumption 3.10 of a coinciding computational procedure gives $Q'_{\xi,\ell} = Q_{\xi,\ell}$, delivering, that the equality $Q'_{\xi,\ell} = Q_\xi$ is here as probable as $Q_{\xi,\ell} = Q_\xi$, i.e., formally $\mathbb{P}(Q'_{\xi,\ell} = Q_\xi|A) = \mathbb{P}(Q_{\xi,\ell} = Q_\xi|A)$.
Taking the complement A^c into account, the definition directly states $|X - X'_\ell| > \gamma^\ell$ and hence $|X - X'_\ell| < |X'_\ell - \xi|$ due to the selective refinement property (3.15). This entails $Q'_{\xi,\ell} = Q_\xi$ by Proposition 3.11. Thus $\mathbb{P}(Q'_{\xi,\ell} = Q_\xi|A^c) = 1$. Since $\mathbb{P}$ is a probability measure $\mathbb{P}(Q_{\xi,\ell} = Q_\xi|A^c) \leq 1$ holds true. This completes the proof. □

At the end of this subsection and for the sake of completeness the discussion in Remark 3.6 shall be made more precise within the next proposition.

Proposition 3.14 (Convergence Rates of $Q'_{\xi,\ell}$)
Let the selective refinement property (3.15) be fulfilled.

i) Then it hold $\left|\mathbb{E}\left[Q'_{\xi,\ell} - Q_\xi\right]\right| \lesssim \gamma^\ell$ and $\mathrm{Var}\left(Q'_{\xi,\ell} - Q'_{\xi,\ell-1}\right) \lesssim \gamma^\ell$.

ii) If the expectation of the error additionally follows $\left|\mathbb{E}\left[Q'_{\xi,\ell} - Q_\xi\right]\right| \gtrsim \gamma^\ell$, *then the convergence rate of the variance in part i) cannot be improved.*

Proof. As already discussed in Remark 3.6 one can redo the proof of Proposition 3.5 in a straightforward manner. □

3.2.2 Implementation

A suitable procedure to obtain an approximation X'_ℓ fulfilling the selective refinement property is built on Proposition 3.11, i.e., if $|X - X'_\ell| < |X'_\ell - \xi|$ can be achieved before the highest possible accuracy is reached, there is no need for further computations, since $Q'_{\xi,\ell}$ is already exact and finer approximations will not improve the accuracy. In [EHM16] a way of formalizing this has been proposed, which is made more precise in Algorithm 2. The proposition directly afterwards ensures that the labeling in line 5 is legitimate.

Algorithm 2 Selective Refinement Algorithm

Input: Level ℓ, realization ω^i_ℓ, threshold ξ and refinement parameter γ.
Output: $X'_\ell(\omega^i_\ell)$

1: Let $j = 0$ and compute $X_0(\omega^i_\ell)$, such that $|X(\omega^i_\ell) - X_0(\omega^i_\ell)| \leq 1$ holds.
2: **while** $j < \ell$ and $\gamma^j \geq |X_j(\omega^i_\ell) - \xi|$
3: Let $j = j + 1$ and compute $X_j(\omega^i_\ell)$, such that $|X(\omega^i_\ell) - X_j(\omega^i_\ell)| \leq \gamma^j$ holds.
4: **end while**
5: Set $X'_\ell(\omega^i_\ell) = X_j(\omega^i_\ell)$.

Proposition 3.15
The approximations X'_ℓ *calculated by Algorithm 2 satisfy the selective refinement property* (3.15)

Proof. As there are two possible reasons for leaving the while loop, namely $j = \ell$ and $\gamma^j < |X_j(\omega^i_\ell) - \xi|$, two cases have to be considered.
The last computation in the first case is $X_\ell(\omega^i_\ell)$, such that $|X(\omega^i_\ell) - X_\ell(\omega^i_\ell)| \leq \gamma^\ell$ holds, which corresponds, adapting notation (line 5), to the first inequality in Assumption 3.9.
In the second case $X_j(\omega^i_\ell)$ is calculated lastly, satisfying both $\gamma^j < |X_j(\omega^i_\ell) - \xi|$ and $|X(\omega^i_\ell) - X_j(\omega^i_\ell)| \leq \gamma^j$. Adjusting the notation once more this yields $|X(\omega^i_\ell) - X'_\ell(\omega^i_\ell)| < |X'_\ell(\omega^i_\ell) - \xi|$, matching the second inequality in Assumption 3.9. □

3.2.3 Cost Reduction

The natural question arising is how selective refinement and especially Algorithm 2 affect the cost of computing the failure probability functional $Q'_{\xi,\ell}$. It is reasonable to assume that, in expectation, it is cheaper to compute $Q'_{\xi,\ell}$ than $Q_{\xi,\ell}$, since many realizations can be solved with lower accuracy. This is corroborated by the next proposition.

Proposition 3.16 (Cost for Q'_ℓ and X'_ℓ)
Let Assumption 3.7 on the cost of full refinement hold. Then the expected cost for computing the failure probability functional with Algorithm 2 is bounded by

$$\mathcal{C}[Q'_{\xi,\ell}] = \mathcal{C}[X'_\ell] \lesssim \sum_{j=0}^{\ell} \gamma^{-(q-1)j} \lesssim \gamma^{-(q-1)\ell}, \tag{3.18}$$

where the second inequality holds for $q \in [1, \infty)$.

Proof. Consider the iteration, where X_j will be calculated with tolerance γ^j, and denote the event that a realization enters this iteration by E_j. For $0 < j < \ell$ one has, using Assumption 1.1,

$$\begin{aligned}\mathbb{P}(E_j) &= \mathbb{P}(\gamma^{j-1} \geq |X_{j-1} - \xi|) = \mathbb{P}(\xi - \gamma^{j-1} \leq X_{j-1} \leq \xi + \gamma^{j-1}) \\ &\leq \mathbb{P}(\xi - 2\gamma^{j-1} \leq X \leq \xi + 2\gamma^{j-1}) = F(\xi + 2\gamma^{j-1}) - F(\xi - 2\gamma^{j-1}) \leq 4C_p\gamma^{j-1}.\end{aligned}$$

The total expected cost adds up to $\mathcal{C}[Q'_{\xi,\ell}] = \mathcal{C}[X'_\ell] = \sum_{j=0}^{\ell} \mathbb{P}(E_j)\gamma^{-qj} \lesssim \sum_{j=1}^{\ell} \gamma^{(1-q)j}$, having used that $\mathbb{P}(E_0) = 1$. □

This tells that selective refinement typically achieves a cost gain of one order of magnitude compared to full refinement, which turns out to hold as well if selective refinement is combined with MLMC. This is outlined in the next subsection.

3.2.4 Multilevel Monte Carlo using Selective Refinement

In order to incorporate the selective refinement idea into a customary MLMC algorithm as, e.g., presented in [CGST11], besides the obvious change to use $Q'_{\xi,\ell}$ instead of $Q_{\xi,\ell}$, it is also necessary to apply technical modifications in the estimation of the expectation and variance of the corrector terms. More details can be found in [EHM16]. This new method is called Multilevel Monte Carlo using Selective Refinement (MLMCSR).
Using the result of Subsection 3.2.3, the computational gain can be quantified.

Theorem 3.17 (Convergence and Computational Cost of MLMCSR)
Let $\gamma \in (0, 1)$ *denote a refinement parameter and assume that Assumption 3.9 and Assumption 3.7 hold. Then, for any* $\epsilon < 1/e$, *there exists a fine level* L *and a sequence* $\{N_\ell\}_{\ell=0}^{L}$ *such that*

$$\mathrm{RMSE}\big(\widehat{P}^{\mathrm{MLMCSR}}_{\xi,L}\big) \leq \epsilon \tag{3.19}$$

and

$$\mathcal{C}\big(\widehat{P}^{\mathrm{MLMCSR}}_{\xi,L}\big) \lesssim \begin{cases} \epsilon^{-2} & \text{if } q < 2 \\ \epsilon^{-2}(\log \epsilon)^2 & \text{if } q = 2 \\ \epsilon^{-q} & \text{if } q > 2 \end{cases}. \tag{3.20}$$

Proof. For $q > 1$ Corollary 3.8 applies with $q - 1$ instead of q due to Proposition 3.14 and Proposition 3.16. Otherwise, for $q \leq 1$, the total cost cannot be larger if a single sample is cheaper. Thus ϵ^{-2} is certainly an upper bound. □

The selective refinement idea presented in this chapter cleverly takes advantage of the failure probability functional's shape and offers a relatively weak but sufficient condition under which the QoI guarantees optimal precision and beyond that reduced cost compared to full refinement.

Chapter 4

Multilevel Preconditioning of the Cross Entropy Estimator

In this chapter the two main ideas for variance reduction from the previous chapters are brought together with the aim of obtaining an efficient estimator for failure probabilities. The foundation of this approach is importance sampling utilizing the cross-entropy method to derive the relevant biasing density. Broadly speaking, this ensures that rare events with probabilities as low as 10^{-9} can be addressed. Since the computations in the cross-entropy algorithm usually are expensive, this is the point where the multilevel idea is exploited in order to reduce the cost. This methodology has been proposed in a similar manner as presented here by Peherstorfer, Kramer and Willcox in [PKW18] and is called multifidelity preconditioned cross-entropy method. In the following it will be mainly limited to the familiar multilevel setting and therefore also referred to as multilevel cross-entropy importance sampling. The multifidelity approach is outlined in the outlook.
A comparison of standard importance sampling using the cross-entropy method and the previously mentioned multilevel variant confirms, that the expected cost savings indeed occur. Furthermore a parallel to multilevel Monte Carlo becomes evident, namely that again due to the failure probability functional's special shape the uniform refinement can be replaced by a weaker selective refinement assumption.
Practical considerations addressing a failure probability-like numerical example, degeneracy issues and inhomogeneous discretizations are covered at the end of this chapter.

4.1 Multilevel Cross-Entropy Importance Sampling

Before the multilevel idea is combined with the standard cross-entropy method from Subsection 2.2.3, its basics shall be repeated quickly.
Since the optimal (zero-variance) importance density q_* is not accessible for importance sampling in practice, the Kullback-Leibler divergence helps to assess whether a different density is at least close. The standard cross-entropy method constructs such a biasing density q_{v_*} among a prescribed parametrizable family $\mathcal{Q}$ of appropriate densities. This is done iteratively by approaching the rare event threshold ξ step-by-step with a strictly decreasing sequence $\{\xi_m\}_{m=1}^{m_L}$. Intermediate biasing densities $\{q_{v_m}\}_{m=1}^{m_L}$ are designed by means of the corresponding thresholds. In each iteration of the algorithm at first the new ξ_m is determined such that the respective rare event has a larger probability with respect

to the probability measure induced by the previous density $q_{v_{m-1}}$. Based upon this the next iterate of the biasing density q_{v_m} is obtained by solving an optimization problem. This process is initialized with the nominal density p and typically carried out on the finest level L to keep the discretization error as small as possible.
It would have been more favorable, if the previous procedure could have already been initialized with a density, which is closer to q_*. This would have reduced the computation time and saved expensive evaluations of the QoI f_L on the finest level L by decreasing the necessary number of iterations in the CE algorithm. However, it is not straightforward to construct such a density. This is the point, where the multilevel idea comes into play and takes effect, consolidating the plan to obtain a density for the fine level initialization from a coarser level. As the quantities appearing in the algorithm arise from approximations of different quality, it is necessary to equip them, additionally to the CE step counter, with a free index for the particular discretization level ℓ, which is chosen to be the first subscript per convention. To avoid confusion the multilevel quantities are endowed with a superscript $^{\mathrm{ML}}$.
This approach is explained in more detail in the following: Rather than initiating the cross-entropy method directly on the final level $\ell = L$, the complete single level method is firstly carried out on the coarsest level $\ell = 0$. This means that, based on the nominal density p, the first initial threshold $\xi_{0,1}^{\mathrm{ML}}$ is determined such that $\mathbb{P}(R_{\xi_{0,1}^{\mathrm{ML}}}) \approx \rho$, again for a $\rho \in (0,1)$ in the order of 10^{-1}. This is done by taking the ρ sample quantile of the QoI's evaluations on level $\ell = 0$, i.e., $\{f_0(z^1), \dots, f_0(z^{N_{\mathrm{MLCE}}})\}$, where the realizations $\{z^1, \dots, z^{N_{\mathrm{MLCE}}}\}$ of Z are distributed independently according to $p\lambda$. In the next step the parameter $v_{0,1}^{\mathrm{ML}}$ of the first biasing density on the coarsest level is estimated by solving the stochastic counterpart of

$$v_{0,1}^{\mathrm{ML}} = \underset{v \text{ s.t. } q_v \in \mathcal{Q}}{\arg\max}\, \mathbb{E}_p\big[I_{\xi_{0,1}^{\mathrm{ML}},0}(Z) \ln q_v(Z)\big]. \tag{4.1}$$

