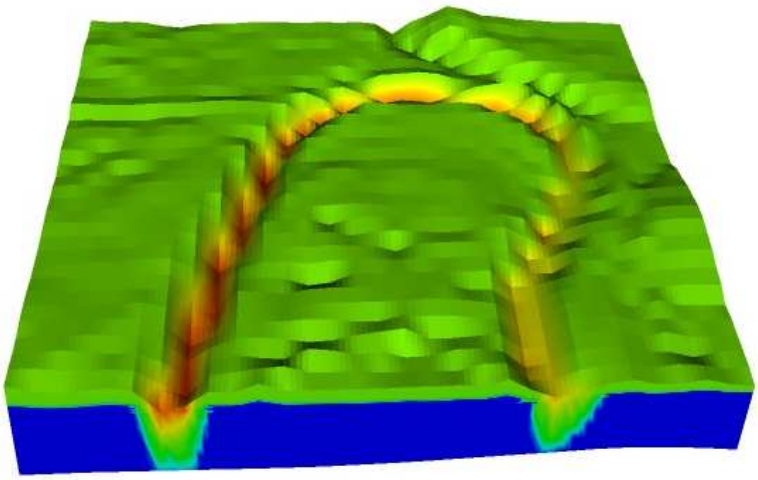
Dissertation zur Erlangung
des akademischen Grades
eines Doktors der Naturwissenschaften
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

On the Stochastic Richards Equation



vorgelegt von
Ralf Forster

Berlin 2011

Betreuer:
Prof. Dr. Ralf Kornhuber, Berlin

Gutachter:
Prof. Dr. Ralf Kornhuber, Berlin
Prof. Dr. Omar Knio, Baltimore

Tag der Disputation: 8. März 2011

Contents

Introduction

Numerical simulation of complex systems is a rapidly developing field in mathematical research with a growing number of applications in engineering, environmental sciences, medicine, and many other areas. This can be traced back to the increasing computer power on the one hand and more and more refined numerical techniques on the other hand. The main goal of these simulations is to make predictions which could reduce the costs of experiments or could be the basis for scientific or political decisions. In light of these consequences, the question of how reliable the obtained results are is of great importance. Since all simulations are, in the best case, only approximations to reality, the occurrence of errors is inevitably in some degree. As a first step, let us classify different types of errors which appear necessarily in the simulation of complex systems (according to [72]).

First, we detect *model errors*, which are rooted in the description of observed real-life processes in mathematical terms. In this stage, some terms which only contribute to a negligible extent to the system are omitted. Moreover, one is often forced to impose certain assumptions to simplify the problem in order to be able to solve it, e.g. if one reduces the dimension of a problem or assumes that certain terms are constant.

Secondly, in order to compute the simulation, one has to discretize the system and to solve it numerically, which leads to *numerical errors*. These errors are also inevitably due to the finite representation of numbers in computers. This concerns the discretization itself and also the utilization of solvers, e.g. iterative solvers which only provide a solution up to a certain tolerance. Time and computer system restrictions can limit the possibilities to pass to more and more refined discretizations or to solve up to maximal accuracy, even if it were possible theoretically.

Thirdly, we encounter *data errors*, which can concern model parameters, boundary conditions or even the geometry of the system. Often, these data are unknown or only estimated roughly since exact measurements are impossible or too costly. In many cases, a perfect knowledge of the data is not possible due to their natural variability. Eventually, the measurements themselves can be incorrect, e.g. by reason of external effects.

The first source of errors is beyond the scope considered here. Regarding the second point, there has been great advantages in the numerical mathematics to control and reduce the error arising from discretization and numerical solution over the last decades. The third error originating from uncertainties in the input data and model parameters, however, has received less attention for a long time. This has changed since the pioneering work of Ghanem and Spanos [47] about stochastic finite elements and polynomial chaos, who presented a possibility to determine the response surfaces of the solution from the stochastic modeling of the uncertain input data. This method proved to be successful for many interesting applications

in physics and engineering (see, e.g., [24] for meteorology, [40, 46, 92] for hydrology, [86] for chemistry, and in particular the recent monograph of Le Maître and Knio [72] with applications in fluid dynamics).

The idea of polynomial chaos consists in extending the solution space by further stochastic coordinates which come from the modeling of the uncertain input data and in discretizing the new space by global polynomial basis functions, while the time and space discretization for the other coordinates can be performed in the usual manner. In this way, this approach can be seen as an extension of deterministic finite element methods for partial differential equations. At the same time, it is in most cases more efficient and more accurate than other methods like, e.g., Monte Carlo or first-order methods (see the overview over alternative methods in [47] and [101]).

There have been many improvements to the original idea from Ghanem and Spanos during the last twenty years, in particular the extension to generalized polynomial chaos [114], to multi-element polynomial chaos [107], to stochastic collocation [13], and to sparse finite element schemes [20]. Apart from the variety of applications as mentioned above, the analysis of this method in a theoretical way, however, mostly concentrated on investigating the diffusion equation with a stochastic diffusion as a prototype of an elliptic partial differential equation (e.g. [11, 29, 87, 103]). To the knowledge of the author, there is in particular no application of this method to variational inequalities.

The focus of this thesis is the application of polynomial chaos methods to the Richards equation modeling groundwater flow in order to quantify the uncertainty arising from certain stochastic model parameters. By means of the Richards equation, one can model water flows in saturated and in unsaturated porous media, where it is parabolic in the unsaturated and elliptic in the saturated regime. The main difficulty consists in the nonlinearities, since the saturation and the hydraulic conductivity are both dependent on the pressure. Moreover, the saturation appears in the time derivative, while the hydraulic conductivity is a factor in the spatial derivative.

Berninger [18] recently presented a method to solve this problem without any linearization techniques by applying a Kirchhoff transformation and an implicit–explicit time discretization to obtain an *elliptic variational inequality of second kind* (see [48, Chapter 1] for the terminology) with a convex, lower semicontinuous functional. This variational inequality can be seen equivalently as a *convex minimization problem*, which in turn allows the proof of existence results for the solution and of convergence results for an appropriate finite element discretization. Moreover, monotone multigrid methods [68] providing fast and reliable solutions of the system are applicable and, since no linearization is involved, turn out to be robust with respect to varying soil parameters.

Let us now consider the case that some parameters in the Richards equation are uncertain and consequently modeled as random functions, e.g. the permeability of the soil, the boundary conditions or source functions. These random functions should have a positive correlation length, which means that stochastic partial differential equations (SPDEs) with white noise like in [58] are not considered here. Our ultimate goal is to quantify the uncertainty of the solution of the Richards equation, viz. of the water pressure, in dependence of the uncertainty of the input parameters such that we are able to answer the following questions:

(Q1) What is the average pressure in the domain, and how is it developing in time?

(Q2) In which areas of the soil do we encounter the largest uncertainty?

(Q3) What is the probability that the soil is saturated after one minute/ one hour/one day at special points in the domain?

(Q4) What is the probability that the pressure exceeds a fixed threshold value within a predetermined time?

To achieve this goal, we first perform a spectral decomposition of the correlation operator of the random input functions (the so-called Karhunen–Loève decomposition) and set up the stochastic solution space. Then, proceeding with Kirchhoff transformation and time discretization as above, we end up with the stochastic Richards equation over a space of tensor products consisting of functions on the computational domain, which will be discretized by finite elements, and functions in the stochastic space, which will be discretized by polynomial chaos.

At this point, the difficulties caused by the nonlinearities become apparent. First, the Kirchhoff transformation yields a lower obstacle condition for the transformed pressure, which is against the nature of global polynomials tending to infinity as the argument becomes larger. Secondly, the convex functional, which has the form of a superposition operator, needs a special discretization and moreover couples all coefficients of the polynomial chaos basis. This means that the solution of the discretized problem in a Galerkin approach cannot be obtained by the reduction to one-dimensional minimization problems by a splitting into subspaces as it is possible in the deterministic case. As will be shown in Chapter 4, this can be achieved nevertheless by an appropriate basis transformation, which connects the stochastic Galerkin approach to stochastic collocation methods. With this connection, we are able to solve the large systems in a robust way with one-dimensional minimization techniques. Moreover, this allows us to extend this method to a multigrid algorithm to speed up the convergence. Note that the theoretical results and techniques developed for the stochastic Richards equation can be carried over to elliptic variational inequalities of second kind with stochastic parameters.

The outline of this thesis is as follows. We start in Chapter 1 with the description of the deterministic Richards equation and the parameters in the hydrological model. We introduce the Kirchhoff transformation and study for deteriorating model parameters the form of the nonlinear parameter functions which specify the saturation and the part of the hydraulic conductivity which depends on the pressure. This is important in the following chapters, since the transformed saturation arises in the derivative of the convex functional later on. Section 1.2 is devoted to the weak formulation of the Richards equation, the reformulation as a variational inequality of second kind, and an overview over known results for the existence and uniqueness of a solution.

Chapter 2 comprises the specification of the input noise as random fields and the derivation of the stochastic Richards equation, in the pointwise formulation as well as in the weak formulation. This requires the definition of tensor spaces and of operators thereon, which is provided in Section 2.2. The main result of this chapter is the insight that the conditions on the existence of solutions for the deterministic Richards equation are sufficient to show the measurability and the integrability of the solution of the stochastic Richards equation if the permeability is modeled as random field.

The main chapter 3 is devoted to the discretization of the stochastic Richards equation. Note that in addition to the time discretization by backward Euler and the spatial discretization by means of finite elements, which are known from the

deterministic problem, we have to consider the approximation of the stochastic input parameters by the Karhunen–Loève expansion and the discretization by polynomial chaos. The Karhunen–Loève expansion is introduced in Section 3.1, and it turns out that the independence of the random variables in this series is of major importance for the following conclusions. Moreover, this expansion introduces a new stochastic space, which is embedded in $\mathbb{R}^M$ and in which the polynomial chaos basis will be established. Beforehand, we investigate the time-discretized stochastic Richards equation after Karhunen–Loève approximation, reformulate it in terms of convex minimization and variational inclusions and show the existence of a unique solution.

The aforementioned discretization in the new stochastic domain is carried out in Section 3.3, where we describe different polynomial chaos schemes with corresponding quadrature formulas. The convex functional needs special attention and we will find a consistent approximation by means of Gaussian quadrature. At the end of this chapter, we turn to the convergence of the polynomial chaos and finite element discretization. We proof the convergence for limit cases, where the problem degenerates to a stochastic obstacle problem, and perform numerical experiments to determine the discretization error.

The numerical solution of this discretized minimization problem is presented in the second main chapter 4. The underlying idea is a successive minimization of the energy in direction of the nodal basis functions in the spatial domain. Due to the dependence on the stochastic basis functions, this turns out to be a Block Gauß–Seidel method, which necessitates an inner minimization within each stochastic block, which is complicated by the fact that the nonlinearity in the convex functional couples all polynomial chaos basis functions. We present a transformation which yields a decoupling within each block and which connects the stochastic Galerkin approach with tensor product polynomial chaos bases to a stochastic collocation method. This reduces the whole problem to a successive minimization in one-dimensional subspaces, and we can prove global convergence for this method. The same idea allows us to establish a multigrid solver in order to accelerate the convergence, where this Block Gauß–Seidel method acts as a fine grid smoother and a Newton linearization is used for the coarse grid corrections. We conclude this chapter by a comparison of our approach with a pure stochastic collocation approach with regard to the convergence rates.

In the final chapter 5, we apply our results to a hydrological problem. We perform our computations on a realistic geometry and explain which further extensions in the spatial solver are advised to treat the anisotropy of the grid. Then, using realistic parameter functions with a lognormally distributed permeability function featuring an exponential covariance, we simulate the infiltration of water from a river into an initially dry soil and demonstrate how the questions (Q1)–(Q4) can be answered on the basis of the results.

Finally, we sum up some facts about the exponential covariance operator in the Karhunen–Loève expansion, some important polynomial sets, and Gaussian quadrature in the appendix.

Acknowledgement. It is a pleasure to thank all the people who supported me during the preparation of this work. First of all, I would like to express my deep gratitude to my supervisor, Prof. Dr. Ralf Kornhuber, for his continuous support, his patience, and his guidance throughout the last years. I also would like to thank Prof. Dr. Omar Knio and Dr. Heiko Berninger, whose expertise helped me in my first steps in this research area. My special thanks go to Dr. Oliver Sander and Carsten Gräser for their invaluable help in struggling with the DUNE code. Moreover, Carsten Gräser provided many clever ideas in convex minimization.
Finally, my deepest gratitude belongs to Luise for her love and for just showing me how wonderful life is.

Chapter 1

Deterministic Richards equation in homogeneous soil

In this introductory chapter 1, we state the Richards equation in its hydrological setting. We explain the parameter functions used throughout this thesis. The first important step in our considerations is done in Section 1.1, where we introduce the Kirchhoff transformation. Its application will allow us to write the Richards equation in a form which is more suitable for the analytical and numerical treatment in the following chapters. Section 1.1 is also devoted to the study of the occurring parameter functions and its limit cases. Afterwards in Section 1.2, we derive the weak formulation of the equation and define boundary conditions. The presentation mainly follows the work of Berninger [18], whereas the facts from hydrology can be read in the book of Bear [15].

1.1 Basic tools in hydrology and Kirchhoff transformation

We consider a computational domain $D \subset \mathbb{R}^3$ and use a coordinate system with points $x = (x_1, x_2, z) \in D$, where we assume that the z-axis points downwards in the direction of gravity such that the relation between the *piezometric head* h and the pressure of the water at x is given by the formula

$$h = \frac{p}{\rho g} - z. \tag{1.1.1}$$

Here, $p = p_w - p_a$ denotes the difference between the pressure of water and the constant atmospheric pressure, while ρ is the *density* of the water and g the *gravitational constant*.

The law of Darcy (first published in [28, p. 576]) states

$$\mathbf{v} = -K_c \nabla h \tag{1.1.2}$$

for the water flux $\mathbf{v}$ through a porous medium at any time t. The coefficient K_c is called *hydraulic conductivity* and is a scalar function if the flow takes place in an isotropic medium and a symmetric positive definite 3×3-matrix for any $x \in D$ in general. In unsaturated soils, it is moreover dependent on the *saturation* $\theta(\cdot)$. This is a function $p \mapsto \theta(p)$ which is monotonically increasing between a minimal

saturation θ_m and a maximal saturation θ_M, where we have $\theta_m > 0$ due to residual water and $\theta_M < 1$ due to residual air in the soil and where the saturation is maximal for hydrostatic pressures $p \geq 0$. Using the symbol η for the *viscosity* of the water, we now split up the hydraulic conductivity according to

$$K_c(x, \theta) = K(x)\, \eta^{-1} \rho g\, kr(\theta) \tag{1.1.3}$$

in the *permeability* $K(\cdot)$ of the soil as a function on D, which is no longer dependent on the fluid, and the function $kr(\cdot)$ which is called *relative permeability*. This is a monotonically increasing function of the saturation θ which assumes values in the range $[0, 1]$.

The work of Richards [91] combined the law of Darcy (1.1.2) with the principle of mass conservation; he stated the continuity equation

$$\mathfrak{n}\theta_t + \operatorname{div} \mathbf{v} = f, \tag{1.1.4}$$

where the function $\mathfrak{n} = \mathfrak{n}(x)$ denotes the *porosity* of the soil and f is a source function (often set equal to zero). Inserting Darcy's law (1.1.2) into (1.1.4) and applying (1.1.1) and (1.1.3), we arrive at the *Richards equation*

$$\mathfrak{n}\theta(p)_t - \operatorname{div}\Big(K\, \eta^{-1} kr(\theta(p)) \nabla\big(p - \rho g z\big)\Big) = f \tag{1.1.5}$$

for the unknown pressure function p on $(0, T) \times D$ with $T > 0$, in which

$$\mathbf{v} = -K\, \eta^{-1} kr(\theta(p)) \nabla\big(p - \rho g z\big) \tag{1.1.6}$$

is the water flux. Since we assume in this thesis that $kr(\theta(p))$ is always positive, it is obvious that the Richards equation is a quasilinear elliptic-parabolic equation which is of elliptic type where the soil is fully saturated and parabolic in the unsaturated regime.

There are different ways how to obtain concrete analytical versions of the parameter functions $p \mapsto \theta(p)$ and $\theta \mapsto kr(\theta)$, and we restrict ourselves in the following to the Brooks–Corey model (see [45] for a discussion of different models). With a soil dependent parameter $\lambda > 0$, which is called the *pore size distribution factor*, and the definition $e(\lambda) := 3 + \frac{2}{\lambda}$, it states

$$kr(\theta(p)) = \left[\frac{p}{p_b}\right]^{-\lambda e(\lambda)} := \begin{cases} \left(\frac{p}{p_b}\right)^{-\lambda e(\lambda)} & \text{for } \; p \leq p_b \\ 1 & \text{for } \; p \geq p_b \end{cases} \tag{1.1.7}$$

for the relative permeability and

$$\theta(p) = \theta_m + (\theta_M - \theta_m) \left[\frac{p}{p_b}\right]^{-\lambda} \tag{1.1.8}$$

for the saturation. The value $p_b < 0$ is the so-called *bubbling pressure*, which indicates the pressure which is necessary to allow air bubbles to enter an originally fully saturated soil. Observe that the Brooks–Corey functions are not differentiable in this point. Figures 1.1 and 1.2 show some typical graphs of these functions.

Analyzing (1.1.5), we detect the nonlinearities in the time derivative and in the spatial derivatives. As a first step, it would be expedient to eliminate the relative permeability $kr(\cdot)$ in front of the gradient. This can be done by means of the Kirchhoff transformation, see Alt and Luckhaus [5]. In our case, this transformation $\kappa : \mathbb{R} \to \mathbb{R}$ is defined as

$$\kappa : p \mapsto u := \int_0^p kr(\theta(q))\, \mathrm{d}q, \tag{1.1.9}$$

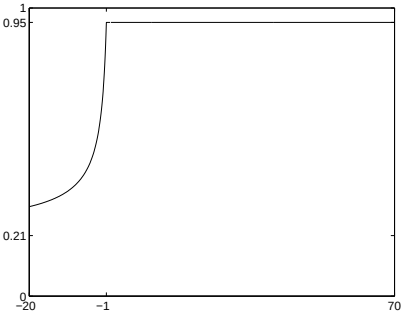


Figure 1.1: $p \mapsto \theta(p)$

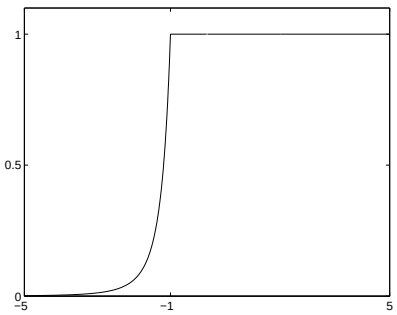


Figure 1.2: $p \mapsto kr(\theta(p))$

and the new variable u is said to be the *generalized pressure*. The saturation as a function of u is denoted by

$$H(u) := \theta(\kappa^{-1}(u)), \tag{1.1.10}$$

and the transformed Richards equation (1.1.5) then reads

$$\mathfrak{n} H(u)_t - \operatorname{div}\Big(K\,\eta^{-1}\big(\nabla u - kr(H(u))\rho g \nabla z\big)\Big) = f. \tag{1.1.11}$$

This can be seen by the chain rule $\nabla u = kr(\theta(p))\nabla p$ if all terms are differentiable; otherwise we refer to [18, Chapter 1.5]. Note that (1.1.11) is now a semilinear equation.

Our choice of the parameter functions in (1.1.7) according to Brooks and Corey comes in handy in combination with the Kirchhoff transformation because the transformation itself as well as its inverse can be given in a closed form by means of (1.1.7) and (1.1.9), viz.

$$u = \kappa(p) = \begin{cases} \frac{p_b}{-\lambda e(\lambda)+1}\left(\frac{p}{p_b}\right)^{-\lambda e(\lambda)+1} + \frac{-\lambda e(\lambda) p_b}{-\lambda e(\lambda)+1} & \text{for } \; p \le p_b \\ p & \text{for } p \ge p_b \end{cases} \tag{1.1.12}$$

and

$$p = \kappa^{-1}(u) = \begin{cases} p_b\left(\frac{u(-\lambda e(\lambda)+1)}{p_b} + \lambda e(\lambda)\right)^{\frac{1}{-\lambda e(\lambda)+1}} & \text{for } \; u_c < u \le p_b \\ u & \text{for } \; u \ge p_b. \end{cases} \tag{1.1.13}$$

Observe that the generalized pressure and the physical pressure are equal in case of full saturation, whereas, in the unsaturated regime, the interval $(-\infty, p_b)$ of the physical pressure corresponds to the bounded interval (u_c, p_b) for the generalized pressure; this lower bound

$$u_c := \frac{\lambda e(\lambda)}{\lambda e(\lambda) - 1}\, p_b \tag{1.1.14}$$

is called the *critical generalized pressure*, and we have $u_c < p_b$. Then, the transformed functions in (1.1.11) are given by

$$\begin{aligned} H(u) &= \theta(\kappa^{-1}(u)) \\ &= \begin{cases} \theta_m + (\theta_M - \theta_m)\left(\frac{u(-\lambda e(\lambda)+1)}{p_b} + \lambda e(\lambda)\right)^{\frac{\lambda}{\lambda e(\lambda)-1}} & \text{for } \; u_c < u \le p_b \\ \theta_M & \text{for } \; u \ge p_b \end{cases} \end{aligned} \tag{1.1.15}$$

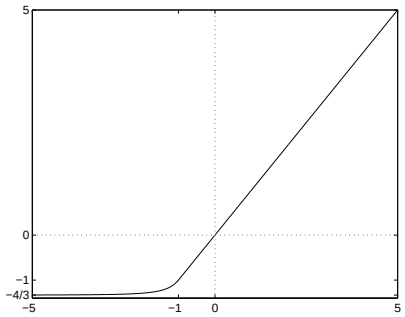


Figure 1.3: $p \mapsto \kappa(p) = u$

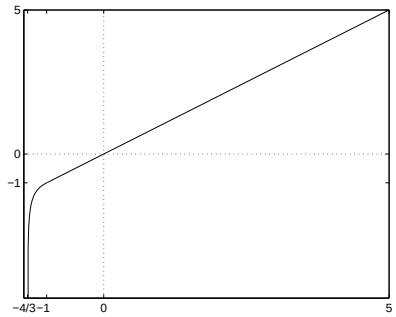


Figure 1.4: $u \mapsto \kappa^{-1}(u) = p$

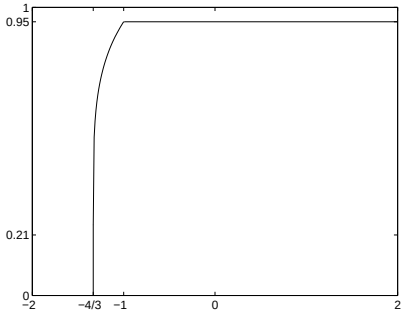


Figure 1.5: $u \mapsto H(u)$

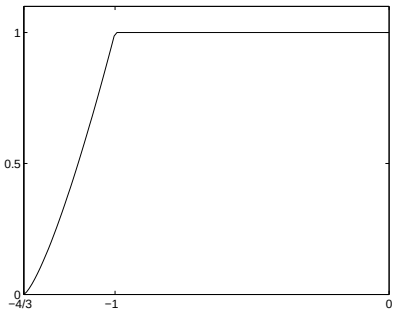


Figure 1.6: $u \mapsto kr(H(u))$

for the saturation as a function of the generalized pressure u, where $H(u) \to \theta_m$ as $u \downarrow u_c$, and

$$\begin{aligned} kr(H(u)) &= \left[\frac{\kappa^{-1}(u)}{p_b}\right]^{-\lambda e(\lambda)} \\ &= \begin{cases} \left(\frac{u(-\lambda e(\lambda)+1)}{p_b} + \lambda e(\lambda)\right)^{\frac{\lambda e(\lambda)}{\lambda e(\lambda)-1}} & \text{for } \ u_c < u \le p_b \\ 1 & \text{for } \ u \ge p_b \end{cases} \end{aligned} \tag{1.1.16}$$

for the transformed relative permeability. The antiderivative Φ of H will turn out to be important in the following. It is convex and has the form

$$\begin{aligned} \Phi(u) &= \int_0^u H(s)\,\mathrm{d}s \\ &= \begin{cases} \theta_m u + (\theta_M - \theta_m)p_b\left(1 + \frac{1}{\lambda(e(\lambda)+1)-1}\right) - \\ \frac{(\theta_M-\theta_m)p_b}{\lambda(e(\lambda)+1)-1}\left(\frac{u(-\lambda e(\lambda)+1)}{p_b} + \lambda e(\lambda)\right)^{\frac{\lambda(e(\lambda)+1)-1}{\lambda e(\lambda)-1}} & \text{for } \ u_c \le u \le p_b \\ \theta_M u & \text{for } \ u \ge p_b. \end{cases} \end{aligned} \tag{1.1.17}$$

We scale the pressure by setting $p_b = -1$. Figures 1.1–1.2 and Figures 1.5–1.6 show some realistic parameter functions according to [18, Section 1.4] with $\theta_M = 0.95$, $\theta_m = 0.21$ and $\lambda = \frac{2}{3}$, whence $u_c = -\frac{4}{3}$; on the one hand for the physical saturation and relative permeability and on the other hand for their transformed counterparts. Moreover, Figures 1.3–1.4 display the Kirchhoff transformation and its inverse.

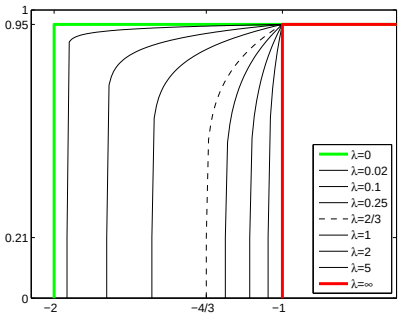


Figure 1.7: $u \mapsto H(u)$ for varying λ

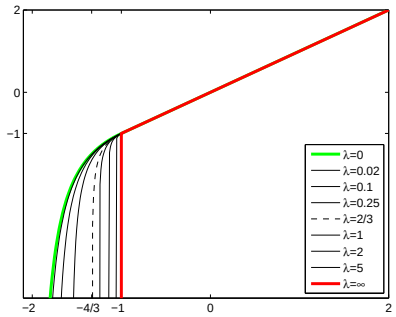


Figure 1.8: $u \mapsto \kappa^{-1}(u)$ for varying λ

Even for realistic hydrological data, the graphs of these functions already look quite steep, and we regard their behavior for deteriorating soil parameter λ. In nature, this value varies approximately between 1.0 for sand and 0.1 for clay, cf. the references in [18], but from a mathematical point of view, we are interested in λ which tends to 0 or ∞. In particular the shape of the generalized pressure H plays an important role for the numerical solution, which is why we put our focus on it. Furthermore, the inverse Kirchhoff transformation κ^{-1} is of interest since it will be applied to obtain the physical solution from the generalized solution at the end of the computation. We underline that these considerations are particularly important with respect to developing numerical solution methods which are *robust*, i.e. the convergence properties of which remain mostly unaffected by the variation of the parameter λ.

For $\lambda \to 0$, we obtain $\lambda e(\lambda) \to 2$ and $u_c \downarrow -2$ by (1.1.14). The unsaturated regime is thus represented by the interval $(-2, -1)$, and the inverse Kirchhoff transformation κ^{-1} in (1.1.13) converges pointwise to the function

$$\kappa_0^{-1} : u \mapsto \begin{cases} -(u+2)^{-1} & \text{for } -2 < u \leq -1 \\ u & \text{for } u \geq -1. \end{cases}$$

The saturation H of the generalized pressure u in (1.1.15) tends pointwise to

$$H_0 : u \mapsto \theta_M \quad \forall u \in (-2, \infty), \tag{1.1.18}$$

which we extend in accordance with H by setting

$$H_0(-2) = H_0\left(\lim_{\lambda \to 0} u_c\right) := \lim_{\lambda \to 0} H(u_c) = \theta_m.$$

An alternative view on this limit case is the observation that the graph of H turns into the monotone graph

$$u \mapsto \begin{cases} [\theta_m, \theta_M] & \text{for } u = -2 \\ \theta_M & \text{for } u > -2 \end{cases} \tag{1.1.19}$$

as $\lambda \to 0$, see the bold green line in Figure 1.7.

In the case $\lambda \to \infty$, we have $\lambda e(\lambda) \to \infty$ and $u_c \uparrow -1$ due to (1.1.14) such that the slopes of the functions κ^{-1} and H increase while the intervals $(u_c, -1)$ representing the unsaturated regime become smaller and smaller. In the limit, the unsaturated regime is vanished and the pointwise limit function H_∞ is constant, viz.

$$H_\infty : u \mapsto \theta_M \quad \forall u \in [-1, \infty). \tag{1.1.20}$$

Again, we observe the monotone graph

$$u \mapsto \begin{cases} [\theta_m, \theta_M] & \text{for } u = -1 \\ \theta_M & \text{for } u > -1 \end{cases} \tag{1.1.21}$$

as the limit of the graph of H as $\lambda \to \infty$, see the bold red line in Figure 1.7. Observe that the structure of (1.1.21) and (1.1.19) agrees. We will show in Remark 3.2.32 and Remark 4.3.8 that these limit cases are also treated by our analytical and numerical approach.

1.2 Boundary conditions and weak formulation

In this section, we specify boundary conditions for the Richards equation (1.1.11) and derive a weak formulation for this problem. Moreover, we cite results from [18] concerning the time discretization and the existence of a unique solution for this equation.

For sake of simplicity, we set $\mathfrak{n} \equiv 1$ and $\eta = 1$ in (1.1.11) and introduce $e_z := \nabla z$. In what is to come, we always assume that

$$K_{\min} \leq K(x) \leq K_{\max} \quad \forall x \in \overline{D} \tag{1.2.1}$$

is satisfied for constants $K_{\max} > K_{\min} > 0$. Furthermore, let $D \subset \mathbb{R}^d$ be an open, bounded, connected and nonempty set with a Lipschitz boundary ∂D such that the normal $\mathbf{n}$, which we assume to be directed outwards, exists almost everywhere on ∂D (cf. [26, pp. 12–14]). For each time $t \in [0, T]$, the boundary ∂D is decomposed into subsets $\Gamma_D(t)$ and $\Gamma_N(t)$. Here, $\Gamma_D(t)$ denotes the set of Dirichlet boundary conditions, which model hydrostatic pressures given by adjacent waters (e.g. rivers or lakes). The set $\Gamma_N(t)$ is corresponding to Neumann boundary conditions, which specify water flow across the boundary of D (e.g. due to precipitation) and often occur as homogeneous Neumann boundary conditions, for instance on the border of an impermeable soil.

Then, for any $t \in (0, T]$, $T > 0$, we consider the boundary value problem

$$H(u)_t - \operatorname{div}\Big(K\nabla u - K\, kr(H(u))\rho g e_z\Big) = f(t) \qquad \text{on } D \tag{1.2.2}$$

$$u = u_D(t) \qquad \text{on } \Gamma_D(t) \tag{1.2.3}$$

$$\mathbf{v} \cdot \mathbf{n} = f_N(t) \qquad \text{on } \Gamma_N(t), \tag{1.2.4}$$

where the flux $\mathbf{v}$ is defined as

$$\mathbf{v} = -(K\nabla u - K\, kr(H(u))\rho g e_z) \tag{1.2.5}$$

and where $u_D(t)$ with $u_D > u_c$ and $f_N(t)$ are given functions on $\Gamma_D(t)$ and $\Gamma_N(t)$, respectively.

The weak formulation of this boundary value problem is straightforward if the involved functions show the regularity of a classical solution. More precisely, if $H, kr : (u_c, \infty) \to \mathbb{R}$ are continuously differentiable real functions, $K \in C^1(\overline{D})$ with (1.2.1), $f(t) \in C^0(\overline{D})$, $f_N(t) \in C^0(\Gamma_N(t))$ and $u(t) \in C^2(\overline{D})$ on a C^1-polyhedron $D \subset \mathbb{R}^d$ with Hausdorff measurable ∂D, then u satisfies the boundary value problem

(1.2.2)–(1.2.4) for t if, and only if, it satisfies the variational inequality

$$\int_D H(u)_t\,(v-u)\,\mathrm{d}x + \int_D K\,\nabla u \nabla(v-u)\,\mathrm{d}x \geq \int_D K\,kr(H(u))\rho g e_z \nabla(v-u)\,\mathrm{d}x + \int_D f\,(v-u(t))\;\mathrm{d}x - \int_{\Gamma_N(t)} f_N\,(v-u)\,\mathrm{d}\sigma \quad \forall v \in \hat{\mathcal{K}}_c(t) \tag{1.2.6}$$

in the convex set

$$\hat{\mathcal{K}}_c(t) := \{w \in C^2(\overline{D}) : w(x) > u_c \;\forall x \in D \wedge w_{|\Gamma_D(t)} = u_D(t)\}, \tag{1.2.7}$$

see Proposition 1.5.3 and the following remarks in [18].

The formulation in terms of weak derivatives necessitates more effort. We recall from [3] the definitions of Sobolev spaces $H^k(D)$ as the space of all functions whose weak derivatives up to order k belong to $L^2(D)$ and of $H_0^k(D)$ as the closure of $C_0^\infty(D)$ w.r.t. $\|\cdot\|_{H^k(D)}$. At a first point, restating the assumptions on D and ∂D as explained before (1.2.2), we define the weak generalization of $\hat{\mathcal{K}}_c(t)$ for $t \in [0,T]$ as

$$\hat{\mathcal{K}}(t) := \{v \in H^1(D) : v \geq u_c \wedge \mathrm{tr}_{\Gamma_D(t)}\, v = u_D(t)\}, \tag{1.2.8}$$

in which $v \geq u_c$ means $v(x) \geq u_c$ almost everywhere on D. Observe that values $v(x) = u_c$ (which correspond to $p(x) = -\infty$) are now possible in contrast to (1.2.7), which ensures that $\hat{\mathcal{K}}(t)$ is closed. The weak Dirichlet boundary condition in (1.2.8) has the meaning that

$$u_D(t) \in \{v = \mathrm{tr}_{\Gamma_D(t)}\, w : w \in H^1(D) \wedge w \geq u_c \text{ a.e.}\} \tag{1.2.9}$$

as an element of $H^{1/2}(\Gamma_D(t))$ with trace operator $\mathrm{tr}_{\Gamma_D(t)} : H^1(D) \to H^{1/2}(\Gamma_D(t))$ (see [22, Sections 2.7–2.9] for the corresponding definitions). The range of $u_D(t)$ has to be contained in $[u_c, \infty)$, now almost everywhere on $\Gamma_D(t)$ and even for an extension of $u_D(t)$ in $H^1(D)$ almost everywhere on D. $\hat{\mathcal{K}}(t)$ has the desired properties.

Proposition 1.2.1 ([18])**.** *$\hat{\mathcal{K}}(t)$ is a nonempty, closed and convex subset of $H^1(D)$.*

With regard to our models of Brooks–Corey type, let the functions $H : [u_c, \infty) \to \mathbb{R}$ and $kr : H([u_c, \infty)) \to \mathbb{R}$ be continuous, monotonically increasing and bounded with a $u_c < 0$. Furthermore, let us assume that the function K satisfies (1.2.1) and that $f(t) \in L^2(D)$ and $f_N(t) \in L^2(\Gamma_N(t))$. We now seek for a solution defined on the open time cylinder $Q := (0,T) \times D$ and employ the solution space

$$L^2(0,T;H^1(D)) := \left\{ v : (0,T) \to D : \left(\int_0^T \|v(t)\|_{H^1(D)}^2 \,\mathrm{d}t \right)^{1/2} < \infty \right\}. \tag{1.2.10}$$

Observe that we require moreover that the solution $u \in L^2(0,T;H^1(D))$ has a regularity such that $H(u)_t \in L^2(D)$ almost everywhere on $(0,T]$; in this case, all the terms in (1.2.11) make sense and we call u the weak solution of the variational inequality (1.2.6) at the time $t \in (0,T]$ if $u(t) \in \hat{\mathcal{K}}(t)$ and

$$\int_D H(u)_t\,(v-u)\,\mathrm{d}x + \int_D K\,\nabla u \nabla(v-u)\,\mathrm{d}x \geq \int_D K\,kr(H(u))\rho g e_z \nabla(v-u)\,\mathrm{d}x + \int_D f\,(v-u)\;\mathrm{d}x - \int_{\Gamma_N(t)} f_N\,(v-u)\,\mathrm{d}\sigma \quad \forall v \in \hat{\mathcal{K}}(t). \tag{1.2.11}$$

In the rest of this section, we will note some insights about the existence of solutions of (1.2.11). First in Remark 1.2.2, we sum up a classical result from Alt, Luckhaus, and Visintin [6], which also provides an alternative formulation (with weak time derivatives). Afterwards, we outline some recent results of Berninger [18] which will be the starting point for the stochastic version in Chapter 2.

Remark 1.2.2. We cite analytic results concerning the Richards equation with Brooks–Corey-like parameter functions first given in [6]. It is assumed that

$$\begin{gathered} u_D \in H^1(Q) \cap C^0(0,T;H^1(D)), \\ H^0 := H(u(0,\cdot)) \in L^\infty(D), \quad H(u) \in L^\infty(Q) \cap H^1(0,T;V') \end{gathered}$$

with $f_N \equiv 0$, $f \equiv 0$. Here, V denotes the space $V = \left\{v \in H^1(D) : v = 0 \text{ on } \Gamma_D\right\}$, and the goal is to find a solution in the set

$$\mathcal{K}_{\mathrm{A}} = \left\{v \in L^2(0,T;H^1(D)) : u = u_D \text{ on } (0,T)\times\Gamma_D\right\}.$$

Furthermore, set

$$W(y) := \sup_{u_c \le z < \infty} \left(yz - \int_0^z H(z')\,\mathrm{d}z' \right).$$

Then, the weak formulation of the problem reads

$$\begin{aligned} u \in \mathcal{K}_{\mathrm{A}} : \int_D \left(W(H^0(x)) - H^0(x)v(0,x)\right)\alpha(0,x)\,\mathrm{d}x + \iint_Q W(H(u))\,\alpha_t\,\mathrm{d}x\,\mathrm{d}t \\ - \iint_Q H(u)(v\alpha)_t\,\mathrm{d}x\,\mathrm{d}t + \iint_Q K\left(\nabla u - kr(H(u))\rho g e_z\right)\nabla\left((v-u)\alpha\right)\,\mathrm{d}x\,\mathrm{d}t \ge 0 \\ \forall v \in C^0(0,T;H^1(D)) \cap H^1(Q) \cap \mathcal{K}_{\mathrm{A}}, \\ \forall \alpha \in C^2(Q) \text{ with } 0 \le \alpha \le 1 \text{ and } \alpha(T,\cdot) = 0 \text{ in } D. \quad (1.2.12) \end{aligned}$$

Indeed, by taking an α which vanishes on Γ_D and partial integration in t and x, one obtains

$$\int_0^T {}_{V'}\langle H(u)_t - \operatorname{div}(K\nabla u - K\,kr(H(u))\rho g e_z), (v-u)\alpha\rangle_V\,\mathrm{d}t \ge 0$$

for all such v as above, where ${}_{V'}\langle\cdot,\cdot\rangle_V$ denotes the duality bracket for (V', V). Consequently, the solution of (1.2.12) satisfies (1.2.2) with $f \equiv 0$ in V', a.e. in $(0,T)$. Moreover, the boundary conditions (1.2.3)–(1.2.4) are attained in a weak sense and $H(u(0,x)) = H^0(x)$ a.e. in D. For this problem (1.2.12), Alt et al. [6] show that there exists at least one solution.

We resume the approach of Berninger [18] who proceeds by applying an implicit time discretization in the diffusion part and an explicit one in the convective part in (1.2.11); more concretely, for a partition $0 = t_0 < t_1 < \ldots < t_{N_T} = T$ of the time interval $[0,T]$ with $\tau_n = t_n - t_{n-1}$, use the backward Euler as implicit scheme. Then, it can be shown that this time-discretized version of (1.2.11) is a variational inequality of second kind, viz.

$$u \in \hat{\mathcal{K}} : \ \hat{a}(u, v-u) - \hat{\ell}(v-u) + \hat{\phi}(v) - \hat{\phi}(u) \ge 0 \quad \forall v \in \hat{\mathcal{K}}. \tag{1.2.13}$$

The terms in (1.2.13) have the meaning

$$\hat{a}(v,w) := \tau_n \int_D K\nabla v \nabla w\,\mathrm{d}x \quad \forall v, w \in H^1(D), \tag{1.2.14}$$

$$\hat{\ell}(v) := \int_D H(u(t_{n-1}))\, v \,\mathrm{d}x + \tau_n \int_D K\, kr(H(u(t_{n-1})))\rho g e_z \,\nabla v \,\mathrm{d}x$$
$$+ \tau_n \int_D f(t_n)\, v \,\mathrm{d}x - \tau_n \int_{\Gamma_N(t_n)} f_N(t_n)\, v \,\mathrm{d}\sigma \quad \forall v \in H^1(D), \quad (1.2.15)$$

and

$$\hat{\phi}(v) := \int_D \hat{\Phi}(v(x)) \,\mathrm{d}x \quad \forall v \in \hat{\mathcal{K}} \quad (1.2.16)$$

with

$$\hat{\Phi}(z) := \int_0^z H(s) \,\mathrm{d}s \quad \forall z \in [u_c, \infty) \quad (1.2.17)$$

and $\hat{\mathcal{K}} := \hat{\mathcal{K}}(t_n)$. We sum up some results from [18] which will be used later.

Proposition 1.2.3. *Let $H : [u_c, \infty) \to \mathbb{R}$ and $kr : H([u_c, \infty)) \to \mathbb{R}$ be monotonically increasing and bounded with a $u_c < 0$. Furthermore, let us assume that the function K satisfies (1.2.1) and that $f(t_n) \in L^2(D)$ and $f_N(t_n) \in L^2(\Gamma_N(t_n))$. Then, the following holds:*

a) The bilinear form $\hat{a}(\cdot,\cdot)$ from (1.2.14) is continuous and coercive; more precisely, there exist C, c, c_1, $c_2 > 0$ such that

$$\hat{a}(v, w) \leq C\|v\|_{H^1(D)}\|w\|_{H^1(D)} \quad \forall v, w \in H^1(D) \quad (1.2.18)$$

and

$$\hat{a}(v, v) \geq c\|v\|^2_{H^1(D)} - c_1\|v\|_{H^1(D)} - c_2 \quad \forall v \in \hat{\mathcal{K}}. \quad (1.2.19)$$

b) The linear form $\hat{\ell}$ from (1.2.15) is continuous on $\hat{\mathcal{K}}$.

c) The functional $\hat{\phi} : \hat{\mathcal{K}} \to \mathbb{R}$ from (1.2.16) is convex and Lipschitz continuous with

$$|\hat{\phi}(v)| \leq L\|v\|_{H^1(D)} \quad \forall v \in \hat{\mathcal{K}}. \quad (1.2.20)$$

This allows to prove the existence of a solution of the time-discretized Richards equation.

Theorem 1.2.4 ([18])**.** *With the assumptions of Proposition 1.2.3, the variational inequality (1.2.13) has a unique solution.*

All these considerations and the numerical solution methods relying upon them assume that the parameters in the Richards equation are known exactly. Our goal of quantifying the uncertainty in the data will lead us to stochastic parameters, and in the following we will show to what extent one can achieve similar results for the stochastic problem. In Chapter 2, we will specify these stochastic parameters and formulate the Richards equation when taking them into consideration. Afterwards, we will discretize the problem in time, space and stochastic dimensions, analyze its properties and show how it can be solved numerically. This will be the purpose of Chapters 3 and 4.

Chapter 2

The stochastic Richards equation

Looking back to the Richards equation (1.1.5), we ask for its solution if there are some parameters which are random, say the permeability $K(x, \omega)$ or the source term $f(t, x, \omega)$ with $\omega \in \Omega$ for a sample space Ω. This is a natural question in the context of our principal goal of quantifying the uncertainty arising from a lack of information of certain input parameters as addressed in the introduction. Obviously, the solution p is consequently also a "random" function $p(t, x, \omega)$ and we ask for its distribution if, say, $K(x, \cdot)$ is only given by certain probabilities $\mathbb{P}(K(x, \cdot) \in A)$ for Borel sets $A \in \mathrm{Bor}(\mathbb{R})$.

Observe that we first have to specify which parameters are assumed to be random and "how random" they are, i.e. which regularity they have. Especially the last question is fundamental in view of how to model the stochastic Richards equation, treating boundary conditions, etc. We discuss this at the end of this chapter in Remark 2.3.9. Beforehand, we begin with some preliminary work in Sections 2.1 and 2.2 and settle notation and definitions that will be useful later on. Then, in Section 2.3, we derive the weak formulation of the stochastic Richards equation. As set out in Chapter 1, the analysis of the Richards equation (1.1.5) with Kirchhoff transformation and implicit–explicit time discretization leads to a variational inequality of the form (1.2.13). We aim for a generalization of this proceeding in the stochastic context. In doing so, the theory and numerics to be developed in this and the following chapters can be transferred to stochastic variational inequalities of the form

$$u \in \mathcal{K}: \quad a(u, v-u) - \ell(v-u) + \phi(v) - \phi(u) \geq 0 \quad \forall v \in \mathcal{K}, \tag{2.0.1}$$

where ϕ is a convex, lower semicontinuous and proper functional and $\mathcal{K}$ is a convex, closed and nonempty subset of a space of L^2 functions defined on $D \times \Omega$.

2.1 Random fields

Let $(\Omega, \mathcal{E}, \mathbb{P})$ be a given complete probability space with a sample space Ω, a sigma-algebra $\mathcal{E} \subset 2^\Omega$ as set of events, and a probability measure $\mathbb{P} : \mathcal{E} \to [0, 1]$. Here, completeness means that for every $B \subset A$ with $A \in \mathcal{E}$ satisfying $\mathbb{P}(A) = 0$, it is $B \in \mathcal{E}$. For a random variable $X : \Omega \to \mathbb{R}$, we denote by $\sigma(X) \subset \mathcal{E}$ the sigma-algebra

generated by X and call P_X with

$$\mathbb{P}_X(B) := \mathbb{P}(X^{-1}(B)) \quad \text{for all } B \in \mathrm{Bor}(\mathbb{R})$$

the *distribution* of X and

$$\mathbb{E}[X] := \int_\Omega X(\omega)\,\mathrm{d}\mathbb{P}(\omega) \tag{2.1.1}$$

the *expectation* of X (w.r.t. $\mathbb{P}$) and

$$\mathrm{Var}[X] := \mathbb{E}[(X - \mathbb{E}[X])^2] \tag{2.1.2}$$

the *variance* of X (w.r.t. $\mathbb{P}$). For $1 \le p \le \infty$, define

$$L^p(\Omega) := \{X : \Omega \to \mathbb{R} : \|X\|_{L^p(\Omega)} < \infty\} \tag{2.1.3}$$

with norms

$$\|X\|_{L^p(\Omega)} := \left(\int_\Omega |X(\omega)|^p\,\mathrm{d}\mathbb{P}(\omega)\right)^{1/p}$$

for $1 \le p < \infty$ and

$$\|X\|_{L^\infty(\Omega)} := \inf\{s \in \mathbb{R} : |X(\omega)| \le s \text{ a.s.}\}$$

for $p = \infty$. In the following, we will mostly take the case $p = 2$; observe that $L^2(\Omega)$ is a Hilbert space with scalar product

$$(X, Y)_{L^2(\Omega)} := \mathbb{E}[XY] = \int_\Omega X(\omega)Y(\omega)\,\mathrm{d}\mathbb{P}(\omega). \tag{2.1.4}$$

Let X_r be a set of random variables indexed by $r \in \{1, \dots, R\}$, $R \le \infty$, defined on $(\Omega, \mathcal{E}, \mathbb{P})$.

Definition 2.1.1. The random variables $\{X_r\}$ are *independent* if for every finite class $(B_{r_1}, \dots, B_{r_n})$ of Borel sets

$$\mathbb{P}\left(\bigcap_{k=1}^n \{X_{r_k} \in B_{r_k}\}\right) = \prod_{k=1}^n \mathbb{P}(X_{r_k} \in B_{r_k}).$$

Equivalently, the random variables $\{X_r\}$ are independent if the induced sigma-algebras $\{\sigma(X_r)\}$ are independent. The following proposition is well known.

Proposition 2.1.2 ([78])**.** *If $X_1, \dots, X_R$ are independent and integrable random variables, then*

$$\mathbb{E}\left[\prod_{r=1}^R X_r\right] = \prod_{r=1}^R \mathbb{E}\left[X_r\right].$$

Now, we come to the main topic of this section. We refer to [4] and [88] for the following definitions and results.

Definition 2.1.3. A *random field* is a parametrized collection of $\mathbb{R}$-valued random variables $\{X_y\}_{y \in \mathcal{Y}}$ defined on a probability space $(\Omega, \mathcal{E}, \mathbb{P})$ for a parameter set $\mathcal{Y} \subset \mathbb{R}^d$. The collection of all measures $\mathbb{P}_{y_1,\dots,y_k}$ on $\mathrm{Bor}(\mathbb{R}^k)$ defined by

$$\mathbb{P}_{y_1,\dots,y_k}(B_1 \times \cdots \times B_k) := \mathbb{P}(X_{y_1} \in B_1, \dots, X_{y_k} \in B_k) \quad \text{for } y_1, \dots, y_k \in \mathcal{Y} \tag{2.1.5}$$

for arbitrary integer k is called the family of *finite-dimensional distributions* of the random field $X = \{X_y\}_{y \in \mathcal{Y}}$.

A random field is a generalization of a stochastic process which is usually defined on the parameter set $\mathcal{Y} = [0, T]$. A random field can be characterized in main parts by its finite-dimensional distributions, and, conversely, one can construct a random field by finite-dimensional distributions. This is stated by the extension theorem of Kolmogorov.

Theorem 2.1.4 ([88]). *For all $y_1, \dots, y_k \in \mathcal{Y}$, $k \in \mathbb{N}$, let $\mathbb{P}_{y_1,\dots,y_k}$ be probability measures on $\mathbb{R}^k$ such that*

$$\mathbb{P}_{y_{\wp(1)},\dots,y_{\wp(k)}}(B_1 \times \cdots \times B_k) = \mathbb{P}_{y_1,\dots,y_k}(B_{\wp^{-1}(1)} \times \cdots \times B_{\wp^{-1}(k)})$$

for all permutations $\wp$ on $\{1, 2, \dots, k\}$ ("symmetry") and

$$\mathbb{P}_{y_1,\dots,y_k}(B_1 \times \cdots \times B_k) = \mathbb{P}_{y_1,\dots,y_k,y_{k+1},\dots,y_{k+m}}(B_1 \times \cdots \times B_k \times \mathbb{R}^m)$$

for all $m \in \mathbb{N}$ ("consistency"). Then, there exist a probability space $(\Omega, \mathcal{E}, \mathbb{P})$ and a random field $\{X_y\}$ on Ω with $X_y : \Omega \to \mathbb{R}$ such that (2.1.5) holds for all $y_i \in \mathcal{Y}$, $k \in \mathbb{N}$ and all Borel sets B_i.

Alternatively, one can also define random fields as function-valued random variables. More precisely, let G^d denote the set of all functions from $\mathbb{R}^d$ to $\mathbb{R}$ and $\mathcal{G}^d$ the sigma-algebra containing all sets of the form $\{g \in G^d : g(y_i) \in B_i,\ i = 1, \dots, k\}$, where k is an arbitrary integer, $y_i \in \mathbb{R}^d$, $B_i \in \mathrm{Bor}(\mathbb{R})$. Then, define a random field as a measurable mapping from $(\Omega, \mathcal{E})$ into $(G^d, \mathcal{G}^d)$. We obtain again the finite-dimensional distributions via (2.1.5) from the measure $\mathbb{P}$ on $\mathcal{E}$, and these finite-dimensional distributions give a unique probability measure on the set $\mathcal{G}^d$. Now, denote by $X(y, \omega)$ the value which the function in G^d corresponding to ω takes at the point y.

Hence, we can look at the random field X from two perspectives: for given $\omega \in \Omega$, $X(\cdot, \omega)$ is simply a deterministic function from $\mathbb{R}^d$ to $\mathbb{R}$, which we refer to as a *realization* of X; on the other hand, for fixed $y \in \mathbb{R}^d$, $X(y, \cdot)$ is a random variable.

2.2 Tensor spaces

In this section, we introduce tensor products, which will help us to clarify the structure of functions depending on $x \in D \subset \mathbb{R}^d$ and $\omega \in \Omega$.

Definition 2.2.1 ([59]). Let H_1, H_2 be two Hilbert spaces with scalar products $(\cdot, \cdot)_{H_1}$ and $(\cdot, \cdot)_{H_2}$, respectively. For $v_1 \in H_1$ and $v_2 \in H_2$, their *tensor product* $v_1 \otimes v_2$ is defined as a conjugate bilinear form

$$v_1 \otimes v_2(w_1, w_2) := (v_1, w_1)_{H_1} (v_2, w_2)_{H_2} \tag{2.2.1}$$

on $H_1 \times H_2$.

For $v_1 \otimes v_2$, $w_1 \otimes w_2 \in S := \mathrm{span}\{v_1 \otimes v_2 : v_1 \in H_1,\ v_2 \in H_2\}$, one can now define

$$(v_1 \otimes v_2, w_1 \otimes w_2)_{H_1 \otimes H_2} := (v_1, w_1)_{H_1} (v_2, w_2)_{H_2} \tag{2.2.2}$$

and linearly extend it to S. It is straightforward to see that (2.2.2) defines a scalar product on S, which justifies the following definition.

Definition 2.2.2. The completion of S with respect to the scalar product defined in (2.2.2) is called the *tensor space* $H_1 \otimes H_2$.