On this basis $\xi_{0,m}^{\mathrm{ML}}$ and $v_{0,m}^{\mathrm{ML}}$ are estimated for $m = 2, 3, \dots$ in an alternating manner, reutilizing the familiar importance sampling idea. This means, $\xi_{0,m}^{\mathrm{ML}}$ is selected such that

$$\mathbb{P}_{q_{v_{0,m-1}^{\mathrm{ML}}}}(R_{\xi_{0,m}^{\mathrm{ML}}}) \approx \rho, \tag{4.2}$$

where the performances of the coarsely discretized QoI $\{f_0(z^1), \dots, f_0(z^{N_{\mathrm{MLCE}}})\}$ enter. $\{z^1, \dots, z^{N_{\mathrm{MLCE}}}\}$ are newly drawn independent realizations of the random variable Z, distributed according to $q_{v_{0,m-1}^{\mathrm{ML}}}\lambda$. Subsequent to this $v_{0,m}^{\mathrm{ML}}$ is obtained via the stochastic counterpart of

$$v_{0,m}^{\mathrm{ML}} = \underset{v \text{ s.t. } q_v \in \mathcal{Q}}{\arg\max}\, \mathbb{E}_{q_{v_{0,m-1}^{\mathrm{ML}}}}\left[\big(I_{\xi_{0,m}^{\mathrm{ML}},0}(Z) \ln q_v(Z)\big)\frac{p(Z)}{q_{v_{0,m-1}^{\mathrm{ML}}}(Z)}\right]. \tag{4.3}$$

This proceeds until the threshold sequence falls below ξ. Then the coarse level calculations are performed one last time with threshold ξ yielding a final level-0-density $q_{v_{0,*}^{\mathrm{ML}}}$.
The successive procedure can be condensed by describing one single step on level ℓ in CE step m. There are two possible scenarios:

i) The computations on level $\ell - 1$ have been completed most recently and the last optimization problem has resulted in a density parameter $v_{\ell-1,*}^{\mathrm{ML}}$, or

ii) the computations carried out lastly have already been executed on level ℓ in the local cross-entropy step $m-1$ having delivered a parameter $v_{\ell,m-1}^{\mathrm{ML}}$.

Case *i)* corresponds to $m = 1$ setting $v_{\ell,0}^{\mathrm{ML}} = v_{\ell-1,*}^{\mathrm{ML}}$. In either case the density corresponding to the parameter $v_{\ell,m-1}^{\mathrm{ML}}$ is now used to steer the generation of independent realizations $\{z^1, \ldots, z^{N_{\mathrm{MLCE}}}\}$ of Z distributed accordingly. They in turn are used to evaluate the QoI on the current level ℓ, meaning that the performances $\{f_\ell(z^1), \ldots, f_\ell(z^{N_{\mathrm{MLCE}}})\}$ are calculated, from which the respective intermediate threshold $\xi_{\ell,m}^{\mathrm{ML}}$ is obtained such that

$$\mathbb{P}_{q_{v_{\ell,m-1}^{\mathrm{ML}}}}(R_{\xi_{\ell,m}^{\mathrm{ML}}}) \approx \rho. \tag{4.4}$$

Afterwards the accompanying biasing density on level ℓ in the CE step m results as usual from

$$v_{\ell,m}^{\mathrm{ML}} = \underset{v \text{ s.t. } q_v \in \mathcal{Q}}{\arg\max}\, \mathbb{E}_{q_{v_{\ell,m-1}^{\mathrm{ML}}}}\left[\left(I_{\xi_{\ell,m}^{\mathrm{ML}},\ell}(Z) \ln q_v(Z)\right) \frac{p(Z)}{q_{v_{\ell,m-1}^{\mathrm{ML}}}(Z)}\right]. \tag{4.5}$$

As previosly, this alternating and (with m) advancing estimation of intermediate threshold and density is kept on going until ξ is reached and a final level-ℓ-density $q_{v_{\ell,*}^{\mathrm{ML}}}$ is obtained, which then serves as starting point for biasing on level $\ell + 1$.
The iterative procedure is carried forward until also on the finest available and practicable level in the hierarchical discretization order, $\ell = L$, the true failure probability threshold ξ is reached. As final result the optimal multilevel cross-entropy (MLCE) density $q_{v_{L,*}^{\mathrm{ML}}}$ is obtained.
The preceding steps are summarized in the following pseudocode.

Algorithm 3 Multilevel Cross-Entropy Algorithm

Input: Nominal density parameter u, family of pdfs $\mathcal{Q}$, finest level L, parametric versions of discretized QoIs $\{f_\ell\}_{\ell=0}^{L}$, sample size N_{MLCE}, threshold ξ, quantile parameter ρ and minimal stepwidth δ.

Output: Parameter $v_{L,*}^{\mathrm{ML}}$ determining optimal MLCE density $q_{v_{L,*}^{\mathrm{ML}}}$

for $\ell = 0, \ldots, L$
  Initialize density $q_{v_{\ell,0}^{\mathrm{ML}}}$ with $v_{0,0}^{\mathrm{ML}} = u$ for $\ell = 0$ and with $v_{\ell,0}^{\mathrm{ML}} = v_{\ell-1,*}^{\mathrm{ML}}$ for $\ell > 0$. Then (re)set CE step counter to $m = 1$.
  while 1
    Generate i.i.d. realizations $\{z^i\}_{i=1}^{N_{\mathrm{MLCE}}}$ of Z distributed according to $q_{v_{\ell,m-1}^{\mathrm{ML}}}\lambda$.
    Evaluate the QoIs on level ℓ, i.e., compute $\{f_\ell(z^i)\}_{i=1}^{N_{\mathrm{MLCE}}}$.
    Set the intermediate threshold $\xi_{\ell,m}^{\mathrm{ML}}$ to the minimum of the ρ-quantile of the previous results and, if existent, $\xi_{\ell,m-1}^{\mathrm{ML}} - \delta$.
    if $\xi_{\ell,m}^{\mathrm{ML}} > \xi$ **then**
      Solve for $v_{\ell,m}^{\mathrm{ML}}$ using $\{f_\ell(z^i)\}_{i=1}^{N_{\mathrm{MLCE}}}$ and set $m = m + 1$.
    else
      Set $\xi_{\ell,m}^{\mathrm{ML}} = \xi$, solve for $v_{\ell,m}^{\mathrm{ML}}$ using $\{f_\ell(z^i)\}_{i=1}^{N_{\mathrm{MLCE}}}$ and **break** while loop with $v_{\ell,*}^{\mathrm{ML}} = v_{\ell,m}^{\mathrm{ML}}$.
    end if
  end while
end for

As the derivation of the biasing density is just a means to an end for the actual challenge of estimating the failure probability, the respective estimator making use of the optimal MLCE density for biasing is given in the subsequent definition.

Definition 4.1 (Multilevel Cross-Entropy Importance Sampling Estimator)
Let $\{Z^i\}_{i=1}^N$ *be i.i.d. copies of the random variable* Z *distributed according to the optimal MLCE density* $q_{v_{L,*}^{\mathrm{ML}}}\lambda$. *Then*

$$\widehat{P}_{\xi,L}^{\mathrm{MLCEIS}} = \frac{1}{N}\sum_{i=1}^{N} I_{\xi,L}(Z^i)\frac{p(Z^i)}{q_{v_{L,*}^{\mathrm{ML}}}(Z^i)} \tag{4.6}$$

defines the multilevel cross-entropy importance sampling (MLCEIS) estimator for the failure probability $P_{\xi,L}$.

Since $\widehat{P}_{\xi,L}^{\mathrm{MLCEIS}}$ is just an importance sampling estimator with a sampling density derived in a designated and moreover doable way, it is unbiased for $P_{\xi,L}$, provided that $v_{L,*}^{\mathrm{ML}}$ leads to a feasible density. Beyond that it approximates the exact failure probability P_ξ.
In the following, Example 2.3 is addressed with multilevel cross-entropy importance sampling.

Example 4.2 (Tail Estimate of the Normal Distribution – Fourth Part)
Similarly to Example 2.12 the importance density is constructed over the course of the procedure utilizing the introduced multilevel cross-entropy method, whose level structure is as described in Example 2.3. Apart from that the same setting as for the single level version is used. Figure 4.1 illustrates the method using additionally to the different opacities, indicating the progress of the CE step counter, a separate color for each level. Note that each density entails an intermediate threshold, from which the last one on every level coincides with the green marked ξ.

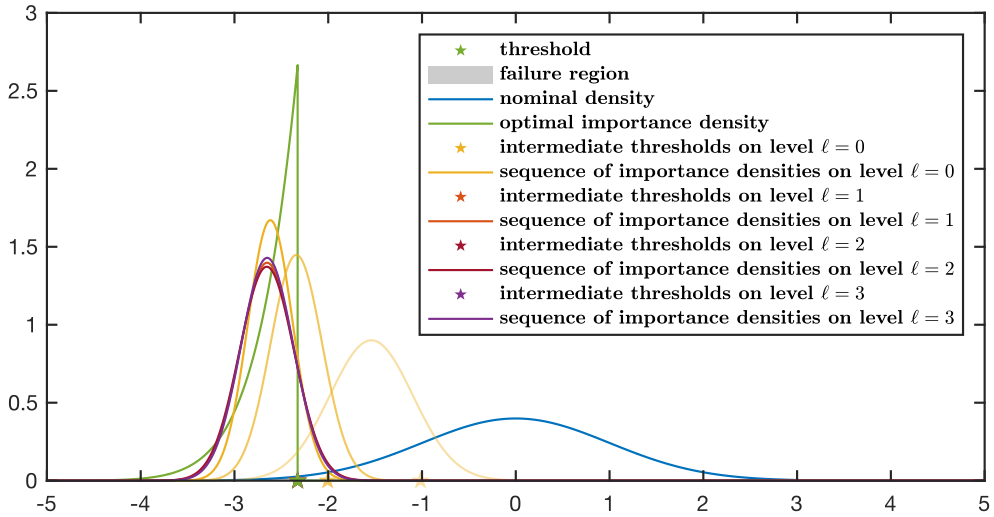


Figure 4.1: Tail estimate of the normal distribution (Example 4.2): Importance sampling utilizing the MLCE method to obtain an importance density. The parameters used in the MLCE method are $L = 3$, $\rho = 0.15$, $\delta = 10^{-2}$ and $N_{\mathrm{MLCE}} = 10^3$. $\mathcal{Q}$ is the family of Gaussian densities with variance larger than $5 \cdot 10^{-2}$.

The figure tells that an extensive part of the approaching procedure towards the threshold can be carried out on the coarsest level $\ell = 0$. In the present example the levels in-between do not have much effect, but this could change in general, if the discretizations are less homogeneous. One further observes that the MLCE method as well as the single level CE method arrive at similar densities, compare, e.g., Figure 2.3.

Table 4.1 presents the results in the usual way, taking $N_{\mathrm{MLCE}} = N$.

N	**Run 1**	**Run 2**	**Run 3**	**Run 4**	**estimated** CV based on 10^4 runs
10^1	$1.1197 \cdot 10^{-4}$	$3.1821 \cdot 10^{-5}$	$8.7080 \cdot 10^{-5}$	$1.3028 \cdot 10^{-4}$	0.6650 (−0.0114)
10^2	$1.0870 \cdot 10^{-4}$	$8.2439 \cdot 10^{-5}$	$9.1418 \cdot 10^{-5}$	$9.2844 \cdot 10^{-5}$	0.1492 (−0.0022)
10^3	$1.0083 \cdot 10^{-4}$	$1.0387 \cdot 10^{-4}$	$9.5043 \cdot 10^{-5}$	$1.0837 \cdot 10^{-4}$	0.0362 (−0.0006)
10^4	$1.0005 \cdot 10^{-4}$	$1.0122 \cdot 10^{-4}$	$9.9863 \cdot 10^{-5}$	$1.0035 \cdot 10^{-4}$	0.0113 (+0.0001)

Table 4.1: Tail estimate of the normal distribution (Example 4.2): Results of the MLCEIS estimation. The parameters used in the MLCE method are $L = 3$, $\rho = 0.1$ and $\delta = 10^{-2}$. $\mathcal{Q}$ is the family of Gaussian densities with variance larger than 10^{-1}.