Proposition 2.2.3 ([59]). *The tensor space $H_1 \otimes H_2$ is a Hilbert space. If $\{e_j\}$ and $\{f_k\}$ are bases of Hilbert spaces H_1 and H_2, then $\{e_j \otimes f_k\}_{j,k\in\mathbb{N}}$ constitute a basis of $H_1 \otimes H_2$.*

Define the usual norm by $\|\cdot\|_{H_1\otimes H_2} := (\cdot,\cdot)^{1/2}_{H_1\otimes H_2}$. From (2.2.2), it follows immediately that

$$\|v_1 \otimes v_2\|_{H_1\otimes H_2} = \|v_1\|_{H_1}\|v_2\|_{H_2} \tag{2.2.3}$$

if $v_1 \in H_1$ and $v_2 \in H_2$.

Let H be a separable Hilbert space. Denote by $L^2(X,\mu;H)$ the space of H-valued square integrable functions on X (w.r.t. a measure μ), equipped with scalar product

$$(f,g)_{L^2(X,\mu;H)} = \int_X (f(x),g(x))_H \, \mathrm{d}\mu(x), \tag{2.2.4}$$

and write shortly $L^2(X,\mu) = L^2(X,\mu;\mathbb{R})$.

Theorem 2.2.4 ([59]). *Let (X,μ) and (Y,ν) be measure spaces such that $L^2(X,\mu)$ and $L^2(Y,\nu)$ are separable. Then, the following holds:*

a) There exists a unique isometric isomorphism

$$L^2(X,\mu)\otimes L^2(Y,\nu) \cong L^2(X\times Y,\mu\times\nu)$$

which maps $f\otimes g$ to $f(x)g(y)$.

b) For any separable Hilbert space H, there exists a unique isometric isomorphism

$$L^2(X,\mu)\otimes H \cong L^2(X,\mu;H)$$

which maps $f\otimes h$ to $f(x)\cdot h$.

Proof. Let $\{e_j(x)\}$ and $\{f_k(y)\}$ be bases of $L^2(X,\mu)$ and $L^2(Y,\nu)$, respectively. Then, the family $\{e_j(x)f_k(y)\}_{j,k\in\mathbb{N}}$ constitutes a basis of $L^2(X\times Y,\mu\times\nu)$, while $\{e_j\otimes f_k\}_{j,k\in\mathbb{N}}$ is a basis of $L^2(X,\mu)\otimes L^2(Y,\nu)$ according to Proposition 2.2.3. Hence, the map

$$U : f\otimes g \mapsto f(x)g(y)$$

can be extended uniquely to a unitary operator from $L^2(X,\mu)\otimes L^2(Y,\nu)$ onto $L^2(X\times Y,\mu\times\nu)$. This proves a).

In order to show part b), let $\{e_j\}$ be a basis of H. Thus, we have

$$\lim_{n\to\infty}\left\| g(x) - \sum_{j=1}^{n}(g(x),e_j)_H e_j \right\| = 0$$

for every $g\in L^2(X,\mu;H)$ such that the linear span of $\{g_j(x)\cdot e_j : g_j \in L^2(X,\mu)\}$ is dense in $L^2(X,\mu;H)$. Now, the map

$$U : \sum_{j=1}^{n}(g_j\otimes e_j) \mapsto \sum_{j=1}^{n} g_j(x)\cdot e_j$$

preserves the scalar products by the definitions (2.2.2) and (2.2.4) and is defined on a dense subspace of $L^2(X,\mu)\otimes H$; hence, it can be extended uniquely to a unitary operator from $L^2(X,\mu)\otimes H$ onto $L^2(X,\mu;H)$. □

In this thesis, we usually take the tensor product of the spaces $H^1(D)$ and $L^2(\Omega)$ and obtain by Theorem 2.2.4 the isomorphism

$$H^1(D) \otimes L^2(\Omega) \cong L^2(\Omega; H^1(D) \cong H^1(D; L^2(\Omega)) \tag{2.2.5}$$

with scalar product

$$(v, w)_{1,0} := \int_\Omega (v, w)_{H^1(D)} \, \mathrm{d}\mathbb{P}$$

and norm

$$\|v\|_{1,0} := (v, v)_{1,0}^{1/2} = \left(\int_\Omega \|v\|_{H^1(D)}^2 \, \mathrm{d}\mathbb{P} \right)^{1/2}.$$

The same construction can be done with tensor products of spaces like L^p for $1 \leq p \leq \infty$ and Banach spaces like C^k, see [12]; e.g. we have

$$L^\infty(D) \otimes L^p(\Omega) := \left\{ v : D \times \Omega \to \mathbb{R} : \operatorname*{ess\,sup}_{x \in D} \|v(x, \cdot)\|_{L^p(\Omega)}^2 < \infty \right\}.$$

Note moreover that spaces like $L^2(0, T; H^1(D))$ from (1.2.10) are defined in the same way. As the notation suggests, the theorem of Fubini holds for $v \in L^2(D) \otimes L^2(\Omega)$ and we have [65]

$$\int_D \int_\Omega v(x, \omega) \, \mathrm{d}\mathbb{P} \, \mathrm{d}x = \int_\Omega \int_D v(x, \omega) \, \mathrm{d}x \, \mathrm{d}\mathbb{P},$$

and this is how the scalar product

$$(v, w)_{0,0} := \int_\Omega (v, w)_{L^2(D)} \, \mathrm{d}\mathbb{P} = \int_D \int_\Omega v(x, \omega) w(x, \omega) \, \mathrm{d}\mathbb{P} \, \mathrm{d}x$$

with norm $\|v\|_{0,0} := (v, v)_{0,0}^{1/2}$ is defined in $L^2(D) \otimes L^2(\Omega)$.

Denote by $C_0^\infty(S)$ for $S \subset \mathbb{R}^k$ the space

$$C_0^\infty(S) := \left\{ \varphi \in C^\infty(S) : \operatorname{supp}(\varphi) := \overline{\{y : \varphi(y) \neq 0\}} \subset S \text{ is compact} \right\}.$$

It is well known (e.g. [109]) that $C_0^\infty(D)$ is dense in $H_0^1(D)$ and that $C_0^\infty(\Omega)$ is dense in $L^2(\Omega)$ if $\Omega \subset \mathbb{R}^k$. Hence, the following statement is not surprising.

Lemma 2.2.5. *Let $\Omega \subset \mathbb{R}^k$. Then, the set*

$$\mathscr{D} := \bigcup_{N=1}^{\infty} \left\{ \sum_{i=1}^{N} \varphi_i^D \varphi_i^\Omega : \varphi_i^D \in C_0^\infty(D), \, \varphi_i^\Omega \in C_0^\infty(\Omega) \right\} \tag{2.2.6}$$

is dense in $H_0^1(D) \otimes L^2(\Omega)$.

Proof. For given $v \in H_0^1(D) \otimes L^2(\Omega)$ and $\varepsilon > 0$, we can find by definition of the tensor space a $v_N = \sum_{i=1}^N v_i^D v_i^\Omega$ with $v_i^D \in H_0^1(D)$, $v_i^\Omega \in L^2(\Omega)$ and

$$\|v_N - v\|_{1,0} \leq \varepsilon/2.$$

For density reasons, we can find for every v_i^D a sequence $(\varphi_{i,n}^D)_n$ with $\varphi_{i,n}^D \to v_i^D$ in $H_0^1(D)$ and for every v_i^Ω a sequence $(\varphi_{i,n}^\Omega)_n$ with $\varphi_{i,n}^\Omega \to v_i^\Omega$ in $L^2(\Omega)$.

First, we show that for $\varepsilon' := \frac{\varepsilon}{2N}$ and each i, we can find an $n^{(i)}$ such that

$$\|\varphi_{i,n}^D \varphi_{i,n}^\Omega - v_i^D v_i^\Omega\|_{1,0} < \varepsilon' \quad \text{for all } n > n^{(i)}. \tag{2.2.7}$$

Choose n_1 such that $\|\varphi_{i,n}^D - v_i^D\|_{H^1(D)} \leq \frac{\varepsilon'}{2\|v_i^\Omega\|_{L^2(\Omega)}}$ for $n > n_1$, whence

$$\|\varphi_{i,n}^D\| \leq \|\varphi_{i,n}^D - v_i^D\|_{H^1(D)} + \|v_i^D\|_{H^1(D)} \leq \frac{\varepsilon'}{2\|v_i^\Omega\|_{L^2(\Omega)}} + \|v_i^D\|_{H^1(D)}.$$

Now, choose an $n_2 > n_1$ with

$$\|\varphi_{i,n}^\Omega - v_i^\Omega\|_{L^2(\Omega)} \leq \frac{\varepsilon'}{2}\left(\frac{\varepsilon'}{2\|v_i^\Omega\|_{L^2(\Omega)}} + \|v_i^D\|_{H^1(D)}\right)^{-1}$$

for $n > n_2$. Then, (2.2.3) provides for all $n > n_2 =: n^{(i)}$

$$\begin{aligned}\|\varphi_{i,n}^D\varphi_{i,n}^\Omega - v_i^D v_i^\Omega\|_{1,0} &\leq \|\varphi_{i,n}^D(\varphi_{i,n}^\Omega - v_i^\Omega)\|_{1,0} + \|(\varphi_{i,n}^D - v_i^D)v_i^\Omega\|_{1,0} \\ &= \|\varphi_{i,n}^D\|_{H^1(D)}\|\varphi_{i,n}^\Omega - v_i^\Omega\|_{L^2(\Omega)} + \|\varphi_{i,n}^D - v_i^D\|_{H^1(D)}\|v_i^\Omega\|_{L^2(\Omega)} \leq \varepsilon'.\end{aligned}$$

By means of (2.2.7), one can estimate

$$\left\|\sum_{i=1}^N \varphi_{i,n}^D\varphi_{i,n}^\Omega - \sum_{i=1}^N v_i^D v_i^\Omega\right\|_{1,0} \leq \sum_{i=1}^N \left\|\varphi_{i,n}^D\varphi_{i,n}^\Omega - v_i^D v_i^\Omega\right\|_{1,0} < \varepsilon/2$$

for all $n \geq \max_i n^{(i)} + 1 =: n_0$. Defining $\varphi = \sum_{i=1}^N \varphi_{i,n_0}^D\varphi_{i,n_0}^\Omega$, we have $\varphi \in \mathscr{D}$ with

$$\|\varphi - v\|_{1,0} \leq \|\varphi - v_N\|_{1,0} + \|v_N - v\|_{1,0} \leq \varepsilon.$$

□

Now, we consider operators on tensor spaces. For $i = 1, 2$, let K_i be Hilbert spaces and A_i be densely defined linear operators from H_i to K_i on the domains $\operatorname{dom} A_i \subset H_i$. Furthermore, let us denote by $\operatorname{dom} A_1 \otimes \operatorname{dom} A_2$ the linear span of $\{v_1 \otimes v_2 : v_i \in \operatorname{dom} A_i \text{ for } i = 1, 2\}$. Then, $\operatorname{dom} A_1 \otimes \operatorname{dom} A_2$ is dense in $H_1 \otimes H_2$. Define

$$A_1 \otimes A_2(v_1 \otimes v_2) := A_1 v_1 \otimes A_2 v_2 \tag{2.2.8}$$

for $v_i \in \operatorname{dom} A_i$ and extend it to a linear operator on $\operatorname{dom} A_1 \otimes \operatorname{dom} A_2$.

Proposition 2.2.6 ([59])**.** *The* operator tensor product $A_1 \otimes A_2$ *is a densely defined linear operator from $H_1 \otimes H_2$ to $K_1 \otimes K_2$. Furthermore, if A_1 and A_2 are bounded operators on H_1 and H_2, respectively, then*

$$\|A_1 \otimes A_2\| = \|A_1\|\|A_2\|.$$

It is obvious that, by induction, one can define tensor products of any finite number of Hilbert spaces and tensor products of any finite number of operators thereon. The following proposition will be helpful.

Proposition 2.2.7. *Let H_r be Hilbert spaces and $(A_n^r)_n$ be a sequence of linear bounded operators on H_r for $r = 1, \ldots, R$ with*

$$\|A_n^r v_r - v_r\|_{H_r} \to 0 \text{ for all } v_r \in H_r \text{ as } n \to \infty. \tag{2.2.9}$$

Then, it holds

$$\|A_n v - v\|_H \to 0 \text{ for all } v \in H := \bigotimes_{r=1}^R H_r \text{ as } n \to \infty \tag{2.2.10}$$

for $A_n = \bigotimes_{r=1}^R A_n^r$.

Proof. Let I^r be the identity operator on H_r and $I = \bigotimes_{r=1}^{R} I^r$. Inductively, we have

$$\begin{aligned}
I - A_n &= \bigotimes_{r=1}^{R} I^r - \bigotimes_{r=1}^{R} A_n^r \\
&= \left(\bigotimes_{r=1}^{R-1} I^r - \bigotimes_{r=1}^{R-1} A_n^r\right) \otimes I^R + \bigotimes_{r=1}^{R-1} A_n^r \otimes \left(I^R - A_n^R\right) \\
&= \left(\bigotimes_{r=1}^{R-2} I^r - \bigotimes_{r=1}^{R-2} A_n^r\right) \otimes I^{R-1} \otimes I^R + \bigotimes_{r=1}^{R-2} A_n^r \otimes \left(I^{R-1} - A_n^{R-1}\right) \otimes I^R + \\
&\quad \bigotimes_{r=1}^{R-1} A_n^r \otimes \left(I^R - A_n^R\right) \\
&= \dots \\
&= \sum_{r=1}^{R} \left(\bigotimes_{s=1}^{r-1} A_n^s\right) \otimes (I^r - A_n^r) \otimes \left(\bigotimes_{s=r+1}^{R} I^s\right).
\end{aligned}$$

Now, take a $v \in H$ of the form $v = \sum_{i=1}^{N} v_i^1 \otimes \cdots \otimes v_i^R$ with $v_i^r \in H_r$. The calculation above allows us to write

$$\begin{aligned}
\|v - A_n v\|_H &= \left\| \left(\sum_{r=1}^{R} \left(\bigotimes_{s=1}^{r-1} A_n^s\right) \otimes (I^r - A_n^r) \otimes \left(\bigotimes_{s=r+1}^{R} I^s\right) \right) v \right\|_H \\
&\leq \sum_{i=1}^{N} \sum_{r=1}^{R} \left\| \left(\bigotimes_{s=1}^{r-1} A_n^s v_i^s\right) \otimes (v_i^r - A_n^r v_i^r) \otimes \left(\bigotimes_{s=r+1}^{R} v_i^s\right) \right\|_H \\
&= \sum_{i=1}^{N} \sum_{r=1}^{R} \left(\prod_{s=1}^{r-1} \|A_n^s v_i^s\|_{H_s}\right) \cdot \|v_i^r - A_n^r v_i^r\|_{H_r} \cdot \left(\prod_{s=r+1}^{R} \|v_i^s\|_{H_s}\right).
\end{aligned}$$

The sequence $(\|A_n^s v_i^s\|_{H_s})_n$ is convergent and thus bounded for all $s = 1, \dots, R$. Consequently, we can conclude $\|v - A_n v\|_H \to 0$ from (2.2.9). The totality of such finite sums being dense in H, (2.2.10) is verified. □

2.3 Weak formulation

In this section, we derive the Richards equation for stochastic parameters. We concentrate on a stochastic permeability K and discuss other random parameters in Remark 2.3.6.

Now, let the permeability K be a random field defined on $D \times \Omega$. For each $\omega \in \Omega$, $K(\cdot, \omega)$ is a realization for which (1.2.1) holds, i.e.

$$K_{\min} \leq K(x, \omega) \leq K_{\max} \quad \forall x \in \overline{D} \quad \forall \omega \in \Omega, \tag{2.3.1}$$

and for which the Richards equation (1.1.5) makes sense. The solution $p(\omega)$ is a function on $(0, T) \times D$, and we take the saturation $\theta(p(\omega))$ and the relative permeability $kr(\theta(p(\omega)))$ as in (1.1.8) and (1.1.7), respectively. We can also define pointwise the Kirchhoff transformation $\kappa : \mathbb{R} \to \mathbb{R}$ as

$$\kappa : p(\omega) \mapsto u(\omega) := \int_0^{p(\omega)} kr(\theta(q))\, \mathrm{d}q. \tag{2.3.2}$$

We obtain the generalized pressure which satisfies (1.2.2)–(1.2.4) for each realization, i.e. for any $\omega \in \Omega$ and any $t \in (0, T]$ consider the boundary value problem

$$H(u(t,\cdot,\omega))_t + \operatorname{div}(\mathbf{v}) = f(t) \qquad \text{on } D \tag{2.3.3}$$
$$u(t,\cdot,\omega) = u_D(t) \qquad \text{on } \Gamma_D(t) \tag{2.3.4}$$
$$\mathbf{v}\cdot\mathbf{n} = f_N(t) \qquad \text{on } \Gamma_N(t), \tag{2.3.5}$$

where the flux $\mathbf{v}$ is defined as

$$\mathbf{v} = -\left(K(\cdot,\omega)\nabla u(t,\cdot,\omega) - K(\cdot,\omega)\, kr(H(u(t,\cdot,\omega)))\rho g e_z\right) \tag{2.3.6}$$

and where $u_D(t)$ with $u_D > u_c$ and $f_N(t)$ are given functions on $\Gamma_D(t)$ and $\Gamma_N(t)$, respectively, which are assumed to be deterministic for the moment (until Remark 2.3.6).

Remark 2.3.1. Note that the Kirchhoff transformation only works, since K is a random field, which gives a sense to single realizations $\omega \in \Omega$. The same holds for the inverse Kirchhoff transformation which is necessary to transform back from our computational solution u to our physical solution p, or, in our setting, to transform back statistics like $\mathbb{E}[u]$ to $\mathbb{E}[p]$. Observe that we have

$$\mathbb{E}[p] = \mathbb{E}[\kappa^{-1}(u)] \neq \kappa^{-1}(\mathbb{E}[u])$$

in general, but by Jensen's inequality (e.g. [78, p. 159]) the estimate

$$\mathbb{E}[p] \leq \kappa^{-1}(\mathbb{E}[u]), \tag{2.3.7}$$

since κ is convex.

For the weak formulation, we can exploit the results from the deterministic case and claim that for each realization $\omega \in \Omega$, the function u satisfies the variational inequality (1.2.11) or its time-discretized counterpart (1.2.13). We want to investigate for the latter problem which conditions are necessary to derive the measurability of u and some regularity results.

To this end, define $\hat{a}^\omega(\cdot,\cdot)$ and $\hat{\ell}^\omega(\cdot)$ as $\hat{a}(\cdot,\cdot)$ and $\hat{\ell}(\cdot)$ in (1.2.14) and (1.2.15), respectively, by replacing the deterministic K by the random field and consider for each $\omega \in \Omega$ the variational inequality

$$u \in \hat{\mathcal{K}}: \quad \hat{a}^\omega(u, v-u) - \hat{\ell}^\omega(v-u) + \hat{\phi}(v) - \hat{\phi}(u) \geq 0 \quad \forall v \in \hat{\mathcal{K}}. \tag{2.3.8}$$

Now, define the (possibly set-valued) map $U : \Omega \to H^1(D)$ by

$$\omega \mapsto U(\omega) := \{u \in \hat{\mathcal{K}} : u \text{ solves } (2.3.8)\}.$$

First, we can state the following.

Proposition 2.3.2. *Let K be a random field satisfying (2.3.1). Then, with the assumptions of Proposition 1.2.3, the map U is measurable.*

Proof. By Theorem 1.2.4, there is a unique solution u of (2.3.8) for each $\omega \in \Omega$. Moreover, the assertions of Proposition 1.2.3 are still valid for $\hat{a}^\omega$ and $\hat{\ell}^\omega$ due to (2.3.1). According to [33, Prop. II.2.2], the variational inequality (2.3.8) is equivalent to

$$u \in \hat{\mathcal{K}}: \quad \hat{a}^\omega(v, u-v) - \hat{\ell}^\omega(u-v) + \hat{\phi}(u) - \hat{\phi}(v) \leq 0 \quad \forall v \in \hat{\mathcal{K}}.$$

Observe that $\hat{\mathcal{K}}$ is separable, since $H^1(D)$ is. With a set $\{v_i\}_{i\in\mathbb{N}}$ which is dense in $\hat{\mathcal{K}}$ and by continuity, we see that

$$u \in U(\omega) \quad \Leftrightarrow \quad \hat{a}^\omega(v_i, u - v_i) - \hat{\ell}^\omega(u - v_i) + \hat{\phi}(u) - \hat{\phi}(v_i) \leq 0 \quad \forall i \in \mathbb{N}.$$

The set

$$U_i(\omega) := \{u \in \hat{\mathcal{K}} : \hat{a}^\omega(v_i, u) - \hat{\ell}^\omega(u) + \hat{\phi}(u) \leq \hat{a}^\omega(v_i, v_i) - \hat{\ell}^\omega(v_i) + \hat{\phi}(v_i)\}$$

is closed by the continuity of $\hat{a}^\omega$, $\hat{\ell}^\omega$ and $\hat{\phi}$. Thus, we can conclude from [8, Theorem 8.2.4] that

$$U(\omega) = \bigcap_{i\in\mathbb{N}} U_i(\omega)$$

is measurable. □

By this proposition, $U(\omega) \in \hat{\mathcal{K}}$ is the unique solution of (2.3.8) for each $\omega \in \Omega$. We can say more.

Proposition 2.3.3. *With the assumptions of Proposition 2.3.2, the map U satisfies $U \in L^p(\Omega; H^1(D))$ for all $1 \leq p \leq \infty$.*

Proof. Again, the assertions of Proposition 1.2.3 are still valid for $\hat{a}^\omega$ and $\hat{\ell}^\omega$ due to (2.3.1). Hence, we have by this proposition

$$\hat{a}^\omega(U(\omega), U(\omega)) \geq c\|U(\omega)\|^2_{H^1(D)} - c_1\|U(\omega)\|_{H^1(D)} - c_2 \tag{2.3.9}$$

on the one hand and by (2.3.8)

$$\begin{aligned} \hat{a}^\omega(U(\omega), U(\omega)) &\leq \hat{a}^\omega(U(\omega), v) - \hat{\ell}^\omega(v - U(\omega)) + \hat{\phi}(v) - \hat{\phi}(U(\omega)) \\ &\leq C\|U(\omega)\|_{H^1(D)}\|v\|_{H^1(D)} + \|\hat{\ell}^\omega\| \left(\|U(\omega)\|_{H^1(D)} + \|v\|_{H^1(D)}\right) \\ &\quad + L\left(\|U(\omega)\|_{H^1(D)} + \|v\|_{H^1(D)}\right) \end{aligned} \tag{2.3.10}$$

for all $v \in \hat{\mathcal{K}}$ on the other hand. Combining (2.3.9) and (2.3.10) provides

$$\begin{aligned} &c\|U(\omega)\|^2_{H^1(D)} \leq \\ &\quad \left(c_1 + C\|v\|_{H^1(D)} + \|\hat{\ell}^\omega\| + L\right)\|U(\omega)\|_{H^1(D)} + \left(\|\hat{\ell}^\omega\| + L\right)\|v\|_{H^1(D)} + c_2. \end{aligned}$$

With the assumptions made on the functions occurring in $\hat{\ell}^\omega$, in particular that K satisfies (2.3.1) and that f and f_N are deterministic, we can state that $\|\hat{\ell}^\omega\|$ is uniformly bounded by a constant. Consequently

$$\|U(\omega)\|_{H^1(D)} \leq \tilde{C}\left(1 + \|v\|_{H^1(D)}\right)$$

for an arbitrary fixed $v \in \hat{\mathcal{K}}$, where the constant $\tilde{C}$ is independent of ω. This yields $U \in L^p(\Omega; H^1(D))$ for all $1 \leq p \leq \infty$. □

Now, instead of testing with $v \in \hat{\mathcal{K}}$, we take $v = V(\omega)$, where $V \in L^2(\Omega; \hat{\mathcal{K}})$, and integrate on Ω, which is possible by Proposition 2.3.3. Then, U is the solution of the variational inequality

$$\begin{aligned} U \in L^2(\Omega; \hat{\mathcal{K}}) : \ \mathbb{E}[\hat{a}^\omega(U, V - U)] - \mathbb{E}[\hat{\ell}^\omega(V - U)] + \mathbb{E}[\hat{\phi}(V)] - \mathbb{E}[\hat{\phi}(U)] \geq 0 \\ \forall\, V \in L^2(\Omega; \hat{\mathcal{K}}). \end{aligned} \tag{2.3.11}$$

Remembering (2.2.5), we detect that

$$\begin{aligned} L^2(\Omega;\hat{\mathcal{K}}) &= \{v \in L^2(\Omega;H^1(D)) : v(\omega) \in \hat{\mathcal{K}} \text{ for almost all } \omega \in \Omega\} \\ &= \left\{v \in H^1(D) \otimes L^2(\Omega) : v \geq u_c \text{ a.e. on } D \times \Omega \wedge \right. \\ &\qquad \left. \operatorname{tr}_{\Gamma_D(t)} v(\cdot,\omega) = u_D(t) \text{ for almost all } \omega \in \Omega\right\}. \end{aligned} \tag{2.3.12}$$

We write this set (2.3.12) in a more compact form as

$$\mathcal{K}(t) := \{v \in H^1(D) \otimes L^2(\Omega) : v \geq u_c \wedge \operatorname{tr}_{\Gamma_D(t)} v = u_D(t)\}. \tag{2.3.13}$$

It is clear that $\mathcal{K}(t)$ is convex and nonempty (note the assumptions on $u_D(t)$ in (1.2.9)). To prove the closedness, we first see that for any $w \in H^1(D) \otimes L^2(\Omega)$ with $w < u_c$ on a subset $E_1 \times E_2 \subset D \times \Omega$ with positive product measure $\mathrm{d}x \times \mathrm{d}\mathbb{P}$, cf. Theorem 2.2.4, there exists for any $\varepsilon > 0$ a second subset $E_3 \times E_4 \subset E_1 \times E_2$ with positive product measure such that $w < u_c - \varepsilon$, whence

$$0 < C \leq \|w - v\|_{0,0} \leq \|w - v\|_{1,0}$$

uniformly for each $v \in \mathcal{K}(t)$. With the same argument for a $w \in H^1(D) \otimes L^2(\Omega)$ with $\operatorname{tr}_{\Gamma_D(t)} w \neq u_D(t)$ on a subset $E_1 \times \Omega \subset D \times \Omega$ with positive product measure, one obtains

$$\begin{aligned} 0 < C &\leq \| \operatorname{tr}_{\Gamma_D(t)} w - \operatorname{tr}_{\Gamma_D(t)} v\|_{L^2(\Gamma_D(t))\otimes L^2(\Omega)} \\ &\leq \| \operatorname{tr}_{\Gamma_D(t)} w - \operatorname{tr}_{\Gamma_D(t)} v\|_{H^{1/2}(\Gamma_D(t))\otimes L^2(\Omega)} \\ &\leq \tilde{C}\|w - v\|_{1,0} \end{aligned}$$

uniformly for each $v \in \mathcal{K}(t)$, where we used the Sobolev embedding theorem (see [22, Theorem 2.19]) in the second line and the trace theorem from [22, Theorem 2.24] in the third line. Hence, we have proven the following analogue to Proposition 1.2.1.

Proposition 2.3.4. *$\mathcal{K}(t)$ is a nonempty, closed and convex subset of the tensor space $H^1(D) \otimes L^2(\Omega)$.*

Let us return to the time-continuous case. If we apply this procedure to problem (1.2.11), it is clear that we seek on $Q \times \Omega = (0,T) \times D \times \Omega$ for a solution which is an element of

$$L^2(0,T;H^1(D) \otimes L^2(\Omega)) := \left\{ v : (0,T) \to D \times \Omega : \left(\int_0^T \|v(t)\|_{1,0}^2 \,\mathrm{d}t \right)^{1/2} < \infty \right\}.$$

As in the deterministic case, the time derivative $H(u)_t$ is problematic and we have to assume that $H(u)_t \in L^2(D) \otimes L^2(\Omega)$ almost everywhere on $(0,T)$. Then, u is called the weak solution of the *stochastic Richards equation* at the time $t \in (0,T]$ if $u(t) \in \mathcal{K}(t)$ and

$$\begin{aligned} &\mathbb{E}\left[\int_D H(u(t))_t \,(v - u(t))\,\mathrm{d}x\right] + \mathbb{E}\left[\int_D K\,\nabla u(t) \nabla\,(v - u(t))\,\mathrm{d}x\right] \\ &\quad \geq \mathbb{E}\left[\int_D K\,kr(H(u(t)))\rho g e_z \nabla\,(v - u(t))\,\mathrm{d}x\right] + \mathbb{E}\left[\int_D f(t)\,(v - u(t))\,\mathrm{d}x\right] \\ &\qquad\qquad - \mathbb{E}\left[\int_{\Gamma_N(t)} f_N(t)\,(v - u(t))\,\mathrm{d}\sigma\right] \quad \forall v \in \mathcal{K}(t). \end{aligned} \tag{2.3.14}$$

If the same time discretization as in Section 1.2 is carried out, we can transfer problem (2.3.14) to problem (2.3.11); this will be shown in Section 3.2.

At this point, it is time for several remarks.

Remark 2.3.5. Having a closer look to the proofs of Propositions 2.3.2 and 2.3.3, we can state some generalizations as it is done in [54] where a similar problem with $\phi \equiv 0$ is examined. First, we can replace $H^1(D)$ with an arbitrary separable Hilbert space H, in particular if H is a subspace of $L^2(D)$ or $H^1(D)$. The source function f can be chosen random, too, which demands the condition $\|f\|_{L^2(D)} \in L^2(\Omega)$. More generally, we can replace $\hat{a}^\omega$, $\hat{\ell}^\omega$, and $\hat{\phi}$ by $a(\cdot,\cdot,\omega)$, $\ell(\cdot,\omega)$, and $\phi(\cdot,\omega)$ provided that they are all Carathéodory maps and satisfy

(i) $(\omega \mapsto \|a(\cdot,\cdot,\omega)\|) \in L^p(\Omega)$,

(ii) $(\omega \mapsto \|\ell(\cdot,\omega)\|) \in L^p(\Omega)$,

(iii) $\phi(\cdot,\omega)$ is Lipschitz continuous with a Lipschitz constant L independent of ω.

Here, ϕ is said to be a *Carathéodory map* if $\phi(x,\cdot)$ is measurable for every $x \in H$ and $\phi(\cdot,\omega)$ is continuous for every $\omega \in \Omega$; with analogous definitions for a and ℓ. Finally, observe that the proofs do not work if the Dirichlet boundary function u_D is also random, since the convex set $\hat{\mathcal{K}}$ is then depending on ω. In this case, different techniques are necessary to obtain the results, see [55].

Remark 2.3.6. In this remark, we discuss the consequences if parameters other than the permeability K are random.

If the relative permeability $kr(\cdot)$ and the saturation $\theta(\cdot)$ are stochastic, say $kr(\cdot,\omega)$ and $\theta(\cdot,\omega)$, it is not clear how to define the Kirchhoff transformation; however, κ will be dependent on ω and so will be the generalized saturation H; and even if kr, θ, and K are independent (in the stochastic sense), $kr(H(u))$ and K will not be in general — apart from the question of how to define $kr \circ H \circ u$ if these are all random fields. In particular the randomness of the nonlinear function H causes problems which are beyond the scope of this thesis. Thus, we will henceforth assume that $kr(\cdot)$ and $\theta(\cdot)$ are deterministic.

We now turn to stochastic source functions f and stochastic Neumann boundary functions f_N. Observe that formulation (2.3.14) already includes them provided that they are integrable. Hence, we assume that

$$f \in L^2(0,T;L^2(D) \otimes L^2(\Omega)), \quad f_N \in L^2(0,T;L^2(\Gamma_N) \otimes L^2(\Omega)), \tag{2.3.15}$$

which corresponds to the considerations in Remark 2.3.5. We will indicate at appropriate locations throughout this thesis how to treat these functions, confer Remarks 2.3.5, 3.1.28 and 4.1.2.

By the formulation in (2.3.12), a stochastic Dirichlet boundary functions u_D is feasible if it can be written as the trace of a function $w \in H^1(D) \otimes L^2(\Omega)$ with $w \geq u_c$ almost everywhere on $D \times \Omega$ analogously to (1.2.9), i.e.

$$u_D(t) \in \{v = \mathrm{tr}_{\Gamma_D(t)}\, w : w \in H^1(D) \otimes L^2(\Omega) \wedge w \geq u_c \text{ a.e.}\}, \tag{2.3.16}$$

where the trace operator $\mathrm{tr}_{\Gamma_D(t)} : H^1(D) \otimes L^2(\Omega) \to H^{1/2}(\Gamma_D(t)) \otimes L^2(\Omega)$ is the generalization of the deterministic trace operator on $H^1(D)$ by creating the operator tensor product with the identity operator on $L^2(\Omega)$. For its discretization, similar to that of stochastic initial conditions u^0, we refer to the Remarks 3.3.3 and 3.3.4 and the reference in Remark 3.1.28.

Remark 2.3.7. The formulation (2.3.8) provides a direct way to compute the expectation $\mathbb{E}[u]$ of the solution of the time-discrete stochastic Richards equation by means of the *Monte Carlo method*. Indeed, we need to sample independent and

identically distributed (i.i.d.) realizations of $K(x,\omega_n)$ or, more generally, of $\hat{a}^{\omega_n}(\cdot,\cdot)$ and $\hat{\ell}^{\omega_n}(\cdot)$ for $n = 1,\dots,N_{\mathrm{MC}}$ and to solve (2.3.8) for each ω_n. We denote the solution by $u(\cdot,\omega_n)$ and compute the average

$$\bar{u}(x) := \frac{1}{N_{\mathrm{MC}}} \sum_{n=1}^{N_{\mathrm{MC}}} u(x,\omega_n).$$

Then, it is well known that

$$\mathbb{P}\left(\lim_{N_{\mathrm{MC}}\to\infty} \bar{u}(x) = \mathbb{E}[u](x) \right) = 1$$

for all $x \in D$ by the strong law of large numbers (e.g. [78, Section 16]). The great advantages of this method are that it is easy to implement and that it can be run independently of the considered setting (e.g. also in high-dimensional spaces). Unfortunately, the convergence speed is rather poor. For this question, we refer to Subsection 4.1.3, where we carry out a more refined analysis of the Monte Carlo approach.

Remark 2.3.8. There are numerous investigations about how the random permeability K is typically distributed. Traditionally (and fortified by a number of experiments), it is often assumed that it has a lognormal distribution, i.e.

$$\log(K(x,\cdot)) \propto \mathcal{N}(\mu_K, \sigma_K^2),$$

see e.g. [27, 39, 57, 100]. This is expedient since it ensures the positivity of K and allows easy computation of moments. Other authors refer to mismatches with real-life data [50, 104] and prefer other models, in particular gamma distributed permeabilities, see e.g. the references in [77]. Comparisons of lognormally and gamma distributed permeability can be found in [77] on the basis of experimental data and in [84] on the basis of field data; [99] provides theoretical estimates about the bias when using false distributions.

We emphasize that, by means of the Karhunen–Loève expansion in Section 3.1, our approach does not depend on a particular choice of the distribution but allows to compute directly on the basis of measured data (see Subsection 3.1.3).

The last remark provides a short survey about generalized random fields and white noise. We refer to the stated references for further details.

Remark 2.3.9. Random fields as defined in Definition 2.1.3 are not sufficient to model highly erratic processes. To this purpose, one introduces *generalized random fields* on D, which are continuous linear mappings from a space S of test functions on D to a space of random variables on Ω (for the exact definition, see [43]). Concerning white noise, one has to deal moreover with generalized random variables which are no longer in $L^2(\Omega)$. If K or u is modeled by a generalized random field, it is not obvious how to interpret the product $K(x,\omega)\nabla u(x,\omega)$ which arises in (2.3.6). We come back to this topic after concretizing "white noise".

Let $\mathscr{S}$ be the Schwartz space of rapidly decreasing smooth functions [109, p. 183] and its dual $\mathscr{S}'$ the space of tempered distributions. In order to model white noise, choose as probability space the set $\Omega = \mathscr{S}'(\mathbb{R}^d)$ equipped with the weak*-topology and Borel sigma-algebra and the measure μ_1 such that

$$\int_{\mathscr{S}'(\mathbb{R}^d)} e^{i\,{}_{\mathscr{S}'(\mathbb{R}^d)}\langle \omega,\varphi\rangle_{\mathscr{S}(\mathbb{R}^d)}} \,\mathrm{d}\mu_1(\omega) = e^{-\|\varphi\|^2/2} \quad \forall \varphi \in \mathscr{S}(\mathbb{R}^d), \tag{2.3.17}$$

which exists due to the Bochner–Minlos theorem [58, Section 2.1]. On this space, the white noise is defined as expansion

$$W(x,\omega) := \sum_{k=1}^{\infty} \eta_k(x) H_{\alpha_k}(\omega), \tag{2.3.18}$$

where $\{\eta_k\}$ is a basis of $L^2(\mathbb{R}^d)$ and $\{H_\alpha\}$ is a basis of $L^2(\mathscr{S}'(\mathbb{R}^d))$, see [58, p. 38]. For each $x \in \mathbb{R}^d$, $W(x,\omega) \in \langle\mathscr{S}\rangle'$, which is the dual space of $\langle\mathscr{S}\rangle := \mathscr{S}(\mathscr{S}'(\mathbb{R}^d,\mu_1))$, cf. [71, Section 3.3] for the precise definitions.

Resuming the question of the multiplication $K(x,\omega)\nabla u(x,\omega)$ if K and u are generalized random fields, this is possible in the usual pointwise way only under further assumptions. For example, it works for $u \in L^2(D) \otimes \langle\mathscr{S}\rangle$ if K is C^∞ in space and has special regularity on Ω, cf. [19, Theorem 6.18]. However, usually one has to take *Wick products* $f \diamond g$ for f, $g \in \langle\mathscr{S}\rangle'$, cf. [58, Section 2.4]. Once having defined the Wick product, one can extend it to analytic functions by its power series, e.g. the Wick exponential $\exp^\diamond X := \sum_{n\in\mathbb{N}} \frac{1}{n!} X^{\diamond n}$ for $X \in \langle\mathscr{S}\rangle'$.

As an example, the Wick SPDE of the simple pressure equation in saturated soil is according to [58, Section 4.6] given by

$$-\operatorname{div}(K(x) \diamond \nabla p(x)) = f(x) \qquad \text{on } D \tag{2.3.19}$$

$$p(x) = 0 \qquad \text{on } \Gamma_D. \tag{2.3.20}$$

Here, $f(x)$ has to be $\langle\mathscr{S}\rangle'$-valued and K is modeled as

$$K(x,\cdot) = \exp^\diamond W(x,\cdot),$$

where W is the white noise process from (2.3.18) such that $K(x,\cdot) \in \langle\mathscr{S}\rangle'$. With appropriate assumptions on the regularity of f, it is possible to prove that (2.3.19)–(2.3.20) has a unique solution $p \in C^2(D) \otimes \langle\mathscr{S}\rangle'$. Further convergence results for linear elliptic SPDEs with white noise can be found in [17].

There are several reasons why this approach does not work for the Richards equation. The main reason is that it no longer allows a pointwise interpretation of the realizations of K and consequently for p and u. Therefore, neither the Kirchhoff transformation $\kappa(p)$ nor the saturation $H(u)$ in (1.1.9) and (1.1.10) make sense, in particular since these functions cannot be written as power series like the Wick exponential function above. Beside from this, it is doubtful whether this Wick product approach provides the correct modeling. First, it is $\mathbb{E}[X \diamond Y] = \mathbb{E}[X]\mathbb{E}[Y]$ for X, $Y \in \langle\mathscr{S}\rangle'$, see [58, p. 64], which means that the mean of the solution of (linear) Wick SPDEs is not influenced by higher statistical moments of K. Secondly, comparisons with SPDEs with usual interpretation of the multiplication, e.g. in [58, Section 3.5], show that the solution from the Wick models do not agree with Monte Carlo simulations in general. For this reason, we will henceforth concentrate solely on random fields K with the usual multiplication.

Chapter 3

Discretization

In this chapter, we describe the discretization of the parabolic nonlinear SPDE to find $u(t) \in \mathcal{K}(t)$ satisfying

$$\begin{aligned}\mathbb{E}\left[\int_D H(u(t))_t\,(v-u(t))\,\mathrm{d}x\right] + \mathbb{E}\left[\int_D K\,\nabla u(t)\nabla\,(v-u(t))\,\mathrm{d}x\right]\\ \geq \mathbb{E}\left[\int_D K\,kr(H(u(t)))\rho g e_z \nabla\,(v-u(t))\,\mathrm{d}x\right] + \mathbb{E}\left[\int_D f(t)\,(v-u(t))\,\mathrm{d}x\right]\\ - \mathbb{E}\left[\int_{\Gamma_N(t)} f_N(t)\,(v-u(t))\,\mathrm{d}\sigma\right] \quad \forall v \in \mathcal{K}(t), \quad (3.0.1)\end{aligned}$$

as introduced in (2.3.14). The structure of the equation and the number of different variables necessitate a separation of the discretization into several steps which are carried out independently. A main difficulty consisting in the nonlinearity in the spatial derivative was already treated in (2.3.2) by the Kirchhoff transformation. As next steps, we have to take care of the random field $K(x,\omega)$ first; we use a Karhunen–Loève expansion to represent it in a countable number of random variables which concurrently specify a finite-dimensional stochastic domain. This is done in Section 3.1. Next, we introduce a suitable time discretization in Section 3.2, where we treat the gravitational term explicitly in order to simplify the following spatial and stochastic problems. It also allows us to write the arising time-discrete problems as variational inequalities, which can be seen as convex minimization problems. Therefor, we use finite element ansatz functions in the spatial domain and polynomial chaos ansatz functions in the stochastic domain which are introduced consecutively in Section 3.3. While this discretization in both spatial and stochastic direction is straightforward for the linear parts of the equation, one has to turn special attention to the nonlinear term, which is done in Subsection 3.3.3. Furthermore, we present some convergence proofs. Finally in Section 3.4, we give some numerical results concerning the discretization error.

3.1 The Karhunen–Loève expansion

A major difficulty when incorporating random fields into differential equations is the problem that one has to deal with abstract measure spaces Ω which are possibly infinite-dimensional or where distributions and density functions—if existent—might only be known approximately by sampling. In addition, one is particularly

interested in handling functions on these abstract spaces, namely random variables defined on the sigma-algebra of random events in Ω. The simplest and mostly used method is the aforementioned Monte Carlo method (Remark 2.3.7), which consists of sampling these functions at randomly chosen points $\omega_n \in \Omega$. The Karhunen–Loève approach in this section pursues another idea by expanding a function K on $D \times \Omega$ in a series of the form

$$K(x,\omega) = \bar{K}(x) + \sum_{r=1}^{\infty} \sqrt{\lambda_r} g_r(x) \xi_r(\omega),$$

where $\{g_r(x)\}$ is an orthonormal set of functions on D and $\{\xi_r(\omega)\}$ is a special set of random variables on Ω. It is named after Kari Karhunen [62] and Michel Loève [78] who developed it independently of each other in the late 1940s.

3.1.1 Definition

Let $K \in L^\infty(D) \otimes L^2(\Omega)$ be a random field. If $K(x,\cdot) \in L^2(\Omega)$ for every $x \in D$, then $K(x,\cdot)$ is often called a *second order random variable*. The family $\{K(x,\cdot)\}_x$ of second order random variables is then called a *second order random field* defined on D. We set the following notations.

Definition 3.1.1. Let $K(x,\cdot)$ be a second order random field. Then, we denote the expectation value of K by $\bar{K}(x) = \mathbb{E}[K(x,\cdot)]$. The function V_K defined on $D \times D$ by

$$V_K(x_1,x_2) := \mathbb{E}\left[\big(K(x_1,\omega) - \bar{K}(x_1)\big)\big(K(x_2,\omega) - \bar{K}(x_2)\big)\right]$$

is the *covariance* of K and the function C_K defined on $D \times D$ by

$$C_K(x_1,x_2) := \mathbb{E}\left[K(x_1,\omega)K(x_2,\omega)\right]$$

is the *correlation* of K.

If the second order random field is centered at expectations, the second moments $\mathbb{E}[K(x,\omega)^2]$ are variances and the correlation and the covariance coincide. In the case $\bar{K}(x) \not\equiv 0$, one can easily see that

$$V_K(x_1,x_2) = C_K(x_1,x_2) - \bar{K}(x_1)\bar{K}(x_2).$$

This means that we have a one-to-one correspondence between these two statistics, and the following statements work with each of them. For a survey of second order properties—those which can be defined or determined by means of covariances—we refer to [78, Chapter X]. In particular, we cite the following two propositions.

Proposition 3.1.2 ([78])**.** *A function $C(x_1,x_2)$ on $D \times D$ is a correlation if, and only if, it is of nonnegative type, i.e. for every finite subset $D_n \subset D$ and every function $h(x)$ on D_n, we have*

$$\sum_{x_1 \in D_n} \sum_{x_2 \in D_n} C(x_1,x_2)h(x_1)h(x_2) \geq 0.$$

Definition 3.1.3. A second order function $K(x,\cdot)$ is *continuous in quadratic mean* (in q.m.) at $x \in D$ if

$$\mathbb{E}\left[\big(K(x+\varepsilon h,\cdot) - K(x,\cdot)\big)^2\right] \to 0 \quad \text{as } \varepsilon \to 0,\ x+\varepsilon h \in D.$$

Proposition 3.1.4 ([78]). *$K(x,\cdot)$ is continuous in q.m. at $x \in D$ if, and only if, $C_K(x_1,x_2)$ is continuous at (x,x).*

We now assume that $K(x,\cdot)$ is continuous in q.m. and consider the operator $\mathcal{C}_K : L^2(D) \to L^2(D)$ defined by

$$g \mapsto (\mathcal{C}_K g)(x_1) := \int_D C_K(x_1,x_2) g(x_2)\,\mathrm{d}x_2. \tag{3.1.1}$$

We will see that the correlation C_K is a Fredholm kernel and $\mathcal{C}_K$ is a trace class operator. First, $\mathcal{C}_K$ is self-adjoint since $C_K(x_1,x_2) = C_K(x_2,x_1)$. Moreover, $\mathcal{C}_K$ is positive, i.e. $(\mathcal{C}_K g, g)_{L^2(D)} \geq 0$ for all $g \in L^2(D)$, due to Proposition 3.1.2. Finally, we show the continuity of the kernel C_K on $D \times D$. By Proposition 3.1.4 we know that C_K is continuous at every diagonal point $(x_1,x_1) \in D \times D$. For $(x_1,x_2) \in D \times D$ with $x_1 \neq x_2$, we observe that $K(x_1+\varepsilon_1 h_1,\cdot) \to K(x_1,\cdot)$ and $K(x_2+\varepsilon_2 h_2,\cdot) \to K(x_2,\cdot)$ in $L^2(\Omega)$ as $\varepsilon_1,\varepsilon_2 \to 0$ implies

$$\begin{aligned} C_K(x_1+\varepsilon_1 h_1, x_2+\varepsilon_2 h_2) &= \mathbb{E}\left[K(x_1+\varepsilon_1 h_1,\cdot)K(x_2+\varepsilon_2 h_2,\cdot)\right] \\ &\to \mathbb{E}\left[K(x_1,\cdot)K(x_2,\cdot)\right] = C_K(x_1,x_2), \end{aligned}$$

where the convergence of the expectation follows from [78, p. 469]. Thus, by Mercer's theorem [109, Theorem VI.4.2], it holds

$$C_K(x_1,x_2) = \sum_{r=1}^{\infty} \lambda_r g_r(x_1) g_r(x_2), \tag{3.1.2}$$

where the series converges absolutely and uniformly on $D \times D$, and the continuous functions g_r are the eigenfunctions of $\mathcal{C}_K$ corresponding to eigenvalues λ_r, i.e.

$$\int_D C_K(x_1,x_2) g_r(x_2)\,\mathrm{d}x_2 = \lambda_r g_r(x_1). \tag{3.1.3}$$

All eigenvalues are nonnegative and can be ordered as $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$. Eigenfunctions corresponding to (necessarily finitely) multiple eigenvalues are written with distinct indices, and they are orthonormalized on D according to

$$\int_D g_r(x) g_s(x)\,\mathrm{d}x = \delta_{rs}. \tag{3.1.4}$$

Employing (3.1.2) and the uniform convergence, we get

$$\sum_{r=1}^{\infty} \lambda_r = \sum_{r=1}^{\infty} \lambda_r \int_D g_r(x) g_r(x)\,\mathrm{d}x = \int_D C_K(x,x)\,\mathrm{d}x < \infty.$$

Let us assume for a moment that $\bar{K}(x) = \mathbb{E}[K(x,\omega)] \equiv 0$. We now define for $r = 1,2,\ldots$

$$\xi_r(\omega) := \frac{1}{\sqrt{\lambda_r}} \int_D K(x,\omega) g_r(x)\,\mathrm{d}x. \tag{3.1.5}$$

These integrals exist, since K (in q.m.) and g_r are continuous on the domain D. The $\{\xi_r\}$ are orthonormal on Ω due to (3.1.3) and (3.1.4), since

$$\begin{aligned} \mathbb{E}[\xi_r \xi_s] &= \frac{1}{\sqrt{\lambda_r \lambda_s}} \int_D \int_D \mathbb{E}[K(x_1,\omega)K(x_2,\omega)] g_r(x_1) g_s(x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 \\ &= \frac{1}{\sqrt{\lambda_r \lambda_s}} \int_D \lambda_s g_r(x_1) g_s(x_1)\,\mathrm{d}x_1 \\ &= \delta_{rs}. \end{aligned} \tag{3.1.6}$$

The same calculation shows

$$\mathbb{E}[K(x,\omega)\xi_r(\omega)] = \sqrt{\lambda_r} g_r(x), \tag{3.1.7}$$

which represents the coefficients when expanding $K(x,\cdot)$ in the orthonormal directions ξ_r. This directs us to the partial sum

$$K_M(x,\omega) = \sum_{r=1}^{M} \sqrt{\lambda_r} g_r(x)\xi_r(\omega).$$

Exploiting the orthogonality, (3.1.7) and (3.1.2), we see by

$$\begin{aligned}\mathbb{E}\left[\left(K(x,\omega) - K_M(x,\omega)\right)^2\right] &= \mathbb{E}\left[K(x,\omega)^2\right] + \mathbb{E}\left[K_M(x,\omega)^2\right] \\ &\quad - 2\mathbb{E}\left[\sum_{r=1}^{M} \sqrt{\lambda_r} g_r(x)\xi_r(\omega) K(x,\omega)\right] \\ &= C_K(x,x) - \sum_{r=1}^{M} \lambda_r g_r(x) g_r(x) \to 0\end{aligned}$$

that $K_M \to K$ in $L^2(\Omega)$ as $M \to \infty$, uniformly on D. These considerations give rise to the following theorem.

Theorem 3.1.5 ([78]). *A second order random field $K(x,\omega) \in L^\infty(D) \otimes L^2(\Omega)$ continuous in q.m. on a domain D with centered expectations has an orthogonal decomposition*

$$K(x,\omega) = \sum_{r=1}^{\infty} \sqrt{\lambda_r} g_r(x)\xi_r(\omega) \tag{3.1.8}$$

with (3.1.4) and (3.1.6) if, and only if, the λ_r are the eigenvalues and the g_r are the orthonormalized eigenfunctions of its correlation operator. Then the series converges in $L^2(\Omega)$ uniformly on D.

Proof. We still need to show the "only if" assertion. If K has the decomposition (3.1.8), we have

$$C_K(x_1,x_2) = \lim_{M\to\infty} \mathbb{E}[K_M(x_1,\omega)K_M(x_2,\omega)] = \sum_{r=1}^{\infty} \lambda_r g_r(x_1) g_r(x_2)$$

and consequently

$$\int_D C_K(x_1,x_2) g_r(x_2)\,\mathrm{d}x_2 = \lambda_r g_r(x_1).$$

□

We now drop the assumption $\bar{K}(x) = \mathbb{E}[K(x,\omega)] \equiv 0$. To this end, we split up $K(x,\omega) = \bar{K}(x) + \tilde{K}(x,\omega)$ and perform the orthogonal decomposition for the random part $\tilde{K}$. Another possibility is to use directly the covariance kernel V_K, which is identical to the correlation kernel $C_{K-\bar{K}}$. In either case, defining the ξ_r according to

$$\xi_r(\omega) := \frac{1}{\sqrt{\lambda_r}} \int_D \left(K(x,\omega) - \bar{K}(x)\right) g_r(x)\,\mathrm{d}x \tag{3.1.9}$$

instead of (3.1.5), Theorem 3.1.5 holds verbatim if one replaces (3.1.8) with

$$K(x,\omega) = \bar{K}(x) + \sum_{r=1}^{\infty} \sqrt{\lambda_r} g_r(x)\xi_r(\omega). \tag{3.1.10}$$

Definition 3.1.6. The orthogonal decomposition (3.1.10) is called the *Karhunen–Loève expansion* (KL expansion) of the random field $K \in L^\infty(D) \otimes L^2(\Omega)$.

Definition 3.1.7. The partial sum

$$K_M(x,\omega) := \bar{K}(x) + \sum_{r=1}^{M} \sqrt{\lambda_r} g_r(x) \xi_r(\omega) \tag{3.1.11}$$

is called the *M-th truncate of the Karhunen–Loève expansion* (3.1.10).

Remark 3.1.8. The preceding definitions can be extended easily to random fields K in $L^2(D) \otimes L^2(\Omega)$, see [95, Theorem 2.5]. In particular, it holds

$$\|K - K_M\|_{0,0}^2 = \mathbb{E}\left[\int_D \left(K(x,\omega) - K_M(x,\omega)\right)^2 \,\mathrm{d}x\right] \to 0. \tag{3.1.12}$$

In Proposition 3.1.12, Corollaries 3.1.14 and 3.1.18, and Theorem 3.1.19, we will give conditions for the convergence of the KL expansion in $L^\infty(D) \otimes L^\infty(\Omega)$.