For the comparison of the numerical results to the single level version only one aspect shall be illuminated at this point. Namely, one notices that the estimated CVs are very similar for both methods. This encourages that MLCEIS does not entail accuracy loss compared to CEIS. A further comment on the cost is postponed to the discussion in Example 4.6.

4.2 Theoretical Analysis

The multilevel framework comes with extra effort. This includes a more comprehensive data structure, the need of approximations of different quality and foremost a theoretically expected and numerically observed higher absolute number of total iterations. However, it can be offset in many cases by the cost savings induced by the multilevel structure. In order to formalize this the following analysis considers worst case bounds on the number of iterations on the different levels and therefore also on the cost of the method. To do so the worst case number of CE steps on level ℓ is defined by

$$M_\ell = 1 + \left(\frac{\xi_{\ell,1}^{\mathrm{ML}} - \xi}{\delta} \right)^+ . \tag{4.7}$$

That this is well-defined, is evident. Clearly it holds $m_\ell \leq M_\ell$, where m_ℓ labels the actual number of CE steps on level ℓ. Typically the upper bound is firstly very generous, but secondly up to a factor a good measure for m_ℓ. The resulting bound of the total cost of MLCEIS can be directly deduced:

$$\mathcal{C}\big[\widehat{P}_{\xi,L}^{\mathrm{MLCEIS}}\big] \leq \sum_{\ell=0}^{L} \Big(\big(M_\ell N_{\mathrm{MLCE}} + \delta_{\ell L} N\big) \mathcal{C}[I_{\xi,\ell}(Z)] + M_\ell \mathcal{C}[\text{density optimization on level } \ell] \Big)$$

The following theorem, similar to the respective one in [PKW18, Proposition 1], formalizes the motivation of the multilevel version of the cross-entropy method addressed already briefly at the beginning of the previous section.

Theorem 4.3 (Multilevel Preconditioning of the Cross-Entropy Method)
For $\ell \in \{1, \dots, L\}$ let $v_{\ell-1,*}^{\mathrm{ML}}$ denote the estimated intermediate CE density parameter on level $\ell - 1$. Furthermore, let $\xi_{\ell,1}^{\mathrm{ML}}$ denote the ρ-quantile of $f_\ell(Z)$ for $Z \sim q_{v_{\ell-1,*}^{\mathrm{ML}}} \lambda$ and let $\widetilde{Z} \sim p\lambda$. If

$$\mathbb{P}_{q_{v_{\ell-1,*}^{\mathrm{ML}}}} \big(f_\ell(Z) \leq \xi_{\ell,1}^{\mathrm{ML}} \big) \geq \mathbb{P}_p \big(f_\ell(\widetilde{Z}) \leq \xi_{\ell,1}^{\mathrm{ML}} \big), \tag{4.8}$$

then it holds that the worst case number of CE steps M_ℓ is at least as small if the CE method is initialized with $q_{v^{\mathrm{ML}}_{\ell-1,*}}$ as if it would have been initialized with p.

Proof. It suffices to show $\xi^{\mathrm{ML}}_{\ell,1} \leq \xi_p$, where ξ_p denotes the ρ-quantile of $f_\ell(\widetilde{Z})$, i.e., one has $\mathbb{P}_p\big(f_\ell(\widetilde{Z}) \leq \xi_p\big) = \rho$. By definiton it also holds $\mathbb{P}_{q_{v^{\mathrm{ML}}_{\ell-1,*}}}\big(f_\ell(Z) \leq \xi^{\mathrm{ML}}_{\ell,1}\big) = \rho$.
Exploiting the assumption yields $\mathbb{P}_p\big(f_\ell(\widetilde{Z}) \leq \xi^{\mathrm{ML}}_{\ell,1}\big) \leq \mathbb{P}_p\big(f_\ell(\widetilde{Z}) \leq \xi_p\big)$, delivering the claim, as the map $\tilde{\xi} \mapsto \mathbb{P}_p\big(f_\ell(\widetilde{Z}) \leq \tilde{\xi}\big)$ is the cdf of $f_\ell(\widetilde{Z})$ and thus non-decreasing. □

The additional condition (4.8) in Theorem 4.3 expresses that the density $q_{v^{\mathrm{ML}}_{\ell-1,*}}$ constructed by Algorithm 3 is better suited for importance sampling with respect to the rare event $R_{\xi^{\mathrm{ML}}_{\ell,1}}$ than the nominal density p.
It remains to find a plausible assumption on the discretizations of the QoI in order to meet condition (4.8). One can show that

$$|f_\ell(z) - f_{\ell-1}(z)| \leq \gamma^\ell \quad \text{for all} \quad z \in \Theta \tag{4.9}$$

is sufficient. However, similarly to the situation in the previous chapter it turns out that this is unnecessarily restrictive. To this end a selective refinement assumption for the present setting exploiting the special structure of the parametrized failure probability functional is proposed. Note that the following assumption is slightly modified compared to Assumption 3.9.

Assumption 4.4 (Selective Refinement for MLCEIS)
For a refinement parameter $\gamma \in (0,1)$ accuracy on level $\ell \in \{1,\dots,L\}$ for the parametrized QoI in the sense of selective refinement is defined by requiring, that

$$|f_\ell - f_{\ell-1}| \leq \gamma^\ell \quad \text{or} \quad |f_\ell - f_{\ell-1}| < |f_{\ell-1} - \xi^{\mathrm{ML}}_{\ell,1}| \tag{4.10}$$

holds pointwise on Θ, where $\xi^{\mathrm{ML}}_{\ell,1}$ denotes the ρ-quantile of $f_\ell(Z)$ for $Z \sim q_{v^{\mathrm{ML}}_{\ell-1,*}}\lambda$.

Proposition 4.5
Let $v^{\mathrm{ML}}_{\ell-1,*}$, $\xi^{\mathrm{ML}}_{\ell,1}$, Z and $\widetilde{Z}$ be as in Theorem 4.3, let f_ℓ fulfill Assumption 4.4 and let the consistency condition

$$\mathbb{P}_{q_{v^{\mathrm{ML}}_{\ell-1,*}}}\big(f_{\ell-1}(Z) \leq \xi^{\mathrm{ML}}_{\ell,1}\big) \geq \mathbb{P}_p\big(f_{\ell-1}(\widetilde{Z}) \leq \xi^{\mathrm{ML}}_{\ell,1}\big) \tag{4.11}$$

be satisfied. Then it holds

$$\mathbb{P}_{q_{v^{\mathrm{ML}}_{\ell-1,*}}}\big(f_\ell(Z) \leq \xi^{\mathrm{ML}}_{\ell,1}\big) \geq \mathbb{P}_p\big(f_\ell(\widetilde{Z}) \leq \xi^{\mathrm{ML}}_{\ell,1}\big) - 8C\gamma^\ell. \tag{4.12}$$

Proof. As in the proof of Proposition 2 in [PKW18] and in analogy to the one of Proposition 3.5 the set $B = \{z \in \Theta : |f_{\ell-1}(z) - \xi^{\mathrm{ML}}_{\ell,1}| \leq \gamma^\ell\}$ is defined, on whose complement by the selective refinement property (4.10) for MLCEIS and by definition of the set B, $|f_\ell(z) - f_{\ell-1}(z)| < |f_{\ell-1}(z) - \xi^{\mathrm{ML}}_{\ell,1}|$ is certainly correct. According to Remark 3.12, this immediately implies $I_{\xi^{\mathrm{ML}}_{\ell,1},\ell} = I_{\xi^{\mathrm{ML}}_{\ell,1},\ell-1}$ on B^c, as the setting is completely analogous.
Thus one arrives at

$$\mathbb{P}_p\big(f_\ell(\widetilde{Z}) \leq \xi^{\mathrm{ML}}_{\ell,1}\big) = \int_B I_{\xi^{\mathrm{ML}}_{\ell,1},\ell}(z)p(z)\,\mathrm{d}\lambda(z) + \int_{B^c} I_{\xi^{\mathrm{ML}}_{\ell,1},\ell-1}(z)p(z)\,\mathrm{d}\lambda(z)$$

$$\begin{aligned}
&\leq \int_B I_{\xi_{\ell,1}^{\mathrm{ML}},\ell}(z)p(z)\,\mathrm{d}\lambda(z) + \mathbb{P}_p\big(f_{\ell-1}(\widetilde{Z}) \leq \xi_{\ell,1}^{\mathrm{ML}}\big) \\
&\leq \int_B I_{\xi_{\ell,1}^{\mathrm{ML}},\ell}(z)p(z)\,\mathrm{d}\lambda(z) + \mathbb{P}_{q_{v_{\ell-1,*}^{\mathrm{ML}}}}\big(f_{\ell-1}(Z) \leq \xi_{\ell,1}^{\mathrm{ML}}\big),
\end{aligned} \tag{4.13}$$

having exploited the non-negativity of $I_{\xi_{\ell,1}^{\mathrm{ML}},\ell-1}$ in the first inequality and the consistency condition in the second.
In order to get closer to the final form of the statement, the information content of the second summand in (4.13) shall be transferred to the finer level ℓ. This is possible by admitting an error of the order $\mathcal{O}(\gamma^\ell)$, which is content of the proof's last part. In the first place the calculations are continued as follows.

$$\begin{aligned}
\mathbb{P}_p\big(f_\ell(\widetilde{Z}) \leq \xi_{\ell,1}^{\mathrm{ML}}\big) &\overset{(4.13)}{\leq} \int_B I_{\xi_{\ell,1}^{\mathrm{ML}},\ell}(z)p(z)\,\mathrm{d}\lambda(z) + \int_B I_{\xi_{\ell,1}^{\mathrm{ML}},\ell-1}(z)q_{v_{\ell-1,*}^{\mathrm{ML}}}(z)\,\mathrm{d}\lambda(z) \\
&\quad + \int_{B^c} I_{\xi_{\ell,1}^{\mathrm{ML}},\ell}(z)q_{v_{\ell-1,*}^{\mathrm{ML}}}(z)\,\mathrm{d}\lambda(z) \\
&\leq \int_B I_{\xi_{\ell,1}^{\mathrm{ML}},\ell}(z)p(z)\,\mathrm{d}\lambda(z) + \int_B I_{\xi_{\ell,1}^{\mathrm{ML}},\ell-1}(z)q_{v_{\ell-1,*}^{\mathrm{ML}}}(z)\,\mathrm{d}\lambda(z) \\
&\quad + \mathbb{P}_{q_{v_{\ell-1,*}^{\mathrm{ML}}}}\big(f_\ell(Z) \leq \xi_{\ell,1}^{\mathrm{ML}}\big),
\end{aligned} \tag{4.14}$$

where for verification of the next-to-last inequality the same reasoning as in the beginning, namely that $I_{\xi_{\ell,1}^{\mathrm{ML}},\ell}(z) = I_{\xi_{\ell,1}^{\mathrm{ML}},\ell-1}(z)$ for $z \in B^c$ has been utilized. The last step has just exploited $I_{\xi_{\ell,1}^{\mathrm{ML}},\ell} \geq 0$.
The remainder, controlling the first two terms in (4.14), follows two similar calculations utilizing model regularity in form of Assumption 1.3 and the selective refinement property (4.10). Firstly,

$$\begin{aligned}
\int_B I_{\xi_{\ell,1}^{\mathrm{ML}},\ell}(z)p(z)\,\mathrm{d}\lambda(z) &\leq \mathbb{P}_p\big(\widetilde{Z}^{-1}(B)\big) = \mathbb{P}_p\big(|f_{\ell-1}(\widetilde{Z}) - \xi_{\ell,1}^{\mathrm{ML}}| \leq \gamma^\ell\big) \\
&\leq \mathbb{P}_p\big(|f_\ell(\widetilde{Z}) - \xi_{\ell,1}^{\mathrm{ML}}| \leq 2\gamma^\ell\big) \\
&= F_{p,\ell}(\xi_{\ell,1}^{\mathrm{ML}} + 2\gamma^\ell) - F_{p,\ell}(\xi_{\ell,1}^{\mathrm{ML}} - 2\gamma^\ell) \leq 4C\gamma^\ell.
\end{aligned}$$

The inequality in the second line is based on the triangle inequality, the selective refinement property (4.10) and the definition of the set B. In detail it holds

$$\begin{aligned}
|f_\ell(\widetilde{Z}) - \xi_{\ell,1}^{\mathrm{ML}}| &\leq |f_\ell(\widetilde{Z}) - f_{\ell-1}(\widetilde{Z})| + |f_{\ell-1}(\widetilde{Z}) - \xi_{\ell,1}^{\mathrm{ML}}| \\
&\leq \begin{cases} \gamma^\ell + |f_{\ell-1}(\widetilde{Z}) - \xi_{\ell,1}^{\mathrm{ML}}| \leq 2\gamma^\ell & \text{if} \quad |f_\ell - f_{\ell-1}| \leq \gamma^\ell. \\ 2|f_{\ell-1}(\widetilde{Z}) - \xi_{\ell,1}^{\mathrm{ML}}| \leq 2\gamma^\ell & \text{if} \quad |f_\ell - f_{\ell-1}| < |f_{\ell-1} - \xi_{\ell,1}^{\mathrm{ML}}|. \end{cases}
\end{aligned} \tag{4.15}$$

Secondly it can be shown in an analogous way

$$\begin{aligned}
\int_B I_{\xi_{\ell,1}^{\mathrm{ML}},\ell-1}(z)q_{v_{\ell-1,*}^{\mathrm{ML}}}(z)\,\mathrm{d}\lambda(z) &\leq \mathbb{P}_{q_{v_{\ell-1,*}^{\mathrm{ML}}}}\big(Z^{-1}(B)\big) = \mathbb{P}_{q_{v_{\ell-1,*}^{\mathrm{ML}}}}\big(|f_{\ell-1}(Z) - \xi_{\ell,1}^{\mathrm{ML}}| \leq \gamma^\ell\big) \\
&\leq \mathbb{P}_{q_{v_{\ell-1,*}^{\mathrm{ML}}}}\big(|f_\ell(Z) - \xi_{\ell,1}^{\mathrm{ML}}| \leq 2\gamma^\ell\big) \\
&= F_{q_{v_{\ell-1,*}^{\mathrm{ML}}},\ell}(\xi_{\ell,1}^{\mathrm{ML}} + 2\gamma^\ell) - F_{q_{v_{\ell-1,*}^{\mathrm{ML}}},\ell}(\xi_{\ell,1}^{\mathrm{ML}} - 2\gamma^\ell) \leq 4C\gamma^\ell,
\end{aligned}$$

having replaced $\widetilde{Z}$ by Z in the previous considerations (4.15) to reason the inequality in the second line. This completes the proof. □

Although Proposition 4.5 just guarantees that the prerequisite (4.8) of Theorem 4.3 is fulfilled up to an additive factor in γ^ℓ and additionally demands the consistency condition (4.11), the combination of both statements encourages that MLCEIS is able to reduce the number of necessary iterations in the CE algorithm on the finest (and therefore most expensive) level. The hope and the typical case is that the cost coming up for obtaining the biasing density $q_{v^{\mathrm{ML}}_{L-1,*}}$ is by far small enough to preserve the computational gain.
Up to now, this chapter has provided a method fusing two concepts to realize one common aim, the estimation of very small probabilities. Loosely speaking, importance sampling has been the central pillar ensuring being capable of reaching such small events. This endeavor has been supported by the cross-entropy algorithm offering a feasible way in order to be able to carry out importance sampling. The other pillar, the multilevel concept in combination with the selective refinement idea, has been integrated with the objective of reducing the cost of the method without a loss in precision.

4.3 Practical Considerations

4.3.1 A Realistic Failure Probability Example

To round off this accompanying toy example series on the tail estimate of the normal distribution and to demonstrate MLCEIS for a realistic failure probability, Example 2.3 is taken up once more in a slightly modified setting.
Note that in the remainder of this thesis the numerical calculations and figures are preformed for the same failure probability.

Example 4.6 (Tail Estimate of the Normal Distribution – Fifth Part)
Compared to Example 4.2 besides the central change of the failure probability to 10^{-8} just the restriction to the variance is adjusted to $5 \cdot 10^{-2}$ for the calculations and 10^{-2} for the plot. Figure 4.2 depicts the illustration of one single run.

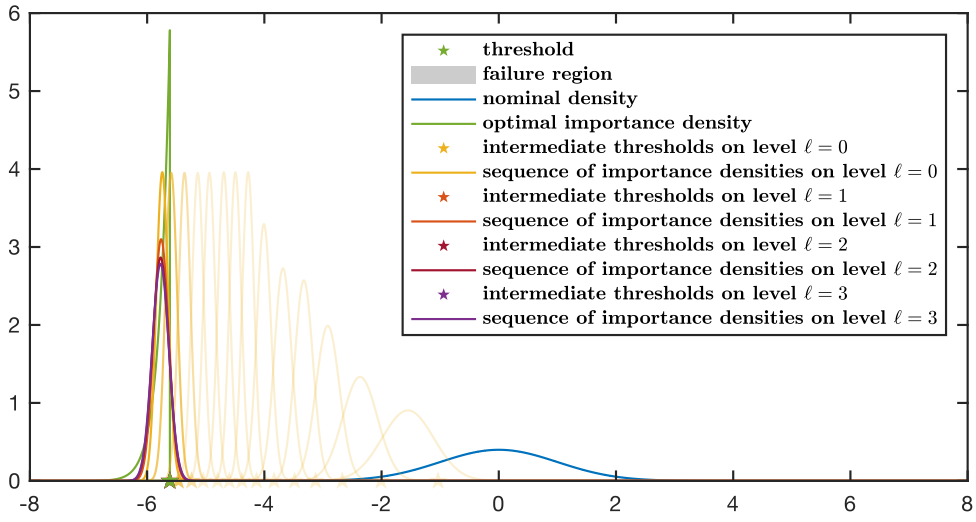


Figure 4.2: Tail estimate of the normal distribution (Example 4.6): Importance sampling utilizing the MLCE method to obtain an importance density in the realistic failure probability case. The parameters used in the MLCE method are $L = 3$, $\rho = 0.15$, $\delta = 10^{-2}$ and $N_{\mathrm{MLCE}} = 10^3$. $\mathcal{Q}$ is the family of Gaussian densities with variance larger than 10^{-2}.

As already addressed in Example 4.2, one observes that the majority of all iterations takes place on the coarsest level $\ell = 0$. Of course, this does not already imply, that also most work is spent on this level, since computations on the finest level are assumed to be way more expensive. However, compared to CEIS this means a significant speedup. A second and subsidiary behavior attracting attention is the change in the biasing densities between the seventh and fourteenth CE step on level $\ell = 0$. In contrast to the steps before the variance does not change any more, which is a result of the variance restriction to avoid degeneracy and a theoretical stagnation of the procedure.
As usual, Table 4.2 presents the numerical results.

N	**Run 1**	**Run 2**	**Run 3**	**Run 4**	**estimated** CV based on 10^4 runs
10^1	$1.0396 \cdot 10^{-8}$	$1.7366 \cdot 10^{-8}$	$2.2769 \cdot 10^{-8}$	$7.1416 \cdot 10^{-9}$	0.8241 (−0.0165)
10^2	$8.5852 \cdot 10^{-9}$	$1.0283 \cdot 10^{-8}$	$9.6539 \cdot 10^{-9}$	$8.7426 \cdot 10^{-9}$	0.1264 (−0.0022)
10^3	$1.0344 \cdot 10^{-8}$	$9.7438 \cdot 10^{-9}$	$1.0262 \cdot 10^{-8}$	$1.0073 \cdot 10^{-8}$	0.0377 (+0.0020)
10^4	$1.0075 \cdot 10^{-8}$	$9.8564 \cdot 10^{-9}$	$1.0001 \cdot 10^{-8}$	$9.9390 \cdot 10^{-9}$	0.0114 (+0.0004)

Table 4.2: Tail estimate of the normal distribution (Example 4.6): Results of the MLCEIS estimation in the realistic failure probability case. The parameters used in the MLCE method are $L = 3$, $\rho = 0.1$ and $\delta = 10^{-2}$. $\mathcal{Q}$ is the family of Gaussian densities with variance larger than $5 \cdot 10^{-2}$.

The results are quite convincing and look surprisingly very similar to the ones for the 4 orders higher failure probability. This convinces that the MLCEIS method is capable of addressing really small failure probabilities.
It remains to investigate whether this is also computationally manageable. To this end Table 4.3 compares the cost of the single level CEIS method and its multilevel variant. The table below has been created as follows: For the familiar four different sample sizes N used throughout the algorithm during the construction of the biasing densities and for the final failure probability estimate, the average m_ℓ, based on 10^4 runs of the estimator, has been determined for each level $\ell \in \{0, \dots, L\}$ with $L = 3$.

N	**Method**	**Density Construction**				**Final Estimation**
		$\ell = 0$	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 3$
10^1	CEIS	–	–	–	4.5572	1
	MLCEIS	4.7634	1.0005	1.0003	1.0003	1
10^2	CEIS	–	–	–	8.0167	1
	MLCEIS	8.0684	1	1	1	1
10^3	CEIS	–	–	–	7.9886	1
	MLCEIS	8.0051	1	1	1	1
10^4	CEIS	–	–	–	8.2275	1
	MLCEIS	8.2352	1	1	1	1

Table 4.3: Tail estimate of the normal distribution (Example 4.6): Cost comparison of CEIS and MLCEIS in the realistic failure probability case. The parameters used in the CE method are $\rho = 0.1$ and $\delta = 10^{-2}$. Additionally, for the MLCE method $L = 3$. $\mathcal{Q}$ is the family of Gaussian densities with variance larger than $5 \cdot 10^{-2}$.

As expected, MLCEIS dramatically reduces the number of iterations on the finest level $\ell = L = 3$ to the absolute minimum of one CE step in almost all cases by outsourcing the bulk of the threshold approaching procedure to the coarsest level $\ell = 0$. Although m_0 for MLCEIS is on average already slightly higher than the CE step counter for CEIS, yielding that the total number of iterations is considerably higher, namely by slightly more than L, the multilevel method pays off as soon as the computational cost on the finest level is high enough.

4.3.2 Degeneracy Issues

The example series on the tail estimate of the normal distribution was accompanied throughout the thesis by mentioning that the variance is restricted in a way that it does not become to small. Formally this has been reasoned by showing that the resulting estimators have infinite variance, see Example 2.7. Ignoring this fact and running MLCEIS without any restriction to counterbalance such issues typically results in a figure as pictured below. The failure probability is still 10^{-8} and with the exception of the skipped restriction to the variance the setting is the same as in Example 4.6.

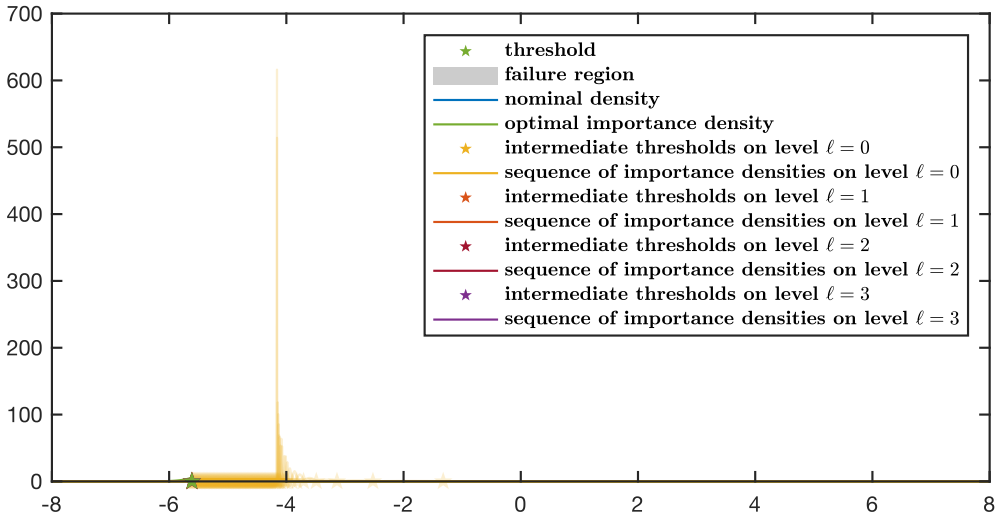


Figure 4.3: Tail estimate of the normal distribution: Degeneracy. An application of the MLCE (or CE) method without any restriction to the variance can deliver a density sequence converging to a Dirac measure, which is unfeasible for further use.