Partial sums K_M of the KL expansion for K are optimal finite-dimensional approximations of K in the mean square sense having an orthonormal basis in D (see [47, p. 24]). A quantitative information provides the following theorem in [95] ($P_{U\otimes V}$ is the projection onto $U \otimes V$ in the L^2 sense).

Theorem 3.1.9. *If $K \in L^2(D) \otimes L^2(\Omega)$ has the KL expansion (3.1.10), then for any $M \in \mathbb{N}$ it holds*

$$\inf_{\substack{U \subset L^2(D) \\ \dim(U)=M}} \|K - P_{U\otimes L^2(\Omega)} K\|_{0,0}^2 = \sum_{r \geq M+1} \lambda_r$$

with equality only for $U = \operatorname{span}\{g_1, g_2, \cdots, g_M\}$, i.e. for the M-th truncate (3.1.11).

The approximation quality of the M-th truncate is thus depending on the decay of the eigenvalues λ_r, which in turn is depending on the covariance kernel V_K. Following the theory described in [66], one can find bounds for the eigenvalue decay in dependence of the regularity of V_K.

Theorem 3.1.10 ([95])**.** *Let $\mathcal{V}_K : L^2(D) \to L^2(D)$ be the operator*

$$g \mapsto (\mathcal{V}_K g)(x_1) = \int_D V_K(x_1, x_2) g(x_2) \,\mathrm{d}x_2 \tag{3.1.13}$$

in analogy to (3.1.1). We denote by $\{(\lambda_r, g_r)\}_{r\geq 1}$ the eigenpair sequence of the operator $\mathcal{V}_K$. Let $\{D_j\}$ be a finite partition of the domain $D \subset \mathbb{R}^d$.

a) If V_K is piecewise analytic on $D \times D$, then there exist positive constants C and $\tilde{C}$, only depending on V_K, such that

$$0 \leq \lambda_r \leq C \exp(-\tilde{C} r^{\frac{1}{d}}) \quad \text{for all } r \geq 1. \tag{3.1.14}$$

b) If V_K is piecewise H^k on $D \times D$ with $k \geq 1$, then there exists a positive constant C, only depending on V_K, such that

$$0 \leq \lambda_r \leq C r^{-\frac{k}{d}} \quad \text{for all } r \geq 1.$$

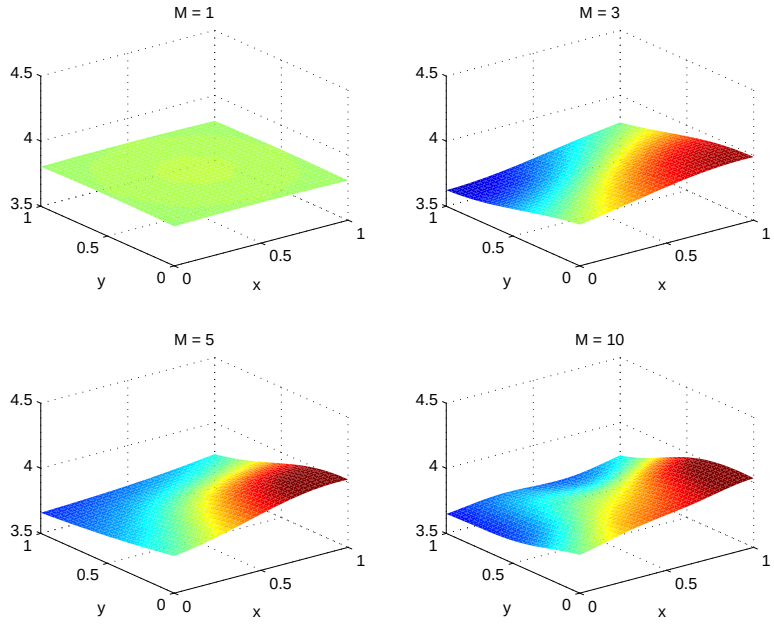


Figure 3.1: Example of a realization of the KL expansion $K_M(x,\omega)$ on $D=(0,1)^2$ with covariance kernel (3.1.16) for different truncations $M=1,3,5,10$.

c) If V_K is piecewise smooth on $D\times D$, then for any $s>0$ there exists a positive constant C, only depending on V_K and s, such that

$$0\le\lambda_r\le Cr^{-s}\quad \textit{for all } r\ge 1.$$

Moreover, if all the domains D_j have the uniform cone property, then for any $s>0$ and any multi-index $\alpha\in\mathbb{N}^d$ there exists a positive constant C, depending on V_K, s and α, such that on each D_j

$$\|\partial^\alpha g_r\|_{L^\infty(D_j)}\le C|\lambda_r|^{-s}\quad \textit{for all } r\ge 1. \tag{3.1.15}$$

Example 3.1.11. Two of the most commonly used covariance kernels (e.g. [9, 47, 63, 95]) are the *exponential covariance* kernel

$$V_K(x_1,x_2)=\exp(-|x_1-x_2|/\gamma), \tag{3.1.16}$$

see Appendix A for more details, and the *Gaussian covariance* kernel

$$V_K(x_1,x_2)=\sigma^2\exp(-(x_1-x_2)^2/\gamma^2). \tag{3.1.17}$$

For $D=(-a,a)$, we obtain the following decay rates, which reflects the regularity of V_K: for the kernel (3.1.16), we have

$$\frac{8a^2\gamma}{(r\gamma\pi)^2+(2a)^2}\le\lambda_r\le\frac{8a^2\gamma}{((r-1)\gamma\pi)^2+(2a)^2}$$

by (A.11) and (A.12) and thus

$$\lambda_r\le Cr^{-2}$$

for a constant $C=C(a,\gamma)$, whereas the analytic kernel (3.1.17) even provides

$$\lambda_r\le C\frac{(1/\gamma)^r}{\Gamma(r/2)}$$

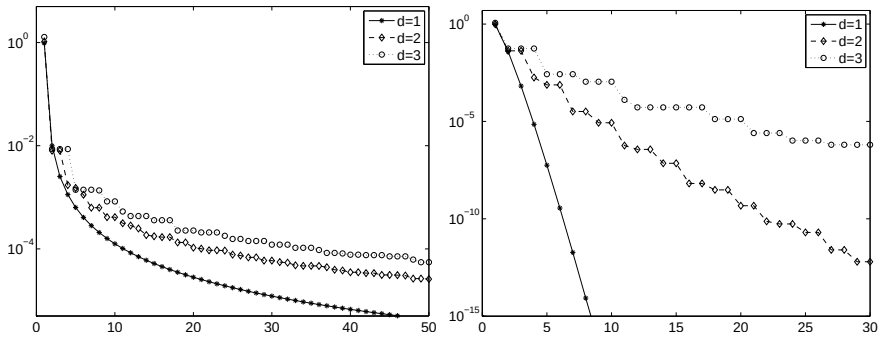


Figure 3.2: Eigenvalue decay for different covariance kernels on $D = (0,1)^d$ for $d = 1, 2, 3$. Left: exponential covariance kernel ($\gamma = 20$). Right: Gaussian covariance kernel ($\sigma = 1$, $\gamma = 2$)

with a constant $C = C(a, \gamma, \sigma)$, see [103], where $\Gamma(\cdot)$ denotes the gamma function interpolating the factorial. In the latter case, the decay of the eigenvalues is even faster than predicted by Theorem 3.1.10. An example is shown in Figure 3.2.

By means of Theorem 3.1.10, we are now ready to give a criterion for the convergence of K_M to K in $L^\infty(D) \otimes L^\infty(\Omega)$.

Proposition 3.1.12. *Let V_K be analytic on $D \times D$ and $K \in L^\infty(D) \otimes L^\infty(\Omega)$. If the random variables ξ_r in the KL expansion (3.1.10) are uniformly bounded, i.e. if there exists a $c \in \mathbb{R}$ such that $\|\xi_r\|_{L^\infty(\Omega)} \leq c$ for all $r \in \mathbb{N}$, then each term in (3.1.10) can be bounded by*

$$\left\| \sqrt{\lambda_r} g_r \xi_r \right\|_{L^\infty(D) \otimes L^\infty(\Omega)} \leq C \exp(-\tilde{C} r^{\frac{1}{d}}) \tag{3.1.18}$$

with positive constants C and $\tilde{C}$, depending only on V_K, and the truncate K_M converges to K in $L^\infty(D) \otimes L^\infty(\Omega)$ with

$$\|K - K_M\|_{L^\infty(D) \otimes L^\infty(\Omega)} \leq C \Gamma(d, \tilde{C} M^{\frac{1}{d}})$$

with positive constants C, depending on V_K, and $\tilde{C}$, depending on V_K and d, where $\Gamma(d, s)$ is the incomplete gamma function [2, p. 260].

Proof. By virtue of the boundedness of ξ_r, we get by (3.1.14) and (3.1.15)

$$\left\| \sqrt{\lambda_r} g_r \xi_r \right\|_{L^\infty(D) \otimes L^\infty(\Omega)} \leq C \lambda_r^{\frac{1}{2} - s} \leq C \exp\left(- \left(\frac{1}{2} - s \right) \tilde{C} r^{\frac{1}{d}} \right)$$

for all $s > 0$.

For the second assertion, (3.1.18) provides

$$\begin{aligned} \|K - K_M\|_{L^\infty(D) \otimes L^\infty(\Omega)} &= \left\| \sum_{r=M+1}^{\infty} \sqrt{\lambda_r} g_r \xi_r \right\|_{L^\infty(D) \otimes L^\infty(\Omega)} \\ &\leq \sum_{r=M+1}^{\infty} C \exp(-\tilde{C} r^{\frac{1}{d}}) \leq \int_M^\infty C \exp(-\tilde{C} x^{\frac{1}{d}}) \, \mathrm{d}x = C d \tilde{C}^{-d} \Gamma(d, \tilde{C} M^{\frac{1}{d}}) \to 0 \end{aligned}$$

as $M \to \infty$. □

We can relax the conditions on the boundedness of ξ_r and the regularity of K and V_K. By using Chebyshev's inequality and the Borel–Cantelli lemma, one can prove the following proposition.

Proposition 3.1.13 ([78]). *Let the random variables X_r be orthogonal. If it holds $\sum_r \log^2(r)\mathbb{E}[X_r^2] < \infty$, then the series $\sum_r X_r$ converges in L^2 sense and almost surely.*

Applying Proposition 3.1.13 to KL expansions with $X_r = \sqrt{\lambda_r} g_r(x)\xi_r$, we get immediately another convergence criterion.

Corollary 3.1.14. *If the functions in (3.1.11) satisfy*

$$\sum_{r=1}^{\infty} \lambda_r \log^2(r) \|g_r\|^2_{L^\infty(D)} \operatorname{Var}(\xi_r) < \infty,$$

then the KL expansion converges in $L^\infty(D) \otimes L^\infty(\Omega)$.

3.1.2 Independent KL expansions and reformulation

We now turn to the functions ξ_r in the Karhunen–Loève expansion (3.1.10). The relationship (3.1.6) means that these random variables ξ_r are mutually uncorrelated and have unit variance. Moreover, by (3.1.9) we obtain immediately

$$\mathbb{E}[\xi_r] = 0 \quad \text{for all } r. \tag{3.1.19}$$

Another important concept when dealing with random variables is independence. If two or more random variables are independent, they are also uncorrelated (this can be seen for random variables $\xi_1, \ldots, \xi_M$ from the KL expansion (3.1.10) by combining (3.1.19) and Proposition 2.1.2), but the opposite is not true in general. In this subsection, we state some consequences if the functions ξ_r are independent. Later on, we have a closer look on the important class of normally distributed random fields where the KL decomposition turns out to consist of independent terms. Finally, we derive an exact formulation of (3.0.1) when K is approximated by a truncated KL series (3.1.11).

A very interesting feature of having independent ξ_r is the fact that this provides further criteria for the $L^\infty(\Omega)$ convergence of the KL expansion. In order to prove this, we recall a famous result from probability theory, *Kolmogorov's zero–one law*. For preparation, let X_r be a set of random variables indexed by $r \in \{1, \ldots, M\}$, $M \leq \infty$, defined on $(\Omega, \mathcal{E}, \mathbb{P})$ with the notations from Section 2.1, and let us consider the sigma-algebras $\sigma_r := \sigma(X_r, X_{r+1}, \ldots)$. The σ_r form a nonincreasing sequence of sigma-algebras, and its intersection

$$\sigma_\infty = \bigcap_{r=1}^{\infty} \sigma_r$$

is called *tail sigma-algebra*. Note that all σ_r and in particular the tail sigma-algebra σ_∞ are contained in $\sigma(X_1, X_2, \ldots)$ induced by the whole sequence $(X_r)_r$. The elements of σ_∞ are called *tail events*, and they describe, loosely speaking, the events which can only be determined by infinitely many X_r. Now, let $\{X_r\}$ be independent. Then $\sigma(X_1, \ldots, X_r)$ is independent of σ_{r+1} and therefore also independent of $\sigma_\infty \subset \sigma_{r+1}$ for arbitrary r. Hence, σ_∞ is independent of $\sigma(X_1, X_2, \ldots)$ and, being contained therein, even independent of itself. Since an event A is independent of itself if, and only if, $\mathbb{P}(AA) = \mathbb{P}(A)\mathbb{P}(A)$, i.e. if $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$, the zero–one law reads:

Theorem 3.1.15 ([78])**.** *On a sequence of independent random variables, the probability of a tail event is either* 0 *or* 1*.*

In other words, the tail sigma-algebra of a sequence of independent random variables is equivalent to $\{\emptyset, \Omega\}$. Since the set of convergence $\{\omega : \sum_{r=1}^{\infty} X_r(\omega) < \infty\}$ is a tail event, we can state:

Corollary 3.1.16. *If $\{X_r\}$ is a sequence of independent random variables, then the series $\sum_{r=1}^{\infty} X_r$ converges almost surely or diverges almost surely.*

In view of this corollary, it is clear that the conditions under which one obtains almost sure convergence is under the assumption of independence by far less restrictive than elsewise. For an arbitrary sequence of random variables, we have the table of convergence below (see [78] for corresponding definitions):

$$\begin{array}{ccccc} & & \text{convergence in } L^2 & & \\ & & \Downarrow & & \\ \text{convergence a.s.} & \Rightarrow & \text{convergence in probability} & \Rightarrow & \text{convergence of laws} \end{array}$$

In contrast, for series of independent random variables, the reverse implications are also true [78, Section 17].

Theorem 3.1.17. *For series $\sum_r X_r$ of independent random variables X_r which are centered at expectations and are uniformly bounded, convergence of laws, convergence in probability, convergence in $L^2(\Omega)$, and almost sure convergence are equivalent.*

Corollary 3.1.18. *If all random variables ξ_r obtained in the KL decomposition (3.1.10) of K are independent and bounded, then it holds*

$$\|K - K_M\|^2_{L^\infty(D)\otimes L^\infty(\Omega)} \to 0.$$

Proof. Combine Theorems 3.1.5 and 3.1.17. □

The question arises whether one can ensure the independence of the uncorrelated random variables ξ_r. Note that their distribution is entirely determined by the distribution of K. An important special case is when $K(x, \cdot)$ is a Gaussian random variable. Since normality is preserved under linear combinations and passages to the limit in q.m. and therefore also under integration in q.m., this property is transferred to the ξ_r. Even more is true: for such random fields, orthogonality becomes independence. The precise formulation reads as follows.

Theorem 3.1.19 ([78])**.** *If the covariance V_K is continuous on $D \times D$, then the random field $K(x, \cdot)$ is Gaussian if, and only if, the random variables ξ_r defined by (3.1.9) are Gaussian. In this case, the ξ_r are independent and the KL expansion (3.1.10) converges in $L^\infty(D) \otimes L^\infty(\Omega)$.*

We stress that the independence of the normally distributed ξ_r is not implied by the fact that they are uncorrelated, which is—albeit sometimes claimed in the literature (see [83])—not true in general, but by the normal distribution of K.

Remark 3.1.20. As we will see below, independence is essential for the formulation of the solution space when K is approximated by a truncated KL expansion K_M. For that, one is sometimes tempted to assume that the ξ_r are Gaussian (e.g. [47, 75]).

Keese [64] proposes another remedy: if there exists a (nonlinear) transformation $G = F^{-1}_{K(x,\cdot)} \circ \mathrm{erf}$, with the distribution function $F_{K(x,\cdot)}(y) = \mathbb{P}(K(x,\omega) < y)$ and the error function $\mathrm{erf}(\cdot)$, see [2, p. 297], such that $K(x,\omega) = G(x, Y(x,\omega))$ for a Gaussian random field Y, the KL expansion reads

$$K(x,\omega) = G\left(x, \bar{Y}(x) + \sum \sqrt{\lambda_{Y,r}} g_{Y,r} \xi_r\right) \tag{3.1.20}$$

with Gaussian ξ_r. The drawback of this idea reveals when using a stochastic Galerkin approach (Subsection 4.1.1), where the nonlinearity of K in ξ_r causes avoidable approximation errors while assembling the stiffness matrix (see Remark 4.1.5).

Another possibility in the case of non-independent ξ_r—which is more related to the purpose of stochastic Galerkin methods—is to expand each ξ_r in a new set of independent Gaussian random variables $\zeta_r = (\zeta_1, \dots, \zeta_{M_r})$ as a polynomial chaos approach

$$\xi_r(\omega) = \sum_{k=1}^{P_r} \xi_{r,k} \Psi_k(\zeta_r(\omega)) \tag{3.1.21}$$

with Hermite polynomials Ψ_k (see Subsection 3.3.1). For the consequences for our approach, we also refer to Remark 4.1.5.

Keeping these options in mind, we make for simplicity the following assumption for the rest of this thesis.

Assumption 3.1.21. The ξ_r in the Karhunen–Loève expansion (3.1.10) are independent.

Now approximate the function K in (3.0.1) by its truncated KL expansion K_M. For a fixed truncation number $M < \infty$, let us denote by $\xi = (\xi_1, \dots, \xi_M)$ the vector of random variables $\xi_r : \Omega \to \mathbb{R}$ and by $\Omega_r := \xi_r(\Omega)$ their range. The set $\Omega^{(M)} := \Omega_1 \times \cdots \times \Omega_M \subset \mathbb{R}^M$ is then a probability space equipped with the sigma-algebra $\mathrm{Bor}(\Omega^{(M)})$ and a unique probability measure $\mathbb{P}^{(M)}$ satisfying

$$\mathbb{P}^{(M)} = \prod_{r=1}^{M} \mathbb{P}_{\xi_r},$$

see [16, Theorem 4.14] for the uniqueness of $\mathbb{P}^M$. One may identify

$$L^2(\Omega, \sigma(\xi_1, \dots, \xi_M), \mathbb{P}) \cong L^2(\Omega^{(M)}, \mathrm{Bor}(\Omega^{(M)}), \mathbb{P}^{(M)}). \tag{3.1.22}$$

In turn, the right-hand side can be identified with

$$L^2(\Omega^{(M)}, \mathrm{Bor}(\Omega^{(M)}), \mathbb{P}^{(M)}) \cong \bigotimes_{r=1}^{M} L^2(\Omega_r, \mathrm{Bor}(\Omega_r), \mathbb{P}_{\xi_r}), \tag{3.1.23}$$

see Theorem 2.2.4. We assume that the random variables ξ_r have known density functions $\mathrm{pdf}_r : \Omega_r \to \mathbb{R}^+$ with $\mathrm{pdf}_r \in L^\infty(\Omega_r)$. Then

$$\mathrm{pdf}(y) = \prod_{r=1}^{M} \mathrm{pdf}_r(y_r) \quad \text{for all } y = (y_1, \dots, y_M) \in \Omega^{(M)} \tag{3.1.24}$$

is the joint probability density function of ξ. Due to the identification (3.1.22), the expectation of a function $v \in L^2(\Omega^{(M)})$ can be written as

$$\begin{aligned}
\mathbb{E}[v] &= \int_{\Omega} v(\xi(\omega)) \,\mathrm{d}\mathbb{P}(\omega) \\
&= \int_{\Omega^{(M)}} v(y) \,\mathrm{d}\mathbb{P}^{(M)}(y) \\
&= \int_{\Omega_M} \cdots \int_{\Omega_1} v(y_1, \ldots, y_M) \,\mathrm{d}\mathbb{P}_{\xi_1}(y_1) \cdots \,\mathrm{d}\mathbb{P}_{\xi_M}(y_M) \qquad (3.1.25) \\
&= \int_{\Omega^{(M)}} v(y) \mathrm{pdf}(y) \,\mathrm{d}y \\
&= \int_{\Omega_M} \cdots \int_{\Omega_1} v(\xi_1, \ldots, \xi_M) \mathrm{pdf}_1(\xi_1) \,\mathrm{d}\xi_1 \cdots \mathrm{pdf}_M(\xi_M) \,\mathrm{d}\xi_M.
\end{aligned}$$

In the last line we changed the notation from y_r to ξ_r to underline the fact that we introduced "coordinates" $\xi_1, \ldots, \xi_M$ to the space Ω by means of the KL expansion.

Remark 3.1.22. At this point, we can concretize the meaning of the space Ω and the function K in (3.0.1). We replace K by its truncated KL expansion K_M, which results in looking at $(\Omega, \sigma(\xi_1, \ldots, \xi_M))$ instead of $(\Omega, \mathcal{E})$. Thus, we have to find $u_M(t) \in \mathcal{K}(t)$ which satisfies

$$\begin{aligned}
&\mathbb{E}\left[\int_D H(u_M(t))_t \left(v - u_M(t)\right) \,\mathrm{d}x\right] + \mathbb{E}\left[\int_D K_M \nabla u_M(t) \nabla \left(v - u_M(t)\right) \,\mathrm{d}x\right] \\
&\geq \mathbb{E}\left[\int_D K_M kr(H(u_M(t))) \rho g e_z \nabla \left(v - u_M(t)\right) \,\mathrm{d}x\right] + \mathbb{E}\left[\int_D f(t) \left(v - u_M(t)\right) \,\mathrm{d}x\right] \\
&\qquad - \mathbb{E}\left[\int_{\Gamma_N(t)} f_N(t) \left(v - u_M(t)\right) \,\mathrm{d}\sigma\right] \quad \forall v \in \mathcal{K}(t). \qquad (3.1.26)
\end{aligned}$$

By means of the Doob–Dynkin lemma (see [90, p. 7]), the solution u_M of the stochastic variational inequality (3.1.26) can be described by just a finite number of random variables, i.e.

$$u_M(t, x, \omega) = u_M(t, x, \xi_1(\omega), \ldots, \xi_M(\omega)) = u_M(t, x, \xi(\omega)).$$

Consequently, the stochastic part of the solution space is given by (3.1.22) and the set $\mathcal{K}(t)$ from (2.3.13) can be written as

$$\begin{aligned}
\mathcal{K}(t) = \big\{ &v \in H^1(D) \otimes L^2(\Omega) : v(x, \xi(\omega)) \geq u_c \text{ a.e. on } D \times \Omega \wedge \\
&\mathrm{tr}_{\Gamma_D(t)} v(\cdot, \xi(\omega)) = u_D(t) \text{ for almost all } \omega \in \Omega \big\} \qquad (3.1.27)
\end{aligned}$$

or

$$\begin{aligned}
\mathcal{K}(t) = \Big\{ &v \in H^1(D) \otimes L^2(\Omega^{(M)}) : v(x, \xi) \geq u_c \text{ a.e. on } D \times \Omega^{(M)} \wedge \\
&\mathrm{tr}_{\Gamma_D(t)} v(\cdot, \xi) = u_D(t) \text{ for almost all } \xi \in \Omega^{(M)} \Big\}. \qquad (3.1.28)
\end{aligned}$$

Remark and Notation 3.1.23. We will henceforth always operate on the space $(\Omega^{(M)}, \mathrm{Bor}(\Omega^{(M)}), \mathbb{P}^{(M)})$. In light of the identification (3.1.22), we will however mostly write $L^2(\Omega)$ in the following and use from time to time the short hand notation

$$\mathbb{E}[v] = \int_{\Omega} v(\xi) \,\mathrm{d}\mathbb{P}$$

for the expectation of $v \in L^2(\Omega^{(M)})$ from (3.1.25). This implies the restriction of the probability space from $(\Omega, \mathcal{E}, \mathbb{P})$ to $(\Omega, \sigma(\xi_1, \ldots, \xi_M), \mathbb{P})$ as mentioned above. Only in cases, when we want to emphasize the embedding in $\mathbb{R}^M$ attained by (3.1.22), we will use the notation $\Omega^{(M)}$ with coordinates $y \in \Omega^{(M)}$ or $\xi \in \Omega^{(M)}$, e.g. when looking at the space $C^0(\Omega^{(M)}) = \{v : \Omega^{(M)} \to \mathbb{R} : v \text{ continuous}\}$ in Section 3.3.

The first question arising from (3.1.26) is whether this problem is still well-posed. In view of (2.3.1) this means whether K_M is bounded and uniformly coercive, i.e. whether there exist $K_{\min}$, $K_{\max} \in (0, \infty)$ such that

$$\mathbb{P}(K_M(x, \omega) \in [K_{\min}, K_{\max}] \quad \forall x \in \overline{D}) = 1. \tag{3.1.29}$$

The convergence $\|K_M - K\|_{L^\infty(D) \otimes L^2(\Omega)} \to 0$ as $M \to \infty$ in combination with (2.3.1) is of no use, since, first, we are interested in truncations K_M for small M (note that the time-discrete problem will be solved in $d + M$ dimensions later on) and, secondly, we remember the Gibbs phenomenon for Fourier series. To circumvent this, one normally imposes assumptions on the ratio of the expectation $\bar{K}(x)$ and the random part $\tilde{K}(x, \cdot)$ (see [9] or [103]) to ensure (3.1.29) while supposing in addition the convergence $K_M \to K$ in $L^\infty(D) \otimes L^\infty(\Omega)$. For example, the assumption in [9] reads

$$\sigma_{0,M}(x) < \bar{K}(x) - K_{\min} \quad \text{for almost all } x \in D \tag{3.1.30}$$

with $\sigma_{0,M} : D \to \mathbb{R}$ given by

$$\sigma_{0,M}(x) = \sum_{r=1}^{M} \sqrt{\lambda_r} |g_r(x)| \alpha_r,$$

where $\Omega_r = \xi_r(\Omega) \subset (-\alpha_r, \alpha_r)$ is supposed to be bounded.

The violation of (3.1.30) can occur even in "nice" situations, see the example in [9, p. 11] for the exponential covariance kernel from Example 3.1.11.

The independence of the random variables $\{\xi_r\}$ is again the key to the problem.