After a few CE steps on the coarsest level the biasing densities become more and more narrow until they degenerate and, depending on the implementation, the algorithm either breaks or just starts assigning NaN to all quantities. The reason for this is as follows: Before such an infeasible biasing density is delivered, the current biasing density is so narrow, that the threshold sequence proceeds by the minimal stepwidth δ only and, more importantly, none of the realizations of the QoI is smaller than this new threshold. This entails that during the density optimization procedure a 0/0 situation occurs.
Skipping the minimal stepwidth limitation is not an option, as δ is typically already so small, that proceeding with these steps is computationally unaffordable.
The more preferable option is to revise and adjust the chosen family $\mathcal{Q}$, which can be done

by restricting, e.g., the variance or by trying something like the in Example 2.7 presented Student's t variant.

4.3.3 Skewness and Inhomogeneous Discretizations

Until now it has seemed, that most of the relevant computations have taken place either on the coarsest level, where the major part of the approximation of the optimal importance density takes place, or on the finest level, where the final density is calculated and the importance sampling is performed. However, this might not be the case in general.
By considering coarse level approximations, which promote failure, one can construct an example, as this entails that the intermediate thresholds on the coarser levels reach ξ too fast. Consequently, finer discretizations have to redo some of this approaching procedure. A possible modification of the standard Example 2.3 is given by the discretizations

$$\widetilde{X}_\ell = X + \gamma^{\ell+4} \cdot \left(U - \frac{1}{2}\right) - 2\sum_{j=0}^{L-1} \gamma^j \delta_{\ell,j}.$$

Figure 4.4 provides an illustration. A very close look reveals the position of the intermediate thresholds and convinces that the nestedness of the intermediate rare events remains just level-wise.

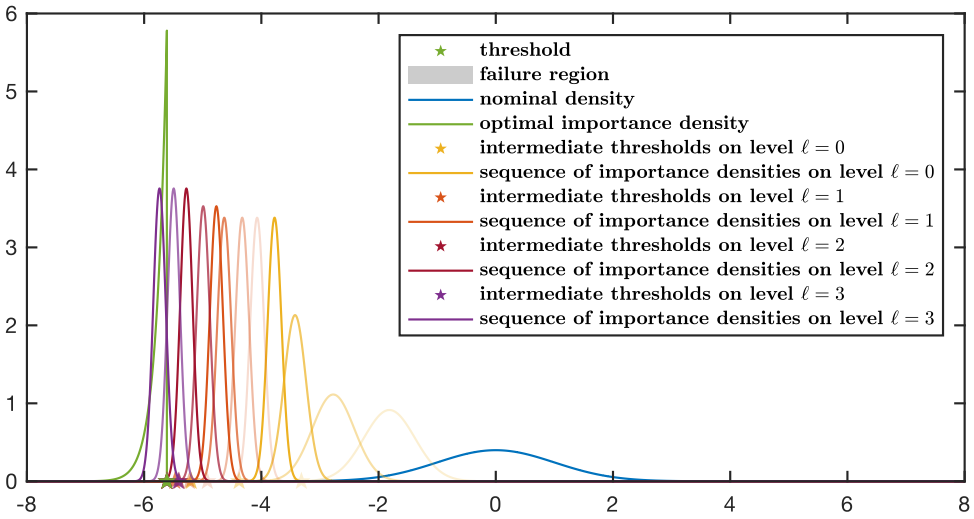


Figure 4.4: Tail estimate of the normal distribution: Skewness of the Discretizations. Different discretizations of the QoI cause different behavior of the MLCE method.

A short résumé to conclude this chapter. The cross-entropy importance sampling method from Subsection 2.2.3 has been taken to multiple levels or phrased more accurately, the search of the biasing density has been executed within a multilevel setting. This promises significant computational savings in several applications.

Chapter 5

Numerical Experiments

This final chapter presents two more evolved numerical settings to demonstrate the MLCEIS method. The first experiment is performed on a one-dimensional heat transfer problem and considers different parameter combinations within the MLCEIS method. The second experiment investigates a more expensive two-dimensional groundwater flow problem to verify the applicability of the method to realistic problems.

5.1 A Heat Transfer Problem

In this first experiment, modifying the one from [PKW18], the stationary heat equation in the domain $D = (0, 1)$ with uncertain heat conductivity a is considered. On the western boundary Γ_w a fixed temperature is imposed by a homogeneous Dirichlet BC, whereas zero heat flux, i.e., a homogeneous Neumann BC is imposed on the eastern boundary Γ_e. For any sample $\omega \in \Omega$ this is modeled by

$$\begin{aligned} -\big(a(x,\omega)\ u'(x,\omega)\big)' &= 1 \quad \text{for } x \in D \\ u(x,\omega)\mathbb{1}_{\Gamma_w}(x) + u'(x,\omega)\mathbb{1}_{\Gamma_e}(x) &= 0 \quad \text{for } x \in \Gamma = \partial D. \end{aligned}$$

The heat conductivity, described by a random field with stochastic dimension k, is, for a fixed vector $v \in \mathbb{R}^k$, given by

$$a(x,\omega) = \sum_{i=1}^{k} \exp\big(Z_i(\omega) - (2^5 + k)|x - v_i|\big)$$

with a normally distributed random vector $Z = (Z_1, \ldots, Z_k)^T \sim \mathcal{N}(\mu, \Sigma)$ with mean $\mu = (1, \ldots, 1)^T \in \mathbb{R}^k$ and covariance matrix $\Sigma = 0.1 \cdot \mathrm{Id}_k \in \mathbb{R}^{k\times k}$.
For the numerical spatial discretization linear finite elements are used on a uniform mesh with width $h_\ell = 2^{-4-\ell}$ for the level sequence $\ell \in \{0, \ldots, L\}$, where the finest level is set to $L = 4$.
The value of the temperature $u : \overline{D} \times \Omega \to \mathbb{R}$ on the eastern boundary $\Gamma_e = \{1\}$ serves as QoI, i.e.,

$$f\big(Z(\omega)\big) = u(1, \omega).$$

In a first setting the stochastic dimension is set to $k = 2$, v is chosen to be $(0.3, 0.8)^T$ and the system is said to fail if the QoI falls below $\xi = 62$. This entails a failure probability

$P_{\xi,L}$ of approximately $9.7 \cdot 10^{-9}$, which has been estimated with importance sampling exploiting prior knowledge about the failure region and a rather high sample size of 10^6. The used importance density has had a mean of $(2.8, 1.2)^T$ and the same covariance matrix as the nominal density.
Multilevel cross-entropy importance sampling is tested within this setting for a few different parameter combinations and the results are summarized in Table 5.1 below. For each parameter tuple, specified in the first four columns, at first the estimated probabilities of two runs are listed before the average in terms of the arithmetic mean of 20 single runs is given as well. At the end the CV is estimated. $\mathcal{Q}$ is the family of Gaussian pdfs with free mean and a variance bounded from below by *restr*.

N	ρ	δ	*restr*	**Run 1**	**Run 2**	**mean** of 20 runs	**estimated** CV based on 20 runs
10^3	0.1	1	$5 \cdot 10^{-2}$	$1.0397 \cdot 10^{-8}$	$9.5509 \cdot 10^{-9}$	$9.7365 \cdot 10^{-9}$	0.0739
10^3	0.25	1	$5 \cdot 10^{-2}$	$9.2970 \cdot 10^{-9}$	$9.6881 \cdot 10^{-9}$	$9.6354 \cdot 10^{-9}$	0.0642
10^4	0.1	1	$5 \cdot 10^{-2}$	$9.6820 \cdot 10^{-9}$	$9.9989 \cdot 10^{-9}$	$9.7468 \cdot 10^{-9}$	0.0329
10^4	0.2	1	$5 \cdot 10^{-2}$	$9.7707 \cdot 10^{-9}$	$9.5764 \cdot 10^{-9}$	$9.6409 \cdot 10^{-9}$	0.0228
10^3	0.1	1	10^{-2}	$8.7981 \cdot 10^{-9}$	$9.3743 \cdot 10^{-9}$	$9.6942 \cdot 10^{-9}$	0.0386
10^3	0.25	1	10^{-2}	$9.6634 \cdot 10^{-9}$	$9.1748 \cdot 10^{-9}$	$9.7450 \cdot 10^{-9}$	0.0360
10^4	0.1	1	10^{-2}	$9.6940 \cdot 10^{-9}$	$9.7830 \cdot 10^{-9}$	$9.7152 \cdot 10^{-9}$	0.0180
10^4	0.2	1	10^{-2}	$9.6514 \cdot 10^{-9}$	$9.8496 \cdot 10^{-9}$	$9.6988 \cdot 10^{-9}$	0.0168

Table 5.1: Heat transfer problem: Results of the MLCEIS estimation for the stochastic dimension $k = 2$.

Instead of trying to interpret too much into the results in Table 5.1 in dependence on the chosen parameters three things shall be mentioned. Firstly one observes that the arithmetic mean of a few runs offers a relatively stable and precise estimate as some special effects resulting from intrinsic randomness taking effect on the procedure of the algorithm are smoothened out very certainly. Secondly, despite not to deniable influence from the chosen parameters the effect in the result seems to be manageably small, as long as severe degeneracy issues are avoided. Lastly, for this rather small stochastic dimension the CV is in the order of the one from the toy example, already for a moderate number of samples N. However, the higher k gets, the more necessary larger sample sizes ensuring the reachability of any direction become as measures tend to be more different in higher dimensions [APSAS17]. This is known as the curse of dimensionality.
To improve the imagination of the heat transfer problem and the MLCEIS method working on this problem Figure 5.1 displays on the left-hand side the failure region and shadows the progression of it during the MLCEIS method by shadowing the intermediate nested failure regions arising within the algorithm. The right-hand side (with data from a run to the parameter tuple $(N = 10^3, \rho = 0.1, \delta = 1, \mathit{restr} = 10^{-2})$) plots the appearing intermediate thresholds, marked with a different color for each level $\ell \in \{0, \dots, 4\}$. Moreover the progression of the computation time is displayed for whole MLCEIS, i.e., the final estimate is included. The runtime measurement was performed on an Intel Core i7-2860QM and 16GB RAM on a single core using a MATLAB implementation.
Figure 5.1 furthermore confirms that MLCEIS spends most iterations on the coarse levels and therefore takes advantage of the relatively cheap model evaluations compared to the ones on the finest level.

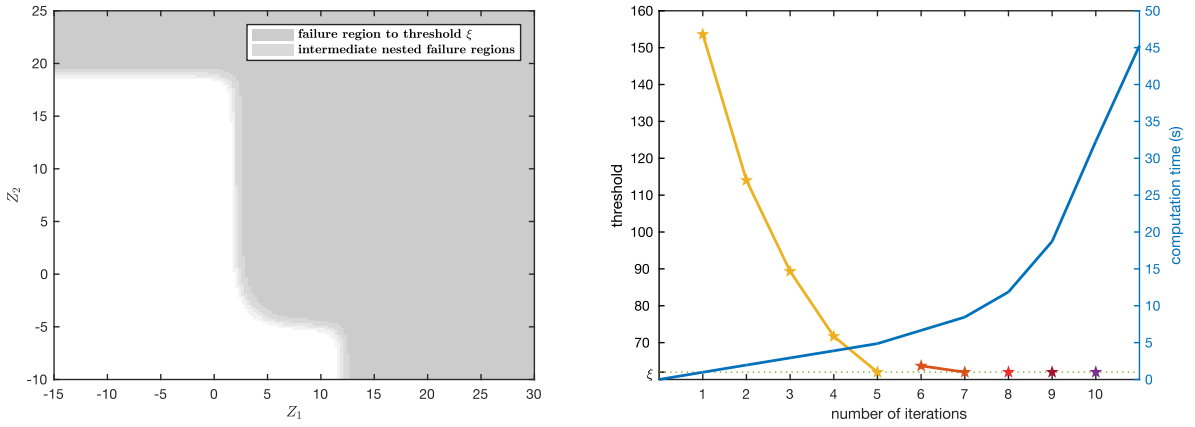


Figure 5.1: Heat transfer problem: Progression of failure region (left): The safe region is colored white, the intermediate failure regions arising in the MLCE method correspond to the different light gray shaded areas enlarging the failure region to threshold ξ, which itself is colored gray. Progression of threshold and computation time (right): The intermediate thresholds determined by Algorithm 3 are plotted with different colors corresponding to the different discretization levels. Additionally the graph of the computation time is depicted in blue.

For a second setting the stochastic dimension is raised to $k = 5$. This involves a modification of v, which is chosen such that the failure probability to the same ξ is roughly as before. $v = (0.28, 0.37, 0.78, 0.84, 0.98)^T$ is a good choice and delivers a $P_{\xi,L}$ of approximately $1.0 \cdot 10^{-8}$. Table 5.2 presents some numerical results in the previous manner.

N	ρ	δ	*restr*	**Run 1**	**Run 2**	**mean** of 20 runs	**estimated** CV based on 20 runs
10^3	0.1	1	$5 \cdot 10^{-2}$	$9.9123 \cdot 10^{-9}$	$9.6518 \cdot 10^{-9}$	$9.9296 \cdot 10^{-9}$	0.6792
10^4	0.1	1	$5 \cdot 10^{-2}$	$1.0153 \cdot 10^{-8}$	$1.0015 \cdot 10^{-8}$	$1.0107 \cdot 10^{-8}$	0.0200
10^3	0.1	1	10^{-2}	$1.0599 \cdot 10^{-8}$	$9.5639 \cdot 10^{-9}$	$1.0231 \cdot 10^{-8}$	0.6747
10^4	0.1	1	10^{-2}	$1.0257 \cdot 10^{-8}$	$9.9389 \cdot 10^{-9}$	$1.0130 \cdot 10^{-8}$	0.0156

Table 5.2: Heat transfer problem: Results of the MLCEIS estimation for the stochastic dimension $k = 5$.

As before, the single runs as well as the arithmetic mean provide good approximations of the failure probability with respect to stability and quality. One furthermore observes that a number of $N = 10^3$ samples is too small to deliver a CV comparable to the respective one for a smaller stochastic dimension. But a sample size of 10^4 suffices and ensures that each stochastic dimension can be reached.

5.2 A Groundwater Flow Problem

The second experiment, aligned to [UP15], investigates a steady-state flow problem in a porous medium in the domain $D = (0, 1) \times (0, 1)$ with uncertain permeability a. Additionally to the system of PDEs, described already in Chapter 1, it is necessary to specify BCs. Whereas $p \equiv 1$ is imposed on the western boundary Γ_w, a homogeneous Dirichlet

condition for p is imposed on the eastern side Γ_e. The remaining horizontal parts Γ_h are equipped with no-flow conditions. The system then reads in its mixed formulation as

$$\begin{aligned} a^{-1}\boldsymbol{u} + \nabla p &= 0 && \text{in } D \\ \operatorname{div}(\boldsymbol{u}) &= g && \text{in } D \\ p\mathbb{1}_{\Gamma_w} + p\mathbb{1}_{\Gamma_e} - (\boldsymbol{n}\cdot\boldsymbol{u})\mathbb{1}_{\Gamma_h} &= \mathbb{1}_{\Gamma_w} && \text{on } \Gamma = \partial D, \end{aligned}$$

where $\boldsymbol{n}$ denotes the outer normal of D and the random coefficient a is modeled as a log-normal random field, such that $\ln(a)$ is a mean-zero Gaussian random field with the exponential covariance function $\rho_{\exp}(x_1, x_2) = \exp(-\lambda^{-1}\|x_1 - x_2\|_1)$. The Karhunen-Loève expansion of a can be obtained analytically (see, e.g., [GS91, p. 29–32]), has the form and properties outlined in Section 1.2 and has to be truncated, say after k terms, in order to permit sampling. This entails a finite dimensional random variable $Z \sim \mathcal{N}\big((0,\dots,0)^T, \mathrm{Id}_k\big)$.
The spatial PDE discretization utilizes mixed $RT_0 - P_0$ finite elements on a simplicial mesh with $h_\ell = 2^{-1-\ell}$ for the level sequence $\ell \in \{0,\dots,L\}$ with $L = 2$. More precisely, $\mathrm{H}(\mathrm{div}; D)$ conforming, lowest order Raviart-Thomas finite elements are used for the Darcy velocity $\boldsymbol{u}$, while the pressure p is discretized with piecewise constants.
The QoI is the time it takes a particle, released in $\boldsymbol{x}_0 = (0, 0.5)^T$, to reach ∂D, i.e., D is chosen as safety zone. For simplicity the porosity of the soil is chosen to be $\varphi \equiv 1$. The theoretical setting is as outlined in Section 1.1.
Numerically this particle tracking is realized by computing the path along the finite element mesh and summing up the times needed for the relevant separate line segments.
The numerical tests, summarized in Table 5.3, are performed for the correlation length $\lambda = 0.64$ and a stochastic dimension of $k = 10$, which captures 81% of the variability of $\ln(a)$. The threshold ξ is set to 0.016, entailing a failure probability $P_{\xi,L}$ of approximately $6.7 \cdot 10^{-9}$. MLCEIS is tested in the same manner as in the previous section, just the arithmetic mean and the estimated CV are based on only 10 runs.

N	ρ	δ	*restr*	**Run 1**	**Run 2**	**mean** of 10 runs	**estimated** CV based on 10 runs
$2 \cdot 10^3$	0.2	10^{-4}	10^{-1}	$9.3302 \cdot 10^{-9}$	$6.8238 \cdot 10^{-9}$	$6.6404 \cdot 10^{-9}$	0.9854
$4 \cdot 10^3$	0.2	10^{-4}	10^{-1}	$6.4959 \cdot 10^{-9}$	$7.5676 \cdot 10^{-9}$	$6.7262 \cdot 10^{-9}$	0.2804

Table 5.3: Groundwater flow problem: Results of the MLCEIS estimation.

Good approximations of the failure probability are especially obtained by the arithmetic mean of a few runs. The single runs, however, show deviations of up to two times the failure probability for $N = 2 \cdot 10^3$ and of up to one half times the failure probability for $N = 4 \cdot 10^3$. This also reflects in the estimated CV.

Conclusions and Outlook

This thesis has been concerned with the estimation of rare event probabilities. To reduce the computational cost, a multilevel approach to cross-entropy based importance sampling has been presented and investigated, both theoretically and numerically. The method has been shown to be suitable for failure probabilities as low as 10^{-9} in a one-dimensional heat transfer and a two-dimensional groundwater flow setting. The numerical tests have focussed on small and moderate stochastic dimensions.
Furthermore, a novel heuristic methodology to avoid degeneracy issues in the cross-entropy method has been introduced. The so-called Student's t variant provides the opportunity to reduce necessary prior knowledge when choosing the distribution family $\mathcal{Q}$ in the cross-entropy method and makes the method more accessible to higher stochastic degrees of freedom.
Moreover, it has been observed that an implementation of selective refinement for the particular application is possible and can bring additional savings in computation time, especially for expensive high accuracy models.
The possibilities for further research, improvements and generalizations are vast. A first idea is to study and test modifications of the cross-entropy algorithm [dBKMR05, Chapter 4] and their applicability to the multilevel setting. To give an example, the fully automated cross-entropy (FACE) algorithm updates the sample size in each CE step, which allows to avoid subtle situations, identify them and optimize the cost. A second suggestion is to analyze the effect of the selected distribution family and further develop the introduced Student's t variant and, moreover, provide an analytical analysis.
A generalization, which has already taken place, is the integration of a larger variety of approximation techniques. Instead of the spatially orientated characterization of approximation quality, also other concepts can be used in order to obtain economical surrogates of the final and in general most expensive model [PKW18]. Arranging them hierarchically, usually according to their approximation quality and computational effort, delivers a structure similar to the familiar multilevel setting. In this so-called multifidelity framework, to give a few examples, simplified models, reduced basis methods, projection methods as well as machine learning techniques like neural networks or support vector machines can be found.

Appendix

This additional part provides a collection of further material referred to in the thesis but skipped there for reasons of clarity and brevity.

A.1 Conventions

Indexing. To keep the notation concise, but as lucidly as possible, the following convention w.r.t. indices is used throughout the thesis: As the multilevel version of the cross-entropy method demands two indices, one for the level ℓ and one for the CE step counter m the respective quantities are equipped with a double index, e.g. $v_{\ell,m}$ denotes the estimated density on level ℓ in CE step m, where ℓ reaches from 0 to L and m from 1 to the level dependent m_ℓ.

Notation. Sometimes when it is clear from the context the hat indicating estimators or estimated quantities is skipped. This affects for example $\xi_{\ell,m}$ and $v_{\ell,m}$.

Parametrizing Random Variable Z. $\mathbb{P}$ induces the distribution $p\lambda$ for Z, see Section 1.1. If Z is assumed to have a different distribution with density q, because Z is, e.g., sampled from such a distribution, it is always stated explicitly via $Z \sim q\lambda$.

A.2 Multivariate Normal and Student's t Distribution

Definition A.1 (Multivariate Normal Distribution)
A k-dimensional random variable Z follows a non-degenerate multivariate normal distribution with mean $\mu \in \mathbb{R}^k$ and a symmetric positive definite covariance matrix $\Sigma \in \mathbb{R}^{k\times k}$, if Z has a pdf p with

$$p(z) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left(-\frac{1}{2}(z-\mu)^T\Sigma^{-1}(z-\mu)\right).$$

Definition A.2 (Multivariate Student's t Distribution)
A k-dimensional random variable Z follows a non-degenerate multivariate Student's t distribution with degree of freedom (dof) $\nu > 1$, mean $\mu \in \mathbb{R}^k$ and a symmetric positive definite shape matrix $\Delta \in \mathbb{R}^{k\times k}$, if Z has a pdf p with

$$p(z) = \frac{\Gamma(\frac{\nu+k}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\det(\nu\pi\Delta)}}\left(1+\frac{1}{\nu}(z-\mu)^T\Delta^{-1}(z-\mu)\right)^{-\frac{\nu+k}{2}}.$$

Remark A.3
The covariance matrix Σ of a multivariate Student's t distribution is $\frac{\nu}{\nu-2}\Delta$ for $\nu > 2$ and else undefined. As ν goes to ∞, the multivariate normal distribution is recovered.

A.3 Proofs, Lemmas and Remarks

Proof of Proposition 2.5 The first part of assertion *i)* results from the linearity of the expectation, the fact that the copies Z^i are i.i.d. according to $q\lambda$, the properties of the importance density q and the equality $\mathbb{E}\big[I_\xi(Z)\big] = P_\xi$. More precisely, it holds

$$\mathbb{E}\big[\widehat{P}_\xi^{\mathrm{IS}}\big] = \mathbb{E}_q\left[\frac{1}{N}\sum_{i=1}^{N} I_\xi(Z^i)\frac{p(Z^i)}{q(Z^i)}\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_q\left[I_\xi(Z^i)\frac{p(Z^i)}{q(Z^i)}\right] = P_\xi.$$

The calculation of the variance is even more direct and just uses the calculation rules for the variance and that the copies Z^i are i.i.d. according to $q\lambda$, i.e.