Lemma 3.1.24 ([103])**.** *Let K_M be the truncated KL expansion (3.1.11) of K. If the random variables $\xi_1, \ldots, \xi_M$ are independent, then it holds*

$$K_M = \mathbb{E}[K | \sigma(\xi_1, \ldots, \xi_M)].$$

Proof. Let $M' > M$. Due to the independence of $\{\xi_r\}$ and (3.1.19), one can state the conditional expectations $\mathbb{E}[\xi_r | \sigma(\xi_1, \ldots, \xi_M)] = \mathbb{E}[\xi_r] = 0$ for $M < r \leq M'$ and hence

$$\mathbb{E}[K_{M'} | \sigma(\xi_1, \ldots, \xi_M)] = K_M. \tag{3.1.31}$$

For an arbitrary $A \in \sigma(\xi_1, \ldots, \xi_M)$, it follows by (3.1.31) and (3.1.12) that

$$\begin{aligned} \int_D \left(\int_A (K_M - K) \, \mathrm{d}\mathbb{P} \right)^2 \mathrm{d}x = \int_D \left(\int_A (K_{M'} - K) \, \mathrm{d}\mathbb{P} \right)^2 \mathrm{d}x \\ \leq \int_D \int_A (K_{M'} - K)^2 \, \mathrm{d}\mathbb{P} \, \mathrm{d}x \to 0 \end{aligned}$$

as $M' \to \infty$. This proofs the assertion. □

Corollary 3.1.25. *If the random variables $\xi_1, \ldots, \xi_M$ are independent, then it holds*

$$\mathbb{P}(K_M(x, \omega) \in [K_{\min}, K_{\max}] \quad \forall x \in \overline{D}) = 1.$$

	$\mathbb{P}(\xi_r \notin \Omega_{r,0})$	estimated by (3.1.32) ($m = 2$)
$\Omega_{r,0} = [-2, 2]$	0.0455	0.25
$\Omega_{r,0} = [-3, 3]$	0.0027	0.111
$\Omega_{r,0} = [-5, 5]$	$5.7 \cdot 10^{-7}$	0.04
$\Omega_{r,0} = [-10, 10]$	$1.5 \cdot 10^{-23}$	0.01

Table 3.1: Loss of information by truncating using $\xi_r \propto \mathcal{N}(0, 1)$.

Proof. This follows from (2.3.1) by Lemma 3.1.24 and the monotonicity of conditional expectations [65, Theorem 8.14], since

$$K_M = \mathbb{E}[K|\sigma(\xi_1, \ldots, \xi_M)] \leq \mathbb{E}[K_{\max}|\sigma(\xi_1, \ldots, \xi_M)] = K_{\max} \quad \text{a.s.},$$

and analogously for $K_{\min}$. □

Remark 3.1.26. At the end of this subsection, we consider truncations of the sets Ω_r. This is helpful, since the restriction (3.1.29) stipulates that all the Ω_r are bounded, more precisely, that there exists an α_r such that

$$\mathbb{P}(\{\omega : |\xi_r(\omega)| < \alpha_r\}) = 1$$

for all $r = 1, \ldots, M$. This would ban many distributions including Gaussian distributed ξ_r although they are widely used in this context (cf. e.g. [47, 75, 92, 113]). This truncation is now "cutting off" the tails of these distributions without affecting their approximation qualities. This is possible due to the trivial result that for all $\varepsilon > 0$, there is a compact set $C \subset \Omega$ with $\mathbb{P}(\Omega \backslash C) < \varepsilon$ [65, Lemma 13.5].

If $\Omega_{r,0} \subset \Omega_r$ and $\Omega_0 = \prod_{r=1}^{M} \Omega_{r,0} \subset \Omega^{(M)}$ denotes the subdomain, the loss of information by truncation can be estimated by Markov's inequality (see [78, p. 158]), which reads

$$\frac{\mathbb{E}[|X|^m] - \varepsilon^m}{\||X|^m\|_{L^\infty(\Omega)}} \leq \mathbb{P}[|X| \geq \varepsilon] \leq \frac{\mathbb{E}[|X|^m]}{\varepsilon^m}$$

with an $m > 0$ for a random variable $X : \Omega \to \mathbb{R}$. With $\Omega_0 = \prod_{r=1}^{M} [-\alpha_r, \alpha_r]$, where $\alpha_1, \ldots, \alpha_M$ are positive real numbers, it provides

$$\mathbb{P}(\xi \notin \Omega_0) = 1 - \prod_{r=1}^{M} \mathbb{P}(|\xi_r| \leq \alpha_r) \leq 1 - \prod_{r=1}^{M} \left(1 - (\alpha_r)^{-m} \mathbb{E}[|\xi_r|^m]\right). \tag{3.1.32}$$

Like the related Chebyshev's inequality, Markov's inequality is in most cases not very strict, confer Table 3.1. The inequality (3.1.32) shows in particular the convergence of $\mathbb{P}(\xi \notin \Omega_0)$ to zero when $\min_{1 \leq r \leq M} \alpha_r \to \infty$. The consequences for (3.1.26) when applying the truncation onto Ω_0 follows immediately: use the expectation

$$\mathbb{E}\left[u(t, x, \xi)\mathbf{1}_{\{\xi \in \Omega_0\}}\right] = \mathbb{E}\left[u(t, x, \xi)|\xi \in \Omega_0\right] \mathbb{P}(\xi \in \Omega_0)$$

instead of the original $\mathbb{E}[u(t, x, \xi)]$.

In the discretization as well as in our computations, this truncation will however not be necessary, cf. the procedure in Subsections 3.3.3 and 3.3.4.

Remark 3.1.27. In hydrological applications, the permeability K is often assumed to be lognormally distributed, see Remark 2.3.8 for a discussion. To approximate such a function K, it is possible to use the methods described in Remark 3.1.20 with

$G(\cdot) = \exp(\cdot)$ in (3.1.20), cf. [64], or with a polynomial chaos approach as in (3.1.21), cf. [53] for approximation results. The most practicable way (used in [13, 40, 87]) in accordance with the common practice in hydrology (see e.g. [27, 39, 57]) is, however, to consider $\tilde{K} = \log(K)$ and to perform a Karhunen–Loève expansion for $\tilde{K}$. This results in

$$K_M(x,\omega) = \exp\left(\bar{K}(x) + \sum_{r=1}^{M} \sqrt{\lambda_r} g_r(x) \xi_r(\omega)\right) \tag{3.1.33}$$

with Gaussian ξ_r and the corresponding probability space $(\Omega^{(M)}, \mathrm{Bor}(\Omega^{(M)}), \mathbb{P}^{(M)})$. Our approach to solve the stochastic Richards equation will work as well if (3.1.11) is replaced by (3.1.33), as will be carried out where needed.

Remark 3.1.28. The Karhunen–Loève expansion was introduced for the permeability function K, as the notation throughout this section suggests. However, it can be applied also to other parts of the Richards equation, even to Dirichlet and Neumann boundary conditions (cf. [96]) and in particular to the right-hand side $f(x,\omega)$. This was done for example in [12, 13, 32, 82] to obtain

$$f(x,\omega) = \bar{f}(x) + \sum_{r=1}^{M^{(f)}} \sqrt{\lambda_r^{(f)}} g_r^{(f)} \zeta_r(\omega), \tag{3.1.34}$$

and one has to assume that $\{\zeta_r\}$ is independent of $\{\xi_r\}$ (which is supported by the hydrological context). Inserting the truncated KL expansion of f in (3.1.26) minimizes the computational effort of our Galerkin approach in Subsection 4.1.1 in a minor way, whereas it yields a further discretization error and a substantial effort to compute (3.1.34). For that reason, we will not pursue this idea.

3.1.3 Computational aspects

In this subsection, we go into some details about the computation of the KL expansion. One has to distinguish between two major tasks: first, parameters like the covariance have to be estimated from experimental data, and, secondly, the first M functions g_r and ξ_r in (3.1.10) have to be computed in an efficient way.

We start with the latter problem and assume that the expectation value $\bar{K}(x)$ and the covariance function $V_K(x_1, x_2)$ from Definition 3.1.1 are known. The functions g_r and numbers λ_r are the eigenfunctions and eigenvalues of the operator $\mathcal{V}_K$ on $L^2(D)$ from (3.1.13), i.e.

$$\int_D \int_D V_K(x_1,x_2) g_r(x_2) \varphi(x_1) \,\mathrm{d}x_2 \,\mathrm{d}x_1 = \int_D \lambda_r g_r(x_1) \varphi(x_1) \,\mathrm{d}x_1 \quad \forall \varphi \in L^2(D) \tag{3.1.35}$$

in weak formulation. Apart from some special cases (e.g. [47, Section 2.3]), exact solutions are not known, and numerical approximations are necessary. To this end, we introduce the finite element space $\mathcal{S}_h \subset L^2(D)$ consisting of all continuous functions in $L^2(D)$ which are linear on each triangle $\mathrm{t} \in \mathcal{T}_h$ for a given triangulation $\mathcal{T}_h$ of D with vertices $\mathcal{N}_h$ and spanned by the nodal basis

$$\Lambda_h := \{s_p : p \in \mathcal{N}_h\},$$

the elements s_p of which are determined by $s_p(q) = \delta_{pq}$ for all $p, q \in \mathcal{N}_h$. This discretization leads to the generalized eigenvalue problem

$$\mathsf{V}\mathsf{g} = \lambda \mathsf{M}\mathsf{g}, \tag{3.1.36}$$

where the matrix V with entries

$$\mathsf{V}_{ij} = \int_D \int_D s_{p_i}(x_1) V_K(x_1, x_2) s_{p_j}(x_2) \, \mathrm{d}x_2 \, \mathrm{d}x_1$$

and the mass matrix M with entries

$$\mathsf{M}_{ij} = \int_D s_{p_i}(x) s_{p_j}(x) \, \mathrm{d}x$$

are both symmetric and positive definite.

Since the number M of approximate eigenpairs in our truncated expansion (3.1.11) is typically much smaller than our discrete dimension $|\mathcal{N}_h|$, iterative solvers like Krylov methods (see, e.g., [93]) are suggested. Moreover, one has to take into account that V is (in contrast to M) a dense matrix in general. For research in this context, we refer to Eiermann et al. [32] who investigated the use of Lanczos-based methods in this setting and Schwab and Todor [95] who created a solver based on the idea of fast multipole methods.

The fact that V is a dense matrix and the limits to the size of $\mathcal{N}_h$ which are effected thereby have another consequence. It is reasonable to use two different discretizations for our spatial domain D for the KL expansion on the one hand and the approximation of the SPDE introduced in 3.3.2 on the other hand or at least to solve the KL eigenproblem on a distinctly coarser grid. Hence, the discrete functions g should be stored in a way that they can be transferred easily to another grid later on. Finally, observe that once the eigenpairs (λ_r, g_r) have been computed, the functions ξ_r are obtained by (3.1.9).

For concluding this section, we provide a short insight into how to recover the expectation value $\bar{K}(x)$ and the covariance function $V_K(x_1, x_2)$ from experimental data. Let $(x_i)_i \in \mathbb{R}^{N_D}$ be a vector of measuring points $x_i \in D$ for $i = 1, \ldots, N_D$. We assume that we can make observations of N_Ω independent realizations of the random field K at these points and store the values in a matrix $\mathsf{K} \in \mathbb{R}^{N_D \times N_\Omega}$ with

$$\mathsf{K}_{il} = K(x_i, \omega_l) \quad \text{for } i = 1, \ldots, N_D, \, l = 1, \ldots, N_\Omega.$$

We define the (vector-valued) sample mean μ^K as

$$\mu_i^K := \frac{1}{N_\Omega} \sum_{l=1}^{N_\Omega} \mathsf{K}_{il} \tag{3.1.37}$$

and the (matrix-valued) sample covariance Σ^K as

$$\Sigma_{ij}^K := \frac{1}{N_\Omega - 1} \sum_{l=1}^{N_\Omega} \left(\mathsf{K}_{il} - \mu_i^K \right) \left(\mathsf{K}_{jl} - \mu_j^K \right). \tag{3.1.38}$$

It is well known (see, e.g., [7]) that μ_i^K is an unbiased and consistent estimator for $\bar{K}(x_i)$, i.e.

$$\mathbb{E}\left[\mu_i^K\right] = \bar{K}(x_i) \quad \text{and} \quad \lim_{N_\Omega \to \infty} \mathbb{P}\left(|\mu_i^K - \bar{K}(x_i)| < \varepsilon \right) = 1 \quad \forall \varepsilon > 0,$$

and that Σ_{ij}^K is an unbiased and consistent estimator for $V_K(x_i, x_j)$, i.e.

$$\mathbb{E}\left[\Sigma_{ij}^K\right] = V_K(x_i, x_j) \quad \text{and} \quad \lim_{N_\Omega \to \infty} \mathbb{P}\left(|\Sigma_{ij}^K - V_K(x_i, x_j)| < \varepsilon \right) = 1 \quad \forall \varepsilon > 0.$$

It is now possible to insert the estimated covariance into (3.1.35), but it is usually more expedient to compute an approximation of the truncated KL expansion directly from μ_i^K and Σ_{ij}^K by performing a spectral decomposition. For details of this method, the so-called principal component analysis (PCA), see [70]. Finally, we refer to Babuška et al. [10] who shed light on the interplay between PCA and the stochastic diffusion equation.

3.2 Time discretization and convex minimization

In this section, the time discretization for the Richards equation is presented. The resulting spatial and stochastic problems will be rewritten as minimization problems and variational inclusions, which points the way to how our numerical solution will be configured. We adopt the approach for the deterministic Richards equation presented in [18, Sections 2.3 and 2.4].

3.2.1 Time discretization

The starting point is the variational inequality (3.1.26) as the weak formulation of the Kirchhoff-transformed stochastic Richards equation (2.3.3)–(2.3.5) for any $t \in (0,T]$ after modeling the permeability K by the truncated Karhunen–Loève expansion (3.1.11). For notational reasons, we omit the truncation parameter M and set $f(t) = 0$ and $f_N(t) = 0$ for all $t \in (0,T]$, cf. Remark 3.2.12. The problem is then to find $u(t) \in \mathcal{K}(t)$ solving

$$\begin{aligned}\mathbb{E}\left[\int_D H(u(t))_t\,(v-u(t))\,\mathrm{d}x\right] + \mathbb{E}\left[\int_D K\nabla u(t)\nabla\,(v-u(t))\,\mathrm{d}x\right]\\ \geq \mathbb{E}\left[\int_D K\,kr(H(u(t)))\rho g e_z \nabla\,(v-u(t))\,\mathrm{d}x\right] \quad \forall v \in \mathcal{K}(t). \quad (3.2.1)\end{aligned}$$

Let $0 = t_0 < t_1 < \ldots < t_{N_T} = T$ be a partition of the time interval $[0,T]$ and denote by $\tau_n = t_n - t_{n-1}$ for $n = 1,\ldots,N_T$ the time step size. We choose our time discretization to be implicit in the diffusion part on the left-hand side and explicit in the convective part on the right-hand side of (3.2.1). The explicit treatment of this convective term coming from the gravitation allows a reformulation in terms of convex minimization as it is carried out in the sequel. Taking the backward Euler for the implicit discretization and substituting the time derivative $H(u(t))_t$ in (3.2.1) by the corresponding differential quotient

$$\frac{H(u(t_n)) - H(u(t_{n-1}))}{\tau_n},$$

we achieve the following time-discrete version of (3.2.1): find $u_n \in \mathcal{K}(t_n)$ with

$$\begin{aligned}\mathbb{E}\left[\int_D H(u_n)\,(v-u_n)\,\mathrm{d}x\right] + \tau_n\mathbb{E}\left[\int_D K\nabla u_n\nabla\,(v-u_n)\,\mathrm{d}x\right]\\ \geq \mathbb{E}\left[\int_D H(u_{n-1})\,(v-u_n)\,\mathrm{d}x\right] + \tau_n\mathbb{E}\left[\int_D K\,kr(H(u_{n-1}))\rho g e_z\nabla\,(v-u_n)\,\mathrm{d}x\right]\\ \forall v \in \mathcal{K}(t_n), \quad (3.2.2)\end{aligned}$$

where u_n is an approximation of $u(t_n)$. In the following, we abstract from the dependence on t and regard (3.2.2) as a steady-state inequality. To that effect, we

set $\mathcal{K} := \mathcal{K}(t_n)$, $\Gamma_D := \Gamma_D(t_n)$ and $\Gamma_N := \Gamma_N(t_n)$ and obtain

$$\mathcal{K} = \{v \in H^1(D) \otimes L^2(\Omega) : v \geq u_c \wedge \mathrm{tr}_{\Gamma_D}\, v = u_D\}. \tag{3.2.3}$$

Recall from Section 2.3 that, with u_D as in (2.3.16), the set $\mathcal{K}$ is a nonempty, closed and convex subset of $H^1(D) \otimes L^2(\Omega)$.

We now take a closer look at the structure of (3.2.2). To this end, we define the bilinear form $a(\cdot, \cdot)$ by

$$a(v, w) := \tau_n \mathbb{E}\left[\int_D K \nabla v \nabla w \,\mathrm{d}x\right] \quad \forall v, w \in H^1(D) \otimes L^2(\Omega) \tag{3.2.4}$$

and recall the definitions of the norms $\|\cdot\|_{0,0}$ and $\|\cdot\|_{1,0}$ from Section 2.2, viz.

$$\|v\|_{0,0} = \mathbb{E}\left[\int_D v^2 \,\mathrm{d}x\right]^{1/2}, \quad \|v\|_{1,0} = \left(\mathbb{E}\left[\int_D (\nabla v)^2 \,\mathrm{d}x\right] + \mathbb{E}\left[\int_D v^2 \,\mathrm{d}x\right]\right)^{1/2}.$$

In light of (3.1.29), $a(\cdot, \cdot)$ is continuous on $H^1(D) \otimes L^2(\Omega)$ with

$$|a(v, w)| \leq \tau_n K_{\max} \|\nabla v\|_{0,0} \|\nabla w\|_{0,0} \leq C \|v\|_{1,0} \|w\|_{1,0} \quad \forall v, w \in H^1(D) \otimes L^2(\Omega). \tag{3.2.5}$$

On the other hand, we get

$$a(v, v) \geq \tau_n K_{\min} \|\nabla v\|_{0,0}^2 \quad \forall v \in H^1(D) \otimes L^2(\Omega).$$

By means of the Poincaré inequality [89, p. 340]

$$\int_D v^2 \,\mathrm{d}x \leq C \int_D (\nabla v)^2 \,\mathrm{d}x \quad \forall v \in H^1_{\Gamma_D}(D)$$

on the space

$$H^1_{\Gamma_D}(D) := \{v \in H^1(D) : \mathrm{tr}_{\Gamma_D}\, v = 0\}, \tag{3.2.6}$$

we obtain the coercivity of $a(\cdot, \cdot)$ on $H^1_{\Gamma_D}(D) \otimes L^2(\Omega)$, i.e. there exists a $c > 0$ such that

$$a(v, v) \geq c \|v\|_{1,0}^2 \quad \forall v \in H^1_{\Gamma_D}(D) \otimes L^2(\Omega). \tag{3.2.7}$$

In order to fulfill the Dirichlet boundary conditions, the customary procedure is to look first for a $w \in H^1(D) \otimes L^2(\Omega)$ with $\mathrm{tr}_{\Gamma_D}\, w = u_D$ and to combine it with an appropriate $\tilde{v} \in H^1_{\Gamma_D}(D) \otimes L^2(\Omega)$ to obtain $v = w + \tilde{v}$. The energy norm of the latter function can be estimated from below by

$$\begin{aligned} a(v, v) &= a(\tilde{v}, \tilde{v}) + 2a(w, \tilde{v}) + a(w, w) \\ &\geq c\|\tilde{v}\|_{1,0}^2 - 2C\|w\|_{1,0}\|\tilde{v}\|_{1,0} - C\|w\|_{1,0}^2 \\ &\geq c\|v\|_{1,0}^2 - 2(C + c)\|w\|_{1,0}\|v\|_{1,0} - (3C - c)\|w\|_{1,0}^2 \end{aligned}$$

using (3.2.5), (3.2.7), and the triangle inequality. Furthermore, we can conclude $\mathcal{K} \subset w + \left(H^1_{\Gamma_D}(D) \otimes L^2(\Omega)\right)$ from $\mathcal{K} - w \subset H^1_{\Gamma_D}(D) \otimes L^2(\Omega)$. We summarize this for sake of quotation.

Lemma 3.2.1. *Let $w \in H^1(D) \otimes L^2(\Omega)$ be a function with $\mathrm{tr}_{\Gamma_D}\, w = u_D$. Then we have $\mathcal{K} \subset w + H^1_{\Gamma_D}(D) \otimes L^2(\Omega)$, and there exist constants c_1, $c_2 > 0$ such that*

$$a(v, v) \geq c\|v\|_{1,0}^2 - c_1\|v\|_{1,0} - c_2 \quad \forall v \in w + H^1_{\Gamma_D}(D) \otimes L^2(\Omega)$$

with the constant c from (3.2.7).

We now turn to the right-hand side of (3.2.2) and define the linear form ℓ on $H^1(D) \otimes L^2(\Omega)$ by

$$\ell(v) := \mathbb{E}\left[\int_D H(u_{n-1})v \,\mathrm{d}x\right] + \tau_n \mathbb{E}\left[\int_D K\, kr(H(u_{n-1}))\rho g e_z \nabla v \,\mathrm{d}x\right] \tag{3.2.8}$$

for all $v \in H^1(D) \otimes L^2(\Omega)$. If $H : [u_c, \infty) \to \mathbb{R}$ and $kr : H([u_c, \infty)) \to \mathbb{R}$ are monotonically increasing and bounded functions and the permeability K satisfies (3.1.29), then $\ell \in \big(H^1(D) \otimes L^2(\Omega)\big)'$. Replacing u_n by u, we can rewrite the variational inequality (3.2.2) as

$$u \in \mathcal{K}: \ \mathbb{E}\left[\int_D H(u)(v-u) \,\mathrm{d}x\right] + a(u, v-u) - \ell(v-u) \geq 0 \quad \forall v \in \mathcal{K} \tag{3.2.9}$$

with $\mathcal{K}$ defined in (3.2.3).

3.2.2 Formulation as a convex minimization problem

It remains to examine the integral in this inequality. It turns out that it can be rewritten in terms of convex functionals, which gives us the possibility to apply methods of convex minimization. Therefore, we recall at the beginning some fundamental definitions.

Definition 3.2.2 ([33])**.** Let V be a real vector space and $C \subset V$ a convex set, i.e. for $y, z \in C$ and $\lambda \in (0,1)$ we have $(1-\lambda)y + \lambda z \in C$. $F : C \to \mathbb{R} \cup \{\pm\infty\}$ is said to be *convex* if for every y and z in C, we have

$$F((1-\lambda)y + \lambda z) \leq (1-\lambda)F(y) + \lambda F(z) \quad \forall \lambda \in [0,1] \tag{3.2.10}$$

whenever the right-hand side is defined, i.e. unless it is $F(z) = -F(y) = \pm\infty$. $F : C \to \mathbb{R}$ is said to be *strictly convex* if it is convex and the strict inequality holds in (3.2.10) for all $y, z \in C$, $y \neq z$ and each $\lambda \in (0,1)$.

The following lemma provides some helpful facts about convex function defined on the real line.

Lemma 3.2.3 ([67])**.** *Let $I \subset \mathbb{R}$ be an interval. Then for $f : I \to \mathbb{R}$ the following holds.*

a) f is convex if, and only if, the inequality

$$\frac{f(z) - f(z_1)}{z - z_1} \leq \frac{f(z_2) - f(z)}{z_2 - z}$$

holds for any $z_1, z, z_2 \in I$ with $z_1 < z < z_2$.

b) If f is convex, then for any $z \in I$ the difference quotient

$$\frac{f(y) - f(z)}{y - z}$$

is a monotonically increasing function of $y \in I \backslash \{z\}$.

We now define the function $\Phi : [u_c, \infty) \to \mathbb{R}$ as

$$\Phi(z) := \int_0^z H(s) \,\mathrm{d}s \quad \forall z \in [u_c, \infty). \tag{3.2.11}$$

We assume $u_c < 0$, which implies that $\Phi(0)$ is defined and equal to zero. A general form of Φ when using Brooks–Corey parameters can be found in (1.1.17). By means of Lemma 3.2.3, it is easy to prove some basic properties of Φ [18, Lemma 2.3.6].

Lemma 3.2.4. *Let $\Phi : [u_c, \infty) \to \mathbb{R}$ be defined as in (3.2.11).*

a) If H is monotonically increasing, then Φ is convex.

b) Φ is differentiable (from the right) in u_c. In addition, we have $\Phi'(u_c) = H(u_c)$ if H is continuous in u_c. Furthermore, Φ is differentiable in $z \in (u_c, \infty)$ if, and only if, H is continuous in z, which is true for all but countably many points, and in this case $\Phi'(z) = H(z)$ holds.

c) If H is bounded, then Φ is Lipschitz continuous with Lipschitz constant $\|H\|_\infty$.

Note that part b) in Lemma 3.2.4 states that differentiable Φ is continuously differentiable at the same time; this assertion then holds for all convex functions since every convex function (with bounded image) has, except for an additive constant, a representation as in (3.2.11).

Using the convex function Φ, we define the functional $\phi : \mathcal{K} \to \mathbb{R}$ by

$$\phi(v) := \mathbb{E}\left[\int_D \Phi\left(v(x, \xi(\omega))\right) \, \mathrm{d}x\right] \quad \forall v \in \mathcal{K}. \tag{3.2.12}$$

In the sequel, we will connect ϕ arising from (3.2.12) to the variational inequality (3.2.9). It is convenient to assume that Φ is Lipschitz continuous. A more general setting (with less strict conditions on the underlying function Φ) is regarded in [68].

Proposition 3.2.5. *If Φ is a convex function, then $\phi : \mathcal{K} \to \mathbb{R}$ is a convex functional. If, in addition, Φ is Lipschitz continuous, then ϕ is also Lipschitz continuous and satisfies*

$$|\phi(v)| \leq C\|v\|_{1,0} \quad \forall v \in \mathcal{K} \tag{3.2.13}$$

with a $C > 0$.

Proof. The convexity of ϕ follows directly from the convexity of the function Φ. To prove the Lipschitz continuity, let $v, w \in \mathcal{K}$. Then, by means of the Cauchy–Schwarz inequality in $L^2(D) \otimes L^2(\Omega)$, it holds

$$\begin{aligned}
|\phi(v) - \phi(w)| &\leq \mathbb{E}\left[\int_D \left|\Phi\left(v(x, \xi(\omega))\right) - \Phi\left(w(x, \xi(\omega))\right)\right| \mathrm{d}x\right] \\
&\leq L \cdot \mathbb{E}\left[\int_D \left|v(x, \xi(\omega)) - w(x, \xi(\omega))\right| \mathrm{d}x\right] \\
&\leq L \cdot \|\mathbf{1}\|_{0,0} \, \|v - w\|_{0,0} \\
&\leq C\|v - w\|_{1,0},
\end{aligned}$$

where L is the Lipschitz constant of Φ. Since $\Phi(0) = 0$ implies $\phi(0) = 0$, we get (3.2.13). □

The following definition can be found, e.g., in [33].

Definition 3.2.6. Let $F : U \to \mathbb{R} \cup \{\pm\infty\}$ be defined on a subset $U \subset V$ of a normed space V with $u \in U$ and $v \in V$.

a) If there is an $\varepsilon > 0$ such that $u + \lambda v \in U$ for all $\lambda \in [0, \varepsilon]$, we call

$$\partial_v F(u) := \lim_{\lambda \downarrow 0} \frac{F(u + \lambda v) - F(u)}{\lambda} \tag{3.2.14}$$

the *directional derivative of F at u in the direction of v* if this limit exists.

b) If there exists a $u' \in V'$ such that

$$\partial_v F(u) = {}_{V'}\langle u', v\rangle_V \quad \forall v \in V, \tag{3.2.15}$$

then we say that F is *Gâteaux-differentiable at* u, call u' the *Gâteaux-derivative at* u *of* F and denote it by $F'(u)$.