$$\mathrm{Var}\big[\widehat{P}_\xi^{\mathrm{IS}}\big] = \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{Var}_q\left[I_\xi(Z^i)\frac{p(Z^i)}{q(Z^i)}\right] = \frac{1}{N}\mathrm{Var}_q\left[I_\xi(Z)\frac{p(Z)}{q(Z)}\right],$$

for some $Z \sim q\lambda$.
Convergence $\mathbb{P}$-a.s. in *ii)* is under the given assumptions a direct result of the L^1-version of the strong law of large numbers, after an application of the Steiner translation theorem and the fact that $L^2(\Omega,\mathcal{F},\mathbb{P}) \subset L^1(\Omega,\mathcal{F},\mathbb{P})$.
In order to show the minimization property in *iii)*, consider an arbitrary, but feasible, density q. It suffices to show that q_* minimizes $\mathrm{Var}_q\big[I_\xi(Z)\frac{p(Z)}{q(Z)}\big]$. Therefore

$$\begin{aligned}
&\mathrm{Var}_{q_*}\left[\frac{I_\xi(Z)p(Z)}{q_*(Z)}\right] + \left(\mathbb{E}_{q_*}\left[\frac{I_\xi(Z)p(Z)}{q_*(Z)}\right]\right)^2 = \mathbb{E}_{q_*}\left[\left(\frac{I_\xi(Z)p(Z)}{q_*(Z)}\right)^2\right] \\
&\quad= \mathbb{E}_{q_*}\left[\left(\frac{I_\xi(Z)p(Z)}{I_\xi(Z)\frac{p(Z)}{P_\xi}}\right)^2\right] = P_\xi^2 = \Big(\mathbb{E}_p\big[I_\xi(Z)\big]\Big)^2 = \left(\mathbb{E}_q\left[\frac{I_\xi(Z)p(Z)}{q(Z)}\right]\right)^2 \\
&\quad\overset{\text{CSI}}{\leq} \mathbb{E}_q\left[\left(\frac{I_\xi(Z)p(Z)}{q(Z)}\right)^2\right] = \mathrm{Var}_q\left[\frac{I_\xi(Z)p(Z)}{q(Z)}\right] + \left(\mathbb{E}_q\left[\frac{I_\xi(Z)p(Z)}{q(Z)}\right]\right)^2 \\
&\quad= \mathrm{Var}_q\left[\frac{I_\xi(Z)p(Z)}{q(Z)}\right] + \left(\mathbb{E}_{q_*}\left[\frac{I_\xi(Z)p(Z)}{q_*(Z)}\right]\right)^2,
\end{aligned}$$

where the first and the next-to-last step are an application of Steiners translation theorem. The last equality follows directly by writing out the notation of $\mathbb{E}_q$ and $\mathbb{E}_{q_*}$. The claim follows by subtracting $\big(\mathbb{E}_{q_*}\big[\frac{I_\xi(Z)p(Z)}{q_*(Z)}\big]\big)^2$ from both sides.
In order to show $\mathrm{Var}_{q_*}\big[\widehat{P}_\xi^{\mathrm{IS}}\big] = 0$ and therefore *iv)* it again suffices to show

$$\mathrm{Var}_{q_*}\left[I_\xi(Z)\frac{p(Z)}{q_*(Z)}\right] = 0.$$

Using again the notation $Q = \operatorname{supp}(q_*)$ the calculation follows

$$\begin{aligned}
\operatorname{Var}_{q_*}\left[I_\xi(Z)\frac{p(Z)}{q_*(Z)}\right] &= \mathbb{E}_{q_*}\left[\left(I_\xi(Z)\frac{p(Z)}{q_*(Z)}\right)^2\right] - \left(\mathbb{E}_{q_*}\left[I_\xi(Z)\frac{p(Z)}{q_*(Z)}\right]\right)^2 \\
&= \int_Q I_\xi^2(z)\frac{p^2(z)}{{q_*}^2(z)}q_*(z)\,\mathrm{d}\lambda(z) - \left(\int_Q I_\xi(z)\frac{p(z)}{q_*(z)}q_*(z)\,\mathrm{d}\lambda(z)\right)^2 \\
&= \int_Q I_\xi^2(z)\frac{p(z)}{q_*(z)}p(z)\,\mathrm{d}\lambda(z) - \left(\int_Q I_\xi(z)p(z)\,\mathrm{d}\lambda(z)\right)^2 \\
&= \int_Q I_\xi^2(z)\frac{p(z)}{I_\xi(z)\frac{p(z)}{P_\xi}}p(z)\,\mathrm{d}\lambda(z) - \left(\int_Q I_\xi(z)p(z)\,\mathrm{d}\lambda(z)\right)^2 \\
&= P_\xi\int_Q I_\xi(z)p(z)\,\mathrm{d}\lambda(z) - \left(\int_Q I_\xi(z)p(z)\,\mathrm{d}\lambda(z)\right)^2 \\
&= P_\xi \cdot P_\xi - P_\xi^2 = 0,
\end{aligned}$$

where the first equality is the definition of the variance, the second arises by exploiting the expectation and the fact that the region $\operatorname{supp}(q_*)^c$ does not contribute to the result. Since it is operated on $\operatorname{supp}(q_*)$ the third equality is clear and the fourth is just plugging in the definition of q_*. The fifth is well-defined due to $\operatorname{supp}(I_\xi \cdot p) \subset \operatorname{supp}(q_*)$. □

Remark A.4
Clearly, the zero variance property implies the variance minimization. But both proofs are worth to be mentioned, since in a more general context, where some integral $\int_\Theta \phi(z)p(z)\,\mathrm{d}\lambda(z)$ shall be evaluated using importance sampling, a variance-minimizing density q_* can be found as well, namely

$$q_*(z) = \frac{|\phi(z)|p(z)}{\int_\Theta |\phi(z)|p(z)\,\mathrm{d}\lambda(z)}.$$

A quick verification shows, that the proof of the minimization property holds as well. The zero variance property proof gets stuck, as soon as ϕ takes negative values, since then $\phi^2(z)/|\phi(z)| = \phi(z)$ is no longer valid.

Lemma A.5
The integral

$$I = \int_{-\infty}^{\xi} \frac{\sigma}{\sqrt{2\pi}} \exp\left(-z^2 + \frac{(z-\mu)^2}{2\sigma^2}\right) d\lambda(z)$$

is finite iff one of the two cases

i) $\sigma > 1/\sqrt{2}$, *or*

ii) $\sigma = 1/\sqrt{2}$ *and* $\mu < 0$

holds.

Proof. For convenience, the notation $l \doteq r$ is introduced, meaning that the left-hand side l is bounded iff the right-hand side r is.
For $\sigma = 1/\sqrt{2}$ one obtains

$$I \doteq \int_{-\infty}^{\xi} \exp(-2z\mu + \mu^2)\,\mathrm{d}\lambda(z) \doteq \int_{-\infty}^{\xi} \exp(-2z\mu)\,\mathrm{d}\lambda(z) = \begin{cases} \frac{\exp(-2\xi\mu)}{-2\mu} & \text{if } \mu < 0 \\ \infty & \text{if } \mu \geq 0 \end{cases}$$

Elsewise for $\sigma \neq 1/\sqrt{2}$:

$$\begin{aligned}
I &\doteq \int_{-\infty}^{\xi} \exp\left(\frac{1-2\sigma^2}{2\sigma^2}z^2 - \frac{z\mu}{\sigma^2} + \frac{\mu^2}{2\sigma^2}\right) \mathrm{d}\lambda(z) \\
&\doteq \int_{-\infty}^{\xi} \exp\left(\frac{1-2\sigma^2}{2\sigma^2}z^2 - \frac{z\mu}{\sigma^2} + \frac{\mu^2}{2\sigma^2(1-2\sigma^2)}\right) \mathrm{d}\lambda(z) \\
&= \int_{-\infty}^{\xi} \exp\left(\frac{1-2\sigma^2}{2\sigma^2}\left(z^2 - \frac{2z\mu}{1-2\sigma^2} + \frac{\mu^2}{(1-2\sigma^2)^2}\right)\right) \mathrm{d}\lambda(z) \\
&= \int_{-\infty}^{\xi} \exp\left(\frac{1-2\sigma^2}{2\sigma^2}\left(z - \frac{\mu}{1-2\sigma^2}\right)^2\right) \mathrm{d}\lambda(z) \\
&= \int_{-\infty}^{\xi-\frac{\mu}{1-2\sigma^2}} \exp\left(\frac{1-2\sigma^2}{2\sigma^2}z^2\right) \mathrm{d}\lambda(z) \\
&\doteq \int_{-\infty}^{0} \exp\left(\frac{1-2\sigma^2}{2\sigma^2}z^2\right) \mathrm{d}\lambda(z) \\
&\doteq \int_{-\infty}^{\infty} \exp\left(\frac{1-2\sigma^2}{2\sigma^2}z^2\right) \mathrm{d}\lambda(z),
\end{aligned}$$

where the non-trivial steps are reasoned as follows: Constant factors were exchanged from the first to second line in order to complete the squares, before the fifth line performs a change of variables. Directly afterwards the extreme value theorem is used on the compact set $\left[\xi - \frac{\mu}{1-2\sigma^2}, 0\right]$ or $\left[0, \xi - \frac{\mu}{1-2\sigma^2}\right]$, depending on the sign of the integration limit. Finally symmetry of the integrand enters.
The remaining integral is typically computed using a trick going back to Poisson, consisting of squaring the integral and applying a polar coordinate transformation. This results for $\frac{1-2\sigma^2}{2\sigma^2} < 0$ in

$$\int_{-\infty}^{\infty} \exp\left(\frac{1-2\sigma^2}{2\sigma^2}z^2\right) \mathrm{d}\lambda(z) = \sqrt{\frac{2\pi\sigma^2}{2\sigma^2-1}}.$$

For $\frac{1-2\sigma^2}{2\sigma^2} > 0$ the integral is obviously unbounded and the case $\frac{1-2\sigma^2}{2\sigma^2} = 0$ was discussed above.
Thus it can be concluded: The integral I is finite iff

$$\sigma > \frac{1}{\sqrt{2}} \quad \text{or} \quad \sigma = \frac{1}{\sqrt{2}} \text{ and } \mu < 0. \qquad \square$$

Remark A.6
The second case *ii)* in Lemma A.5 is, from a numerical point of view, rather uninteresting, since $\sigma = 1/\sqrt{2}$ is computationally not realizable.

Lemma A.7
The integral

$$I = \int_{-\infty}^{\xi} \frac{\left(\Gamma(\frac{\nu}{2})\sqrt{\nu\delta^2}\right)}{\left(2\sqrt{\pi}\Gamma(\frac{\nu+1}{2})\right)} e^{-z^2} \left(1 + \frac{(z-\mu)^2}{\nu\delta^2}\right)^{\frac{\nu+1}{2}} \mathrm{d}\lambda(z)$$

is bounded for any combination of μ, $\delta^2 > 0$ *and* $\nu > 1$.

Proof. For convenience, the notation $l \doteq r$ is introduced, meaning that the left-hand side l is bounded iff the right-hand side r is.

It holds

$$
\begin{aligned}
I &\doteq \int_{-\infty}^{\xi} e^{-z^2} \left(1 + \frac{(z-\mu)^2}{\nu\delta^2}\right)^{\frac{\nu+1}{2}} \mathrm{d}\lambda(z) \\
&\leq \int_{-\infty}^{\xi} e^{-z^2} \left(1 + \frac{(z-\mu)^2}{\nu\delta^2}\right)^{\left\lceil\frac{\nu+1}{2}\right\rceil} \mathrm{d}\lambda(z) \\
&= \int_{-\infty}^{\xi} e^{-z^2} \sum_{j=0}^{\left\lceil\frac{\nu+1}{2}\right\rceil} \binom{\left\lceil\frac{\nu+1}{2}\right\rceil}{j} \frac{(z-\mu)^{2j}}{\nu^j\delta^{2j}} \mathrm{d}\lambda(z) \\
&= \int_{-\infty}^{\xi} e^{-z^2} \sum_{j=0}^{\left\lceil\frac{\nu+1}{2}\right\rceil} \binom{\left\lceil\frac{\nu+1}{2}\right\rceil}{j} \frac{1}{\nu^j\delta^{2j}} \sum_{i=0}^{2j} \binom{2j}{i} z^{2j-i}(-\mu)^i \mathrm{d}\lambda(z) \\
&\doteq \sum_{j=0}^{\left\lceil\frac{\nu+1}{2}\right\rceil} \sum_{i=0}^{2j} \int_{-\infty}^{\xi} e^{-z^2} z^{2j-i} \mathrm{d}\lambda(z) \\
&\doteq \sum_{j=0}^{\left\lceil\frac{\nu+1}{2}\right\rceil} \sum_{i=0}^{2j} \int_{-\infty}^{\xi} e^{-z^2} z^{2j-i} \mathrm{d}\lambda(z) \\
&\leq \sum_{j=0}^{\left\lceil\frac{\nu+1}{2}\right\rceil} \sum_{i=0}^{2j} \int_{-\infty}^{\infty} e^{-z^2} |z|^{2j-i} \mathrm{d}\lambda(z) < \infty,
\end{aligned}
$$

having used, that the prefactor is constant for fixed ν and δ^2 in the first line and the fact that $\int_{-\infty}^{\infty} e^{-z^2}|z|^{\iota}\,\mathrm{d}\lambda(z) < \infty$ for any $\iota \in \mathbb{N}$, what can be checked by repeated partial integration. Consequently, in any case, the integral is bounded. □

Lemma A.8
A natural way to quantify the self-information of a message m that occurs with probability $\varrho(m)$ is described by

$$I(m) = \log\left(\frac{1}{\varrho(m)}\right).$$

Proof. Consider at first one very special case, namely that the content of a message is totally known before, i.e. there is no uncertainty at all. Actually, then the message does not transport any kind of information. Otherwise, if there is uncertainty and the message is not known priorly, the message carries information. It is natural to expect a large amount of self-information if the uncertainty is large and vice versa.
This reasons the approach $I(m) = f\big(\varrho(m)\big)$ for some message m and a to be determined function f. Besides the already addressed properties $\varrho(m) = 1 \Rightarrow I(m) = 0$ and $\varrho(m) < 1 \Rightarrow I(m) > 0$, there is one further fundamental requirement. Note therefore, that messages are in this context events. Let m be a message, which can be decomposed in a way that $m = m_1 \cap m_2$, where m_1 and m_2 are independent from one another. As m conveys the information of both messages m_1 and m_2 it is very reasonable to presuppose

$$I(m) = I(m_1) + I(m_2).$$

Furthermore, by independence of the two messages it follows $\varrho(m) = \varrho(m_1)\varrho(m_2)$ and consequently

$$I(m) = f\big(\varrho(m)\big) = f\big(\varrho(m_1)\varrho(m_2)\big).$$

Combining the two previous statements one arrives at

$$f\big(\varrho(m_1)\big) + f\big(\varrho(m_2)\big) = I(m) = f\big(\varrho(m_1)\varrho(m_2)\big).$$

This functional equation is solved (non-trivially) by the family of logarithmic functions to arbitrary bases. The selection of the base is problem-dependent and typically 2 if related to bits or e in the case of nats. A further freedom is to choose between log and $-\log$. As $\varrho(m) \leq 1$ and $I(m) \geq 0$ the minus sign has to be selected.
Consequently

$$I(m) = -\log\big(\varrho(m)\big) = \log\left(\frac{1}{\varrho(m)}\right)$$

for some base. □

Lemma A.9 (Completion of the proof of Proposition 3.5)
Let $\mathcal{R}(Q_\xi) \subset \{0,1\}$ and analogously for $Q_{\xi,\ell-1}$. Then

$$\big|\mathbb{E}[Q_{\xi,\ell}Q_{\xi,\ell-1} - Q_\xi]\big| \leq 2\big(\big|\mathbb{E}[Q_{\xi,\ell} - Q_\xi]\big| + \big|\mathbb{E}[Q_{\xi,\ell-1} - Q_\xi]\big|\big).$$

Proof. Starting directly with the integral representation of the left-hand side, one obtains

$$\begin{aligned}
\left|\int_\Omega Q_{\xi,\ell}Q_{\xi,\ell-1} - Q_\xi \,\mathrm{d}\mathbb{P}\right| &\leq \left|\int_\Omega Q_{\xi,\ell-1}(Q_{\xi,\ell} - Q_\xi)\,\mathrm{d}\mathbb{P}\right| + \left|\int_\Omega Q_\xi(Q_{\xi,\ell-1} - Q_\xi)\,\mathrm{d}\mathbb{P}\right| \\
&\leq \left|\int_{\{Q_{\xi,\ell}-Q_\xi\geq 0\}} Q_{\xi,\ell-1}(Q_{\xi,\ell} - Q_\xi)\,\mathrm{d}\mathbb{P}\right| + \left|\int_{\{Q_{\xi,\ell}-Q_\xi< 0\}} Q_{\xi,\ell-1}(Q_{\xi,\ell} - Q_\xi)\,\mathrm{d}\mathbb{P}\right| \\
&\quad + \left|\int_{\{Q_{\xi,\ell-1}-Q_\xi\geq 0\}} Q_\xi(Q_{\xi,\ell-1} - Q_\xi)\,\mathrm{d}\mathbb{P}\right| + \left|\int_{\{Q_{\xi,\ell-1}-Q_\xi< 0\}} Q_\xi(Q_{\xi,\ell-1} - Q_\xi)\,\mathrm{d}\mathbb{P}\right| \\
&\leq \left|\int_{\{Q_{\xi,\ell}-Q_\xi\geq 0\}} Q_{\xi,\ell-1}(Q_{\xi,\ell} - Q_\xi)\,\mathrm{d}\mathbb{P}\right| + \left|\int_{\{Q_\xi-Q_{\xi,\ell}> 0\}} Q_{\xi,\ell-1}(Q_\xi - Q_{\xi,\ell})\,\mathrm{d}\mathbb{P}\right| \\
&\quad + \left|\int_{\{Q_{\xi,\ell-1}-Q_\xi\geq 0\}} Q_\xi(Q_{\xi,\ell-1} - Q_\xi)\,\mathrm{d}\mathbb{P}\right| + \left|\int_{\{Q_\xi-Q_{\xi,\ell-1}> 0\}} Q_\xi(Q_\xi - Q_{\xi,\ell-1})\,\mathrm{d}\mathbb{P}\right| \\
&\leq \big|\mathbb{E}[Q_{\xi,\ell} - Q_\xi]\big| + \big|\mathbb{E}[Q_\xi - Q_{\xi,\ell}]\big| + \big|\mathbb{E}[Q_{\xi,\ell-1} - Q_\xi]\big| + \big|\mathbb{E}[Q_\xi - Q_{\xi,\ell-1}]\big|,
\end{aligned}$$

having used the properties of the absolute value and the fact that the range of Q_ξ (and $Q_{\xi,\ell-1}$) is contained in $\{0,1\}$. □

List of Figures

List of Tables

Bibliography

[AB01] S.-K. Au and J.L. Beck. Estimation of small failure probabilities in high dimensions by subset simulation. *Prob. Eng. Mech.*, 16(1):263–277, 2001.

[AB03] S.-K. Au and J.L. Beck. Important sampling in high dimensions. *Struct. Saf.*, 25(2):139–163, 2003.

[AG07] S. Asmussen and P.W. Glynn. *Stochastic simulation: algorithms and analysis*, volume 57 of *Stochastic Modelling and Applied Probability*. Springer, New York, 2007.

[APSAS17] S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso, and A. M. Stuart. Importance sampling: intrinsic dimension and computational cost. *Statist. Sci.*, 32(3):405–431, 2017.

[AW14] S.-K. Au and Y. Wang. *Engineering Risk Assessment with Subset Simulation*. John Wiley & Sons, Inc., Singapore, 2014.

[BK08] Z.I. Botev and D.P. Kroese. An efficient algorithm for rare-event probability estimation, combinatorial optimization, and counting. *Methodol. Comput. Appl.*, 10(4):471–505, 2008.

[CGST11] K.A. Cliffe, M.B. Giles, R. Scheichl, and A.L. Teckentrup. Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients. *Comput. Visual. Sci.*, 14(1):3–15, 2011.

[dBKMR05] P.-T. de Boer, D.P. Kroese, S. Mannor, and R.Y. Rubinstein. A tutorial on the cross-entropy method. *Ann. Oper. Res.*, 134(1):19–67, 2005.

[EEHM14] D. Elfverson, D.J. Estep, F. Hellman, and A. Målqvist. Uncertainty quantification for approximate p-quantiles for physical models with stochastic inputs. *SIAM/ASA J. Uncertain. Quantif.*, 2(1):826–850, 2014.

[EHM16] D. Elfverson, F. Hellman, and A. Målqvist. A multilevel Monte Carlo method for computing failure probabilities. *SIAM/ASA J. Uncertain. Quantif.*, 4(1):312–330, 2016.

[FHMN16] F. Fagerlund, F. Hellman, A. Målqvist, and A. Niemi. Multilevel monte carlo methods for computing failure probability of porous media flow systems. *Adv. Water Resour.*, 94(1):498–509, 2016.

[Gil08] M.B. Giles. Multilevel Monte Carlo path simulation. *Oper. Res.*, 56(3):607–617, 2008.

[Gil15] M.B. Giles. Multilevel Monte Carlo methods. *Acta Numer.*, 24(1):259–328, 2015.

[GPS19] S. Geyer, I. Papaioannou, and D. Straub. Cross entropy-based importance sampling using gaussian densities revisited. *Struct. Saf.*, 76(1):15–27, 2019.

[GS91] R.G. Ghanem and P.D. Spanos. *Stochastic Finite Elements: A Spectral Approach*. Springer-Verlag, Berlin, Heidelberg, 1991.

[HdMR02] T. Homem-de Mello and R.Y. Rubinstein. Estimation of rare event probabilities using cross-entropy. In E. Yücesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, editors, *Winter Simulation Conference Proceedings*, volume 1, pages 310–319, 2002.

[Hei01] S. Heinrich. Multilevel monte carlo methods. In S. Margenov, J. Waśniewski, and P. Yalamov, editors, *Lect. Notes Comput. Sc.*, pages 58–67. Springer Berlin Heidelberg, 2001.

[KPS04] P.S. Koutsourelakis, H.J. Pradlwarter, and G.I. Schuëller. Reliability of structures in high dimensions, part I: algorithms and applications. *Prob. Eng. Mech.*, 19(4):409–417, 2004.

[KRG13] D.P. Kroese, R.Y. Rubinstein, and P.W. Glynn. The cross-entropy method for estimation. In *Machine learning: theory and applications*, volume 31 of *Handbook of Statist.*, pages 19–34. Elsevier/North-Holland, Amsterdam, 2013.

[Kul59] S. Kullback. *Information theory and statistics*. John Wiley and Sons, Inc., New York; Chapman and Hall, Ltd., London, 1959.

[LLX11] J. Li, J. Li, and D. Xiu. An efficient surrogate-based method for computing rare failure probability. *J. Comput. Phys.*, 230(24):8683–8697, 2011.

[LPS14] G.J. Lord, C.E. Powell, and T. Shardlow. *An introduction to computational stochastic PDEs*. Cambridge Texts in Applied Mathematics. Cambridge University Press, New York, 2014.

[LX10] J. Li and D. Xiu. Evaluation of failure probability via surrogate models. *J. Comput. Phys.*, 229(23):8966–8980, 2010.

[MRRT53] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, and A.H. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21(6):1087–1092, 1953.

[Owe13] A.B. Owen. Monte carlo theory, methods and examples. http://statweb.stanford.edu/~owen/mc/, 2013. Accessed on Mai 4, 2018.

[Pan03] L. Paninski. Estimation of entropy and mutual information. *Neural Comput.*, 15(6):1191–1253, June 2003.

[PCMW16] B. Peherstorfer, T. Cui, Y. Marzouk, and K. Willcox. Multifidelity importance sampling. *Comput. Method. Appl. M.*, 300(1):490–509, 2016.

[PKW18] B. Peherstorfer, B. Kramer, and K. Willcox. Multifidelity preconditioning of the cross-entropy method for rare event simulation and failure probability estimation. *SIAM/ASA J. Uncertain. Quantif.*, 6(2):737–761, 2018.

[PWG16] B. Peherstorfer, K. Willcox, and M. Gunzburger. Optimal model management for multifidelity Monte Carlo estimation. *SIAM J. Sci. Comput.*, 38(5):A3163–A3194, 2016.

[PWG18] B. Peherstorfer, K. Willcox, and M. Gunzburger. Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *ArXiv e-prints*, 2018.

[RC04] Ch.P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 2004.

[RK04] R.Y. Rubinstein and D.P. Kroese. *The cross-entropy method*. Information Science and Statistics. Springer-Verlag, New York, 2004. A unified approach to combinatorial optimization, Monte-Carlo simulation, and machine learning.

[RK17] R.Y. Rubinstein and D.P. Kroese. *Simulation and the Monte Carlo method*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, third edition, 2017.

[Rob15] Ch.P. Robert. Importance Sampling with Infinite Variance. https://xianblog.wordpress.com/2015/11/13/, 2015. Accessed on Mai 6, 2018.

[Sha48] C.E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(1):379–423, 623–656, 1948.

[SW49] C.E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. The University of Illinois Press, Urbana, Ill., 1949.

[TK10] S.T. Tokdar and R.E. Kass. Importance sampling: A review. *WiRes. Comput. Stat.*, 2(1):54–60, 2010.

[UP15] E. Ullmann and I. Papaioannou. Multilevel estimation of rare events. *SIAM/ASA J. Uncertain. Quantif.*, 3(1):922–953, 2015.

[Wol18] M.M. Wolf. Mathematical Foundations of Supervised Learning. https://www-m5.ma.tum.de/foswiki/pub/M5/Allgemeines/MA4801_2018S/ML_notes_main.pdf, 2018. Lecture Notes, Technische Universität München. Accessed on July 3, 2018.

[XK02] D. Xiu and G.E. Karniadakis. The Wiener-Askey polynomial chaos for stochastic differential equations. *SIAM J. Sci. Comput.*, 24(2):619–644, 2002.