The uniqueness of the Gâteaux-derivative is obvious from (3.2.15). The case of convex functions if of special interest since the fraction on the right-hand side of (3.2.14) is in that instance a monotone function of λ, which means that this expression always has a limit which can be, however, $\pm\infty$. This fact is utilized in the following proposition.

Proposition 3.2.7. *Let $\Phi : [u_c, \infty) \to \mathbb{R}$ be convex and differentiable. Then, for any $u, v \in \mathcal{K}$ the directional derivative $\partial_{v-u}\phi(u)$ exists and can be written as*

$$\partial_{v-u}\phi(u) = \mathbb{E}\left[\int_D \Phi'\left(u(x,\xi(\omega))\right)\left(v(x,\xi(\omega)) - u(x,\xi(\omega))\right)\,\mathrm{d}x\right] \tag{3.2.16}$$

or, equivalently, as

$$\partial_{v-u}\phi(u) = \mathbb{E}\left[\int_D H\left(u(x,\xi(\omega))\right)\left(v(x,\xi(\omega)) - u(x,\xi(\omega))\right)\,\mathrm{d}x\right]. \tag{3.2.17}$$

Proof. Recall the short hand notation from Remark 3.1.23 with $\xi = \xi(\omega)$. For all $u, v \in \mathcal{K}$, we have $u+\lambda(v-u) \in \mathcal{K}$ for $\lambda \in [0,1]$ since $\mathcal{K}$ is convex. Setting $w := v-u$, we have to look at the difference quotient

$$\frac{\phi(u+\lambda w) - \phi(u)}{\lambda} = \int_\Omega \int_D \frac{\Phi(u(x,\xi) + \lambda w(x,\xi)) - \Phi(u(x,\xi))}{\lambda}\,\mathrm{d}x\,\mathrm{d}\mathbb{P} \tag{3.2.18}$$

as $\lambda \downarrow 0$. By Lemma 3.2.3 b) we obtain

$$\frac{\Phi(u(x,\xi) + \lambda w(x,\xi)) - \Phi(u(x,\xi))}{\lambda} \leq \frac{\Phi(u(x,\xi) + w(x,\xi)) - \Phi(u(x,\xi))}{1} =: G_1(x,\omega)$$

and

$$\frac{\Phi(u(x,\xi) + \lambda w(x,\xi)) - \Phi(u(x,\xi))}{\lambda} \geq \frac{\Phi(u(x,\xi) - w(x,\xi)) - \Phi(u(x,\xi))}{1} =: G_2(x,\omega)$$

for $w(x,\xi) = w(x,\xi(\omega)) \geq 0$ and $\lambda \in (0,1]$ and an analogous result for the case $w(x,\xi) = w(x,\xi(\omega)) \leq 0$.

The function Φ is assumed to be differentiable, thus the integrands

$$\frac{\Phi(u(x,\xi) + \lambda w(x,\xi)) - \Phi(u(x,\xi))}{\lambda} \tag{3.2.19}$$

in (3.2.18) converge to $\Phi'(u(x,\xi))w(x,\xi)$ almost everywhere in $D \times \Omega$ as $\lambda \downarrow 0$, either as a monotonically increasing sequence for $w(x,\xi) < 0$ or as a monotonically decreasing sequence for $w(x,\xi) \geq 0$ due to Lemma 3.2.3 b).

As shown above, the integrand (3.2.19) is bounded by the integrable function $\max(|G_1(\cdot,\cdot)|, |G_2(\cdot,\cdot)|)$ independently of $\lambda \in (0,1]$, thus one can apply the theorem of Lebesgue (see, e.g., [109, Theorem A.3.2]), which ensures the convergence of the integrals in (3.2.18).

Finally, the equivalence of (3.2.16) and (3.2.17) follows from $\Phi' = H$ in Lemma 3.2.4 b). □

Under the assumption that H is continuous, the variational inequality (3.2.9) can be now written as

$$u \in \mathcal{K}: \quad \partial_{v-u}\phi(u) + a(u, v-u) - \ell(v-u) \geq 0 \quad \forall v \in \mathcal{K} \tag{3.2.20}$$

with the notation fixed in (3.2.4) and (3.2.8).

The linear terms on the left-hand side of (3.2.20) are often subsumed under a functional $\mathcal{J} : H^1(D) \otimes L^2(\Omega) \to \mathbb{R}$ defined by

$$\mathcal{J}(v) := \frac{1}{2}a(v, v) - \ell(v) \quad \forall v \in H^1(D) \otimes L^2(\Omega). \tag{3.2.21}$$

This functional $\mathcal{J}$ is quadratic and strictly convex, see [33, Remark II.1.1]. It is also continuous if H and kr are monotonically increasing and bounded functions and K satisfies (3.1.29), see (3.2.5) and the paragraph following (3.2.8). Moreover, by [33, Remark II.2.1], the Gâteaux-derivative exists in $u \in H^1(D) \otimes L^2(\Omega)$ and reads

$$\mathcal{J}'(u)(v) = \partial_v \mathcal{J}(u) = a(u, v) - \ell(v) \quad \forall v \in H^1(D) \otimes L^2(\Omega). \tag{3.2.22}$$

Combining the above results, the functional $F : \mathcal{K} \to \mathbb{R}$ defined by

$$F(v) := \mathcal{J}(v) + \phi(v) \quad \forall v \in \mathcal{K}$$

is strictly convex with existing derivative $\partial_{v-u}F(u)$ for any $u, v \in \mathcal{K}$, and (3.2.20) has the short form

$$u \in \mathcal{K}: \quad \partial_{v-u}F(u) \geq 0 \quad \forall v \in \mathcal{K}. \tag{3.2.23}$$

Now, Lemma 3.2.3 b) states

$$F(v) - F(u) \geq \frac{F(u + \lambda(v-u)) - F(u)}{\lambda}$$

for any $\lambda \in (0, 1]$. Taking the limit as $\lambda \downarrow 0$, it follows

$$F(v) - F(u) \geq \partial_{v-u}F(u) \geq 0.$$

Conversely, if $F(v) - F(u) \geq 0$ for all v, then for any $\lambda \in (0, 1]$ it holds

$$\frac{F(u + \lambda(v-u)) - F(u)}{\lambda} \geq 0.$$

Taking the limit as $\lambda \downarrow 0$, we arrive at

$$\partial_{v-u}F(u) \geq 0$$

for all v and have hence proven the equivalence of the variational inequality (3.2.23) with a convex minimization problem. We state this result in a more general form like in [18].

Proposition 3.2.8. *Let V be a real vector space, $C \subset V$ a convex set and the mapping $F : C \to \mathbb{R}$ a convex functional whose directional derivative $\partial_{v-u}F(u)$ exists for all $u, v \in C$. Then the variational inequality*

$$u \in C: \quad \partial_{v-u}F(u) \geq 0 \quad \forall v \in C \tag{3.2.24}$$

is equivalent to the minimization problem

$$u \in C: \quad F(u) \leq F(v) \quad \forall v \in C. \tag{3.2.25}$$

This reformulation as a convex minimization problem is the key to obtain the main result of this section, viz. the conclusion that there exists a unique solution to our problem (3.2.9). Therefor, we cite the general result in reflexive Banach spaces, which can be found, e.g., in [33, Prop. II.1.2]. We first provide some definition.

Definition 3.2.9. Let V be a real vector space and $C \subset V$ nonempty, closed and convex. A functional $F : C \to \mathbb{R} \cup \{+\infty\}$ is said to be *lower semicontinuous* if $\liminf_{w \to v} F(w) \geq F(v)$ holds for all $v \in C$ (with $w \in C$). A convex functional $F : C \to \mathbb{R} \cup \{+\infty\}$ is called *coercive* if for any sequence $(u_n) \subset C$ with $\|u_n\| \to \infty$ we have $F(u_n) \to +\infty$. It is said to be *proper* if it is not identically equal to $+\infty$. We call the section $\operatorname{dom} F := \{v \in K : \ F(v) < +\infty\}$ the *effective domain* of F.

Proposition 3.2.10 ([33])**.** *Let V be a reflexive Banach space, $C \subset V$ a nonempty, closed and convex subset of V. We assume that $F : C \to \mathbb{R} \cup \{+\infty\}$ is convex, lower semicontinuous and proper. We assume in addition that the set C is bounded or that the functional F is coercive over C. Then the minimization problem (3.2.25) has at least one solution. It has a unique solution if F is strictly convex over C.*

We apply this to our situation.

Theorem 3.2.11. *Let $\mathcal{K} \subset H^1(D) \otimes L^2(\Omega)$, $a(\cdot, \cdot)$ and $\ell(\cdot)$ be defined as in (3.2.3), (3.2.4) and (3.2.8), respectively. If $H : [u_c, \infty) \to \mathbb{R}$ is monotonically increasing, bounded and continuous and $kr : H([u_c, \infty)) \to \mathbb{R}$ is monotonically increasing and bounded and K satisfies (3.1.29), then the variational inequality (3.2.9) has a unique solution. Furthermore, it is equivalent to the minimization problem*

$$u \in \mathcal{K} : \quad \mathcal{J}(u) + \phi(u) \leq \mathcal{J}(v) + \phi(v) \quad \forall v \in \mathcal{K} \tag{3.2.26}$$

with $\mathcal{J}$ and ϕ as defined in (3.2.21) and in (3.2.12), respectively.

Proof. By Proposition 3.2.8, the last assertion is clear and it suffices to show that $\mathcal{K}$ and $F = \mathcal{J} + \phi$ satisfy the conditions of Proposition 3.2.10 on the reflexive Hilbert space $H^1(D) \otimes L^2(\Omega)$. The conditions on $\mathcal{K}$ were verified in Proposition 2.3.4.

The functional F is strictly convex on $\mathcal{K}$ since $\mathcal{J}$ is strictly convex and ϕ is convex. F is proper and continuous as $\mathcal{J}$ and ϕ are, the latter according to Proposition 3.2.5. It remains to check the coercivity of F, which follows from Lemma 3.2.1 and (3.2.13) for $v \in \mathcal{K}$ by

$$\mathcal{J}(v) + \phi(v) \geq \frac{1}{2} a(v, v) - |\ell(v)| - |\phi(v)| \geq \frac{1}{2} c \|v\|_{1,0}^2 - (c_1 + \|\ell\| + C)\|v\|_{1,0} - c_2 \to \infty \tag{3.2.27}$$

as $\|v\|_{1,0} \to \infty$. ∎

Before we proceed by relaxing some conditions on the occurring functions, we give some remarks.

Remark 3.2.12. At the beginning of this section, we set $f_N = 0$ and $f = 0$ for notational reasons. If we drop this assumption, Theorem 3.2.11 remains valid if we guarantee that the functional ℓ, to which both functions solely contribute, is still continuous. This is achieved for $f \in L^2(D) \otimes L^2(\Omega)$ and $f_N \in L^2(\Gamma_N) \otimes L^2(\Omega)$, which holds according to the assumptions in (2.3.15), by means of the Cauchy–Schwarz inequality and the trace theorem in [22, p. 1.61].

Remark 3.2.13. As a further generalization, we can insert a space-dependent porosity function $\mathfrak{n} = \mathfrak{n}(x)$. If $\mathfrak{n}$ is nonnegative and bounded, we just define the

functional ϕ by

$$\phi(v) := \mathbb{E}\left[\int_D \mathfrak{n}(x)\Phi\left(v(x,\xi(\omega))\right)\,\mathrm{d}x\right] \quad \forall v \in \mathcal{K}$$

instead of (3.2.12). All arguments in this section remain valid, and it is only needed to replace $\Phi(\cdot)$ and $H(\cdot)$ by $\mathfrak{n}(\cdot)\Phi(\cdot)$ and $\mathfrak{n}(\cdot)H(\cdot)$, respectively, in the according terms, in particular in (3.2.16) and (3.2.17).

Remark 3.2.14. Under the assumptions of Theorem 3.2.11, the map

$$v \mapsto \mathbb{E}\left[\int_D H\left(u(x,\xi(\omega))\right)\, v(x,\xi(\omega))\,\mathrm{d}x\right]$$

is a bounded linear functional on $H^1(D) \otimes L^2(\Omega)$. Consequently, ϕ in Proposition 3.2.7 is Gâteaux-differentiable and so is $F = \mathcal{J} + \phi$.

In that instance, the strict inequality in (3.2.23) can only occur for $u \in \partial\mathcal{K}$. For an inner point $u \in \operatorname{int}\mathcal{K}$ and an $\varepsilon > 0$ such that

$$B_\varepsilon(u) := \{v \in H^1(D) \otimes L^2(\Omega) : \|u - v\|_{1,0} < \varepsilon\} \subset \operatorname{int}\mathcal{K},$$

we take a $w \in B_\varepsilon(u)$ and obtain

$$\partial_{w-u}F(u) \geq 0 \quad \text{as well as} \quad \partial_{-(w-u)}F(u) \geq 0.$$

This leads to the problem

$$F'(u) = 0$$

if $u \in \operatorname{int}\mathcal{K}$. Note that this is always the case if, for example, $\mathcal{K} = H^1(D) \otimes L^2(\Omega)$.

Looking at the definition of $\mathcal{K}$ in (3.2.3), we detect the obstacle condition $v \geq u_c$, which results from the Kirchhoff transformation. If, however, $H : \mathbb{R} \to \mathbb{R}$ is defined on the whole real line—this occurs for instance in certain hydrological limit cases as described in [18, Section 1.4]—then we obtain the variational equality

$$\tilde{u} \in H^1_{\Gamma_D}(D) \otimes L^2(\Omega) : \ \mathbb{E}\left[\int_D H(w+\tilde{u})\, v\,\mathrm{d}x\right] + a(w+\tilde{u}, v) - \ell(v) = 0 \quad \forall v \in H^1_{\Gamma_D}(D) \otimes L^2(\Omega)$$

for a $w \in H^1(D) \otimes L^2(\Omega)$ with $\operatorname{tr}_{\Gamma_D} w = u_D$. For $u_D \equiv 0$, this simplifies to

$$u \in H^1_{\Gamma_D}(D) \otimes L^2(\Omega) : \ \mathbb{E}\left[\int_D H(u)\, v\,\mathrm{d}x\right] + a(u,v) - \ell(v) = 0 \ \ \forall v \in H^1_{\Gamma_D}(D) \otimes L^2(\Omega). \tag{3.2.28}$$

Remark 3.2.15. Finally, we recall that condition (3.1.29) in Theorem 3.2.11 is satisfied by (2.3.1) and Assumption 3.1.21 according to Corollary 3.1.25.

3.2.3 Variational inclusions

Having a closer look on Theorem 3.2.11, we detect that the continuity of H is not necessary to ensure the coercivity of $\mathcal{J} + \phi$ and therefore the existence of a unique solution but only for the equivalence of the variational inequality (3.2.9) and the minimization problem (3.2.26). In this subsection, the consequences of having a possibly uncontinuous H are investigated more specifically. In the course of this, we arrive at a reformulation of problem (3.2.26) by introducing subdifferentials and

disposing of the convex set $\mathcal{K}$. As a collateral benefit, we are able to proof the well-posedness of (3.2.26).

For the rest of this section, we use the assumptions of Theorem 3.2.11 except for the continuity of H. In view of Lemma 3.2.4 b) and Proposition 3.2.7, the derivative $\partial_v \phi(u)$ needs no longer exist. However, one can proceed by using the following generalization of Proposition 3.2.8, the proof of which uses the same arguments and can be found in [33, Prop. II.2.2].

Proposition 3.2.16. *Let V and C be as in Proposition 3.2.8 and $F = F_1 + F_2$, where F_1, $F_2 : C \to \mathbb{R} \cup \{+\infty\}$ are convex and F_1 possesses directional derivatives $\partial_{v-u} F_1(u)$ for all u, $v \in C$. Then*

$$u \in C: \quad \partial_{v-u} F_1(u) + F_2(v) - F_2(u) \geq 0 \quad \forall v \in C$$

is equivalent to

$$u \in C: \quad (F_1 + F_2)(u) \leq (F_1 + F_2)(v) \quad \forall v \in C.$$

Corollary 3.2.17. *The minimization problem (3.2.26) is equivalent to the variational inequality*

$$u \in \mathcal{K}: \ a(u, v-u) - \ell(v-u) + \phi(v) - \phi(u) \geq 0 \quad \forall v \in \mathcal{K}. \tag{3.2.29}$$

Proof. Take $C = \mathcal{K}$, $F_1 = \mathcal{J}$ and $F_2 = \phi$ in Proposition 3.2.16. □

Remark 3.2.18. In the course of this section, we started with (2.3.14) and performed first a transformation by means of the KL expansion and afterwards the time discretization to obtain (3.2.29). If one performed the transformation by means of the KL expansion directly on the time-discretized stochastic Richards equation (2.3.11), one would achieve the same result (3.2.29). This is mainly due to the fact that K is not dependent on time. This shows that the results in this chapter remain true even if u is not regular enough to formulate the inequality (2.3.14).

Remark 3.2.19. In Propositions 3.2.8 and 3.2.16, the functional F is defined on a convex set C. It is often convenient to have F defined on the whole space V. Therefor, we introduce the *canonical extension* $\bar{F} : V \to \mathbb{R} \cup \{+\infty\}$ of an $F : C \to \mathbb{R} \cup \{+\infty\}$ by setting $\bar{F}(v) = F(v)$ for all $v \in C$ and $\bar{F}(v) = +\infty$ for all $v \in V \backslash C$. Then, since we assume that C is nonempty, closed and convex, $\bar{F}$ is lower semicontinuous and proper if, and only if, F is. Moreover, u solves the minimization problem (3.2.25) if, and only if, u solves the problems

$$u \in V: \ \bar{F}(u) \leq \bar{F}(v) \quad \forall v \in V.$$

In the following, we will use the extension for Φ and ϕ with regard to the convex set $\mathcal{K}$ without indicating it by $\bar{\Phi}$ and $\bar{\phi}$ explicitly.

Before we give a reformulation of (3.2.26) in terms of subdifferentials, we achieve our purpose of abstracting from the convex set $\mathcal{K}$ defined in (3.2.3) by introducing a translation of the Dirichlet values. As in Lemma 3.2.1, we first choose a fixed

$$w \in H^1(D) \otimes L^2(\Omega) \quad \text{with } \operatorname{tr}_{\Gamma_D} w = u_D \tag{3.2.30}$$

and set $u = w + \tilde{u}$ and $v = w + \tilde{v}$. To write it in a more compact form, we introduce

$$F_w(\cdot) := F(w + \cdot) \tag{3.2.31}$$

as the translation for mappings F defined on a vector space V. Obviously, the functional $\mathcal{J}_w + \phi_w$ is convex, lower semicontinuous and proper on

$$\mathcal{K}_{\Gamma_D} := \mathcal{K} - w = \{v \in H^1_{\Gamma_D}(D) \otimes L^2(\Omega) : v \geq u_c - w\}, \tag{3.2.32}$$

and, since $\phi : H^1(D) \otimes L^2(\Omega) \to \mathbb{R} \cup \{+\infty\}$ is the extended functional, even on its superset $H^1_{\Gamma_D}(D) \otimes L^2(\Omega)$. Thus, the following is clear.

Proposition 3.2.20. *The minimization problem (3.2.26) is equivalent to*

$$\tilde{u} \in H^1_{\Gamma_D}(D) \otimes L^2(\Omega) : \quad \mathcal{J}_w(\tilde{u}) + \phi_w(\tilde{u}) \leq \mathcal{J}_w(v) + \phi_w(v) \quad \forall v \in H^1_{\Gamma_D}(D) \otimes L^2(\Omega) \tag{3.2.33}$$

in the sense that the solution u of (3.2.26) equals $w + \tilde{u}$.

We apply Proposition 3.2.16 with $V = C = H^1_{\Gamma_D}(D) \otimes L^2(\Omega)$ and $F_1 = \mathcal{J}_w$ and $F_2 = \phi_w$ to give another formulation.

Proposition 3.2.21. *The minimization problem (3.2.33) is equivalent to the variational inequality*

$$\begin{aligned} \tilde{u} \in H^1_{\Gamma_D}(D) \otimes L^2(\Omega) : \quad & a(w + \tilde{u}, v - \tilde{u}) - \ell(v - \tilde{u}) \\ & + \phi(w + v) - \phi(w + \tilde{u}) \geq 0 \quad \forall v \in H^1_{\Gamma_D}(D) \otimes L^2(\Omega). \end{aligned} \tag{3.2.34}$$

The assertion in Proposition 3.2.21 remains true if one replaces the space $H^1_{\Gamma_D}(D) \otimes L^2(\Omega)$ with the set $\mathcal{K}_{\Gamma_D}$, since

$$\tilde{u} \in \mathcal{K}_{\Gamma_D} : \quad \mathcal{J}(w + \tilde{u}) + \phi(w + \tilde{u}) \leq \mathcal{J}(w + v) + \phi(w + v) \quad \forall v \in \mathcal{K}_{\Gamma_D}$$

is equivalent to (3.2.26).

This result allows us to state that the convex minimization problem (3.2.26) is well-posed with regard to the functional ℓ.

Proposition 3.2.22. *Assume that the conditions in Theorem 3.2.11 are satisfied. Furthermore, for $i \in \{1, 2\}$, let $\ell_i \in \left(H^1(D) \otimes L^2(\Omega)\right)'$ and let u_i be the unique solutions of*

$$u_i \in \mathcal{K} : \quad \frac{1}{2} a(u_i, u_i) - \ell_i(u_i) + \phi(u_i) \leq \frac{1}{2} a(v, v) - \ell_i(v) + \phi(v) \quad \forall v \in \mathcal{K}. \tag{3.2.35}$$

Then it holds

$$\|u_1 - u_2\|_{1,0} \leq c^{-1} \|\ell_1 - \ell_2\|,$$

where c is the coercivity constant of $a(\cdot, \cdot)$ in (3.2.7).

Proof. Due to Proposition 3.2.21, we can rewrite (3.2.35) in the form

$$\tilde{u}_i \in \mathcal{K}_{\Gamma_D} : \quad a(w + \tilde{u}_i, v - \tilde{u}_i) + \phi_w(v) - \phi_w(\tilde{u}_i) \geq \ell_i(v - \tilde{u}_i) \quad \forall v \in \mathcal{K}_{\Gamma_D}.$$

We set $v = \tilde{u}_2$ for $i = 1$ and $v = \tilde{u}_1$ for $i = 2$ and obtain

$$a(w + \tilde{u}_1, \tilde{u}_1 - \tilde{u}_2) - \phi_w(\tilde{u}_2) + \phi_w(\tilde{u}_1) \leq \ell_1(\tilde{u}_1 - \tilde{u}_2) \tag{3.2.36}$$

and

$$a(-w - \tilde{u}_2, \tilde{u}_1 - \tilde{u}_2) - \phi_w(\tilde{u}_1) + \phi_w(\tilde{u}_2) \leq -\ell_2(\tilde{u}_1 - \tilde{u}_2). \tag{3.2.37}$$

Adding (3.2.36) and (3.2.37), one gets

$$a(\tilde{u}_1 - \tilde{u}_2, \tilde{u}_1 - \tilde{u}_2) \leq (\ell_1 - \ell_2)(\tilde{u}_1 - \tilde{u}_2)$$

and thus the assertion by the coercivity (3.2.7) and $\|u_1 - u_2\|_{1,0} = \|\tilde{u}_1 - \tilde{u}_2\|_{1,0}$. □

Instead of circumventing the lack of derivatives of ϕ as in Propositions 3.2.16 and 3.2.21, one can take another path by generalizing the concept of differentiability.

Definition 3.2.23 ([14])**.** Let V be a normed space, $F : V \to \mathbb{R} \cup \{+\infty\}$ a convex functional and $v_0 \in \operatorname{dom} F$. A bounded linear functional g on V with

$$F(v) - F(v_0) \geq {}_{V'}\langle g, v - v_0\rangle_V \quad \forall v \in V \tag{3.2.38}$$

is called a *subgradient of F at v_0*. The set of all subgradients at v_0 is called the *subdifferential of F at v_0* and is denoted by $\partial F(v_0)$. Furthermore, we set $\operatorname{dom} \partial F := \{v \in \operatorname{dom} F : \partial F(v) \neq \emptyset\}$.

It is obvious that the subdifferential ∂F is a multivalued operator from $\operatorname{dom} \partial F$ to $2^{V'}$ and that $\partial F(v)$ is always a closed convex set (possibly empty). The subdifferential is indeed a generalization of the Gâteaux-derivative, since the following can be proven.

Proposition 3.2.24 ([33])**.** *Let F be a convex function of V into $\mathbb{R} \cup \{+\infty\}$. If F is Gâteaux-differentiable at $v \in V$, then it is subdifferentiable at v and $\partial F(v) = \{F'(v)\}$. Conversely, if F is continuous and finite at the point $v \in V$ and has only one subgradient, then F is Gâteaux-differentiable at v and $\partial F(v) = \{F'(v)\}$.*

It follows immediately from Definition 3.2.23 that ∂F is a monotone operator and that $\operatorname{dom} \partial F \subset \operatorname{dom} F$. Even more is true.

Proposition 3.2.25 ([14])**.** *Let V be a real Banach space and F be a lower semicontinuous proper convex functional on V. Then it holds:*

a) The subdifferential ∂F is a maximal monotone operator, i.e.

$$ {}_{V'}\langle g - g_1, v - v_1\rangle_V \geq 0 \quad \forall g \in \partial F(v)\, \forall v \in \operatorname{dom} \partial F$$

implies $v_1 \in \operatorname{dom} \partial F$ and $g_1 \in \partial F(v_1)$.

b) The domain $\operatorname{dom} \partial F$ is a dense subset of $\operatorname{dom} F$.

Remark 3.2.26. In $V = \mathbb{R}$, the converse of part a) is also true: each maximal monotone function is the subdifferential of a lower semicontinuous proper convex function (see [14, p. 60]).

Now, let the scalar function Φ defined in (3.2.11) be canonically extended by $+\infty$ on $(-\infty, u_c)$ according to Remark 3.2.19 and H be monotonically increasing. By Proposition 3.2.25 b), we have $\operatorname{dom} \partial\Phi \cap (-\infty, u_c) = \emptyset$, which agrees with the fact that for $v_0 = z_0 < u_c$, the left-hand side of (3.2.38) is not defined for all $v = z < u_c$. For $z_0 > u_c$, we confirm the convexity of Φ by Lemma 3.2.4 and obtain, by using this lemma in combination with Lemma 3.2.3,

$$\begin{aligned}\frac{\Phi(z_0) - \Phi(z_1)}{z_0 - z_1} = \frac{\int_{z_1}^{z_0} H(z)\,\mathrm{d}z}{z_0 - z_1} &\leq \lim_{y \uparrow z_0} H(y) \\ &\leq \lim_{y \downarrow z_0} H(y) \leq \frac{\int_{z_0}^{z_2} H(z)\,\mathrm{d}z}{z_2 - z_0} = \frac{\Phi(z_2) - \Phi(z_0)}{z_2 - z_0}\end{aligned} \tag{3.2.39}$$

for $z_1 < z_0 < z_2$. Therefore, we have

$$\Phi(z) - \Phi(z_0) \geq g_{z_0}(z - z_0) \quad \forall z \in \mathbb{R} \tag{3.2.40}$$

for all

$$g_{z_0} \in [\lim_{y \uparrow z_0} H(y), \lim_{y \downarrow z_0} H(y)] =: I_{z_0}. \tag{3.2.41}$$

If H is continuous and thus Φ differentiable, this results in $\partial\Phi(z_0) = I_{z_0} = \{H(z_0)\}$. Finally for $z_0 = u_c$, the same consideration as in (3.2.39) with $\Phi(z_1) = +\infty$ for $z_1 < z_0 = u_c$ provides (3.2.40) for all

$$g_{z_0} \in (-\infty, \lim_{y \downarrow u_c} H(y)] =: I_{u_c}. \tag{3.2.42}$$

Altogether, the multivalued function $\tilde{H} : [u_c, \infty) \to 2^{\mathbb{R}}$ defined by

$$\tilde{H}(z_0) := I_{z_0} \quad \forall z_0 \in [u_c, \infty) \tag{3.2.43}$$

with I_{z_0} from (3.2.41) and (3.2.42) is the subdifferential of Φ, i.e. $\partial\Phi = \tilde{H}$, on $\operatorname{dom} \partial\Phi = [u_c, \infty)$, and it is maximal monotone.

Note that the only but decisive assumption for the preceding considerations is the fact that H is monotonically increasing. An example for $\tilde{H}$ are the hydrological limit cases (1.1.19) and (1.1.21).

We now turn to the subdifferential of ϕ. We assume for a moment that ϕ is defined on $L^2(D) \otimes L^2(\Omega)$ and will show that

$$g \in \partial\phi(v_0) \subset L^2(D) \otimes L^2(\Omega) \quad \Leftrightarrow \quad g(x, \xi(\omega)) \in \partial\Phi\left(v_0(x, \xi(\omega))\right) \subset \mathbb{R}. \tag{3.2.44}$$

For $g \in \partial\phi(v_0)$, it holds

$$\begin{aligned} &\mathbb{E}\left[\int_D \Big(\Phi\left(v(x, \xi(\omega))\right) - \Phi\left(v_0(x, \xi(\omega))\right)\Big) \,\mathrm{d}x\right] = \phi(v) - \phi(v_0) \\ &\geq \mathbb{E}\left[\int_D g(x, \xi(\omega)) \left(v(x, \xi(\omega)) - v_0(x, \xi(\omega))\right) \,\mathrm{d}x\right] \quad \forall v \in L^2(D) \otimes L^2(\Omega). \end{aligned} \tag{3.2.45}$$

For any measurable set $E := E_1 \times E_2 \subset D \times \Omega$ define $\tilde{v} = v$ on E and $\tilde{v} = v_0$ on the complement of E. Taking $v = \tilde{v}$ in (3.2.45) leads to

$$\begin{aligned} \int_{E_2} \int_{E_1} \Big(&\Phi\left(v(x, \xi(\omega))\right) - \Phi\left(v_0(x, \xi(\omega))\right) \\ &- g(x, \xi(\omega)) \left(v(x, \xi(\omega)) - v_0(x, \xi(\omega))\right)\Big) \,\mathrm{d}x \,\mathrm{d}\mathbb{P} \geq 0 \end{aligned} \tag{3.2.46}$$

for all $v \in L^2(D) \otimes L^2(\Omega)$, whence, since E was arbitrary,

$$\begin{aligned} &\Phi\left(v(x, \xi(\omega))\right) - \Phi\left(v_0(x, \xi(\omega))\right) \geq g(x, \xi(\omega)) \left(v(x, \xi(\omega)) - v_0(x, \xi(\omega))\right), \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{a.e. on } D \times \Omega. \end{aligned}$$

Thus $g(x, \xi(\omega)) \in \partial\Phi\left(v_0(x, \xi(\omega))\right)$, a.e. on $D \times \Omega$. Conversely, let $g \in L^2(D) \otimes L^2(\Omega)$ such that

$$\Phi(v) - \Phi\left(v_0(x, \xi(\omega))\right) \geq g(x, \xi(\omega)) \left(v - v_0(x, \xi(\omega))\right), \quad \text{a.e. on } D \times \Omega, \quad \forall v \in \mathbb{R}.$$

From this, it is immediately clear that $g \in \partial\phi(v_0)$, as claimed.

From Proposition 3.2.25 b) and [14, Prop. II.2.8], we can further deduce

$$\overline{\operatorname{dom} \partial\phi} = \overline{\operatorname{dom} \phi} = \{u \in L^2(D) \otimes L^2(\Omega) : u(x, \xi(\omega)) \in \overline{\operatorname{dom} \Phi}, \text{ a.e. on } D \times \Omega\}.$$

Let ϕ be now defined on $H^1(D) \otimes L^2(\Omega)$ with $\Phi(v) \in L^2(D) \otimes L^2(\Omega)$. We expect the subdifferential $\partial\phi(v_0) \subset \left(H^1(D) \otimes L^2(\Omega)\right)'$ to be given on a larger domain than

in the previous case. To make this precise, let us denote the subdifferential for ϕ defined on $L^2(D)\otimes L^2(\Omega)$ by $(\partial\Phi(v_0),\cdot)_{0,0}\subset \big(L^2(D)\otimes L^2(\Omega)\big)'$, which is motivated by (3.2.44). It is now obvious that

$$(\partial\Phi(v_0),\cdot)_{0,0}\subset\partial\phi(v_0), \tag{3.2.47}$$

more exactly, $(\partial\Phi(v_0),\cdot)_{0,0}$ is the subset of all elements in $\partial\phi(v_0)$ which are also functionals on $L^2(D)\otimes L^2(\Omega)$.

We are now ready to state the reformulation of (3.2.33) as a variational inclusion.

Proposition 3.2.27. *The minimization problem (3.2.33) is equivalent to the variational inclusion*

$$\tilde{u}\in H^1_{\Gamma_D}(D)\otimes L^2(\Omega):\ \ 0\ \in\ a(w+\tilde{u},\cdot)-\ell(\cdot)+\partial\phi(w+\tilde{u}) \tag{3.2.48}$$

in $\big(H^1_{\Gamma_D}(D)\otimes L^2(\Omega)\big)'$.

Proof. First, the functional $\mathcal{J}$ is differentiable, and as in (3.2.22) we obtain

$$\partial\mathcal{J}_w(v_0)(v)=\mathcal{J}'(w+v_0)(v)=a(w+v_0,v)-\ell(v)$$

and consequently

$$\partial(\mathcal{J}_w+\phi_w)(v_0)(v)=a(w+v_0,v)-\ell(v)+\partial\phi_w(v_0)(v)$$

for all $v_0\in\operatorname{dom}\partial\phi_w\subset H^1_{\Gamma_D}(D)\otimes L^2(\Omega)$ and $v\in H^1_{\Gamma_D}(D)\otimes L^2(\Omega)$.

If $\tilde{u}\in H^1_{\Gamma_D}(D)\otimes L^2(\Omega)$ solves the minimization problem (3.2.33), then we have

$$(\mathcal{J}_w(v)+\phi_w(v))-(\mathcal{J}_w(\tilde{u})+\phi_w(\tilde{u}))\geq 0\quad\forall v\in H^1_{\Gamma_D}(D)\otimes L^2(\Omega) \tag{3.2.49}$$

and thus by Definition 3.2.23

$$0\ \in\ a(w+\tilde{u},\cdot)-\ell(\cdot)+(\partial\phi_w)(\tilde{u}).$$

Conversely, if $\tilde{u}$ solves the variational inclusion (3.2.48), we obtain (3.2.49) by the same argument. □

Finally, we cite a more general existence theorem than the one obtained in Theorem 3.2.11. We choose homogeneous Dirichlet conditions and apply a result from [60, Prop. 3.2.1] to our case.

Theorem 3.2.28. *If $H:I\to\mathbb{R}$ defined on an interval $I\subset\mathbb{R}$ containing zero is monotonically increasing and $a(\cdot,\cdot)$ is a coercive and continuous bilinear form and $\ell\in\big(H^1_{\Gamma_D}(D)\otimes L^2(\Omega)\big)'$, then the variational inclusion*

$$u\in H^1_{\Gamma_D}(D)\otimes L^2(\Omega):\ \ 0\ \in\ a(u,\cdot)-\ell(\cdot)+(\tilde{H}(u),\cdot)_{0,0} \tag{3.2.50}$$

in $\big(H^1_{\Gamma_D}(D)\otimes L^2(\Omega)\big)'$ *has a solution.*

Note that I can be chosen as $I=[u_c,\infty)$ with $u_c<0$ or $I=\mathbb{R}$ and that $\tilde{H}=\partial\Phi$ is the multifunction corresponding to H. In light of Propositions 3.2.20 and 3.2.27, we can deduce from (3.2.47) immediately the following.

Proposition 3.2.29. *If u is a solution of the variational inclusion (3.2.50), then it is also a solution of the minimization problem (3.2.26).*

Since we have

$$(\partial\Phi(v_0), \cdot)_{0,0} \neq \partial\phi(v_0), \tag{3.2.51}$$

the converse of Proposition 3.2.29 is not true in general. However, both subdifferentials coincide when we discretize them, see Remark 3.3.13. Moreover, the uniqueness of the solution of (3.2.26) provides another interesting result.

Proposition 3.2.30. *If $H : I \to \mathbb{R}$ defined on an interval $I \subset \mathbb{R}$ containing zero is monotonically increasing and bounded, then the solution of (3.2.50) is unique.*

We conclude this section with two remarks.

Remark 3.2.31. The assumption of boundedness of H for the statement that the solution in Theorems 3.2.11 or 3.2.28 is unique can be relaxed and replaced by Hölder continuity of H outside of an interval $[-a, a]$, confer [18, p. 69].

Remark 3.2.32. Finally, we consider the Richards equation in the limit cases introduced in (1.1.18) and (1.1.20). In either case, the function Φ is linear on the interval $[u_c, \infty)$ and its subdifferential reads

$$\partial\Phi(u) = \tilde{H}(u) = \begin{cases} \emptyset & \text{for } u < u_c \\ (-\infty, \theta_M] & \text{for } u = u_c \\ \theta_M & \text{for } u > u_c. \end{cases} \tag{3.2.52}$$

The functional ϕ becomes linear, too, and we end up with a linear constrained problem. Indeed, with the function w from (3.2.30), we define

$$\tilde{\ell}(v) := \ell(v) - a(w, v) - (\theta_M, v)_{0,0}$$

and can rewrite (3.2.34) as

$$\tilde{u} \in \mathcal{K}_{\Gamma_D} : \;\; a(\tilde{u}, v - \tilde{u}) - \tilde{\ell}(v - \tilde{u}) \geq 0 \quad \forall v \in \mathcal{K}_{\Gamma_D}. \tag{3.2.53}$$

We come back to this topic in Subsection 3.3.4.

3.3 Polynomial chaos and finite elements

In this section, we pursue the discretization of the minimization problem (3.2.26). At this point, the different nature of the spaces $H^1(D)$ and $L^2(\Omega)$ has to be taken into account. In the former space we have to deal with derivatives and boundary conditions and—as it is known from the numerics of (deterministic) PDEs—a finite element approach is feasible. On the other hand, in the latter space we seek after an approximation of an L^2 function, where the structure of the space is specified by the density functions which are based upon the random variables ξ_r from the KL expansion (3.1.10) as explained in (3.1.22). A suitable tool for this are the so-called polynomial chaos expansions, which are presented in Subsection 3.3.1. After we set notation for our finite element approach in Subsection 3.3.2, we focus on the approximation of the nonlinear functional ϕ from (3.2.12), see Subsection 3.3.3. Putting all these tools together, we can formulate (3.3.49) as the discretized version of minimization problem (3.2.26). Finally, we show the convergence of the discretized solution in special cases.

3.3.1 (Generalized) Polynomial chaos

In this subsection, we present the basis that will be used for our discretization in the stochastic space $L^2(\Omega)$. After a historical excursion, we define our polynomial basis in one dimension and extend it afterwards to M dimensions. We present three different possible bases and explain their construction and properties. At the end, we introduce a Gaussian quadrature which is closely related to our polynomial basis.

The *polynomial chaos* (PC) approach was first applied in similar context developed by Ghanem and Spanos [47] for various problems in mechanics and is based on fundamental work by Wiener [110] in 1938. It allows high-order representation in combination with fast convergence by expanding functions $v \in L^2(\Omega)$ in truncated series of the form

$$v(\omega) = \sum_{k=0}^{P} v_k \Psi_k(\omega) \tag{3.3.1}$$

with global polynomials Ψ_k and coefficients v_k.

Although the term "polynomial chaos" was introduced by Wiener for a construction in ergodic theory, the main idea was revealed by Cameron and Martin [23]. They showed that the set $\{\mathrm{H}_k\}$ of Hermite polynomials (see Appendix B.1) forms, if normalized, a complete orthonormal set and that each function $v \in L^2(\tilde{\Omega})$ can be approximated in L^2 sense by an Hermite series

$$v_P(\omega) := \sum_{k=0}^{P} \left(\int_{\tilde{\Omega}} v(\zeta) \mathrm{H}_k(\zeta) \, \mathrm{d}\tilde{\mathbb{P}}(\zeta) \right) \mathrm{H}_k(\omega), \tag{3.3.2}$$

i.e.

$$\int_{\tilde{\Omega}} |v(\omega) - v_P(\omega)|^2 \, \mathrm{d}\tilde{\mathbb{P}}(\omega) \to 0 \quad \text{as } P \to \infty. \tag{3.3.3}$$

This is valid for arbitrary one-dimensional $\tilde{\Omega}$ and arbitrary measures $\tilde{\mathbb{P}}(\cdot)$ which possess a density function. Thus, by setting $\Psi_k = \mathrm{H}_k$ and v_k as the expression in the parentheses in (3.3.2), the Hermite series v_P allows a PC representation as in (3.3.1). Together with a scheme to create products of Hermite polynomials to be able to approximate also functions on an M-dimensional domain $\tilde{\Omega}$ (see the construction of the classical PC basis $\{\Psi_k^c\}$ below), this was the utilized PC basis in all the early works in this field, cf. [47], [73] or [82]. Nevertheless, one problem persisted: while an exponential convergence in (3.3.3) is experienced if $\tilde{\mathbb{P}}(\cdot)$ is Gaussian, i.e. if it has a density of the form (B.2), see [80], the convergence is rather poor for other distributions [53]. The remedy was surprisingly already suggested by Cameron and Martin and systematized by Xiu and Karniadakis [114] who called it the *generalized polynomial chaos*: since the convergence result (3.3.3) still holds if one replaces the Hermite polynomials by an arbitrary dense sets of functions, one can choose the polynomial set according to the distribution in $\tilde{\mathbb{P}}(\cdot)$. The optimal polynomial set is known for some important distributions (e.g. the Legendre polynomials $\{\mathrm{Le}_k\}$ for the uniform distribution, cf. Wiener–Askey chaos in Table 3.2 and Appendix B for definitions and properties of the most common polynomial sets) and can be computed in other cases, see [108].

Taking this into consideration, we define our PC basis as follows. First, recall the spaces Ω_r constructed by means of the random variables ξ_r defined preceding (3.1.22). For each $r = 1, \ldots, M$, let $\{\psi_k^r\}_{k=0,1,\ldots}$ be orthogonal polynomials in Ω_r, which are normalized according to

$$\mathbb{E}\left[\psi_k^r \psi_l^r\right] = \int_{\Omega_r} \psi_k^r(\xi_r) \psi_l^r(\xi_r) \mathrm{pdf}_r(\xi_r) \, \mathrm{d}\xi_r = \delta_{kl} \tag{3.3.4}$$

random variable ξ_r	Wiener chaos $\{\Psi(\xi_r)\}$	support
Gaussian	Hermite chaos	$(-\infty, \infty)$
uniform	Legendre chaos	$[a, b]$
beta	Jacobi chaos	$[a, b]$
gamma	Laguerre chaos	$[0, \infty)$

Table 3.2: The correspondence of Wiener–Askey PC and their underlying random variables (following [114]).

and which constitute an orthogonal basis in $L^2(\Omega_r)$. Observe that the orthogonality is computed with regard to the probability induced by the ξ_r from the KL expansion (3.1.10). Choosing some multi-index $\alpha^{(k)} = (\alpha_1^{(k)}, \dots, \alpha_M^{(k)})$, the multidimensional polynomials Ψ_k are defined as products of corresponding one-dimensional polynomials by

$$\Psi_k(\xi_1, \xi_2, \dots, \xi_M) := \prod_{r=1}^{M} \psi^r_{\alpha_r^{(k)}}(\xi_r). \tag{3.3.5}$$

Note that these products reflect the structure of $\Omega^{(M)}$ given by (3.1.23). If $k = 0, 1, \dots$ is a counting of all multi-indices in $\mathbb{N}_0^M$, these polynomials form a basis such that each second order random variable $v \in L^2(\Omega^{(M)})$ can be written as

$$v(\xi) = \sum_{k=0}^{\infty} v_k \Psi_k(\xi).$$

In light of the identity (3.1.22) and Remark 3.1.23, we henceforth rewrite this as

$$v(\xi(\omega)) = \sum_{k=0}^{\infty} v_k \Psi_k(\xi(\omega)) = \sum_{k=0}^{\infty} v_k \Psi_k(\xi) \tag{3.3.6}$$

for $v \in L^2(\Omega)$. Since the orthonormality

$$\mathbb{E}\left[\Psi_k \Psi_l\right] = \prod_{r=1}^{M} \mathbb{E}\left[\psi^r_{\alpha_r^{(k)}} \psi^r_{\alpha_r^{(l)}}\right] = \delta_{kl} \tag{3.3.7}$$

follows from (3.3.4), the coefficient v_k in (3.3.6) is obtained by projection on Ψ_k, i.e.

$$v_k = (v, \Psi_k)_{L^2(\Omega)} = \mathbb{E}\left[v \Psi_k\right].$$

For our discretization, we truncate the expansion (3.3.6) having a total of $P + 1$ basis polynomials and denote the solution space by

$$\mathcal{Z}^P := \operatorname{span}\{\Psi_k : k = 0, \dots, P\} \subset L^2(\Omega). \tag{3.3.8}$$

The aforementioned projection on our basis polynomials now allows the definition of a projection operator $\mathcal{P}^P : L^2(\Omega) \to \mathcal{Z}^P$ defined by

$$\mathcal{P}^P v(\xi(\omega)) := \sum_{k=0}^{P} (v, \Psi_k)_{L^2(\Omega)} \Psi_k(\xi(\omega)) = \sum_{k=0}^{P} v_k \Psi_k(\xi(\omega)), \tag{3.3.9}$$

where the generalization for $v \in V \otimes L^2(\Omega)$ is immediate by taking the operator tensor product with the identity operator on V.

According to the one-dimensional polynomials ψ_k^r and the multi-indices α, different schemes are possible. We will discuss three different bases in the following. The

mostly used basis (e.g. in [47, 63, 73]) consists of one-dimensional polynomials ψ_k^r of degree k and multi-indices α with $|\alpha| := \sum_r \alpha_r \leq P_0$ such that we have a total of

$$P + 1 = \frac{(M + P_0)!}{M!P_0!} \tag{3.3.10}$$

multidimensional polynomials. We denote this classical polynomial set as $\{\Psi_k^c\}$ and adopt its traditional ordering (cf. [47]), where $\alpha^{(k)} < \alpha^{(l)}$ if $|\alpha^{(k)}| < |\alpha^{(l)}|$ or if, in the case $|\alpha^{(k)}| = |\alpha^{(l)}|$, the first different index in $\alpha^{(k)}$ is smaller than its counterpart in $\alpha^{(l)}$. In particular, it holds

$$\Psi_0 \equiv 1, \tag{3.3.11}$$

which yields that

$$\mathbb{E}\left[\Psi_0\right] = \int_\Omega \mathbf{1}\,\mathrm{d}\mathbb{P} = |\Omega| = 1 \tag{3.3.12}$$

and

$$\mathbb{E}\left[\Psi_k\right] = \mathbb{E}\left[\Psi_k \Psi_0\right] = 0 \quad \forall k > 0 \tag{3.3.13}$$

due to (3.3.7).

A larger but more flexible scheme (e.g. in [11]) is when using the same polynomials ψ_k^r but with multi-indices α with $\alpha_r \leq P_r$ for all $r = 1, \ldots, M$. We denote it as the tensor product basis $\{\Psi_k^t\}$. It it easy to see that the identities (3.3.11)–(3.3.13) hold as above. We employ the same ordering as for $\{\Psi_k^c\}$ and remark that the space $\mathcal{Z}^P$ can be rewritten as

$$\mathcal{Z}^P = \bigotimes_{r=1}^{M} \mathcal{Z}_r^{P_r} \tag{3.3.14}$$

with one-dimensional global polynomial spaces

$$\mathcal{Z}_r^{P_r} := \left\{v = v(y) \in L^2(\Omega_r) \text{ with } v \in \operatorname{span}\{y^k : k = 0, \ldots, P_r\}\right\} = \mathrm{Pol}^{P_r}(\Omega_r).$$

Hence, the cardinality of $\{\Psi_k^t\}$ is

$$P + 1 = \prod_{r=1}^{M} (P_r + 1). \tag{3.3.15}$$

A variant derived from $\{\Psi_k^t\}$ are biorthogonal polynomials [11]. In one dimension, we denote them by $\{\hat{\psi}_k^r\}$ and claim that they satisfy

$$\mathbb{E}\left[\hat{\psi}_k^r \hat{\psi}_l^r\right] = \int_{\Omega_r} \hat{\psi}_k^r(\xi_r)\hat{\psi}_l^r(\xi_r)\mathrm{pdf}_r(\xi_r)\,\mathrm{d}\xi_r = \delta_{kl}, \tag{3.3.16}$$

$$\mathbb{E}\left[\xi_r \hat{\psi}_k^r \hat{\psi}_l^r\right] = \int_{\Omega_r} \xi_r\, \hat{\psi}_k^r(\xi_r)\hat{\psi}_l^r(\xi_r)\mathrm{pdf}_r(\xi_r)\,\mathrm{d}\xi_r = c_k^r \delta_{kl} \tag{3.3.17}$$

for all $k, l = 0, \ldots, P_r$, where the numbers c_k^r are nonzero constants. The clue of finding these polynomials lies in the representation as linear combinations of orthogonal polynomials

$$\hat{\psi}_k^r = \sum_{l=0}^{P_r} s_{kl}^r \psi_l^r. \tag{3.3.18}$$

Defining the coefficient matrices $\mathsf{S} \in \mathbb{R}^{(P_r+1)\times(P_r+1)}$ and $\mathsf{C} \in \mathbb{R}^{(P_r+1)\times(P_r+1)}$ with entries $\mathsf{S}_{kl} = s_{kl}^r$ and $\mathsf{C}_{kl} = c_k^r \delta_{kl}$ and the mass matrices $\mathsf{M}, \bar{\mathsf{M}} \in \mathbb{R}^{(P_r+1)\times(P_r+1)}$ with entries

$$\mathsf{M}_{kl} = \int_{\Omega_r} \xi_r\, \psi_k^r(\xi_r)\psi_l^r(\xi_r)\mathrm{pdf}_r(\xi_r)\,\mathrm{d}\xi_r,$$

$$\bar{\mathsf{M}}_{kl} = \int_{\Omega_r} \psi_k^r(\xi_r)\psi_l^r(\xi_r)\mathrm{pdf}_r(\xi_r)\,\mathrm{d}\xi_r,$$

P_0	$M=1$		$M=2$		$M=3$		$M=4$	
	$\lvert\{\Psi_k^c\}\rvert$	$\lvert\{\Psi_k^t\}\rvert$	$\lvert\{\Psi_k^c\}\rvert$	$\lvert\{\Psi_k^t\}\rvert$	$\lvert\{\Psi_k^c\}\rvert$	$\lvert\{\Psi_k^t\}\rvert$	$\lvert\{\Psi_k^c\}\rvert$	$\lvert\{\Psi_k^t\}\rvert$
0	1	1	1	1	1	1	1	1
1	2	2	3	4	4	8	5	16
2	3	3	6	9	10	27	15	81
3	4	4	10	16	20	64	35	256
4	5	5	15	25	35	125	70	625
5	6	6	21	36	56	216	126	1296

Table 3.3: The cardinality of the classical PC basis (3.3.10) in comparison with tensor product PC basis (3.3.15) with $P_r = P_0$ for all $r = 1, \dots, M$.

respectively, and denoting by I the identity matrix, we can rewrite (3.3.16)–(3.3.17) by inserting (3.3.18) as

$$\mathsf{S}^T \bar{\mathsf{M}} \mathsf{S} = \mathrm{I}, \quad \mathsf{S}^T \mathsf{M} \mathsf{S} = \mathsf{C},$$

which corresponds in view of $\bar{\mathsf{M}} = \mathrm{I}$ to an eigenvalue problem

$$\mathsf{M}\mathsf{S} = \mathsf{S}\mathsf{C} \tag{3.3.19}$$

for each dimension $r = 1, \dots, M$. Finally, we construct the multidimensional biorthogonal polynomials according to (3.3.5) and denote them by $\{\Psi_k^b\}$. The cardinality is given by (3.3.15), too. Remark that they no longer fulfill the identities (3.3.11)–(3.3.13), but that they possess the advantage of diagonalizing the stiffness matrix for linear problems, cf. Subsection 4.1.1.

For later reference, we note the convergence of the PC discretization. Henceforth, $P \to \infty$ is short hand notation for $P_0 \to \infty$ or $\min_r P_r \to \infty$, respectively.

Theorem 3.3.1. *Let $v \in L^2(\Omega)$ and $v^P := \mathcal{P}^P v$ its projection onto $\mathcal{Z}^P$ as defined in (3.3.9). Then we have*

$$\|v - v^P\|_{L^2(\Omega)} \to 0 \quad \text{as } P \to \infty.$$

The proof is just the multidimensional generalization of the argument in [23] using the tensor product PC basis $\{\Psi_k^t\}$ or, alternatively, a well-known result for orthonormal bases in Hilbert spaces, see e.g. [109].

Remark 3.3.2. Observe that each convergence result as $P \to \infty$ for one of the three presented polynomial sets also holds for the other ones, since we have by construction

$$\operatorname{span}\{\Psi_k^t\} = \operatorname{span}\{\Psi_k^b\} \subset \operatorname{span}\{\Psi_k^c\},$$

if $P_0 \geq \sum_r P_r$, and

$$\operatorname{span}\{\Psi_k^c\} \subset \operatorname{span}\{\Psi_k^t\} = \operatorname{span}\{\Psi_k^b\},$$

if $\min_r P_r \geq P_0$. Nevertheless, the size of the bases and hence the numerical work is distinctly different, see Table 3.3.

Remark 3.3.3. We can now explain how Dirichlet boundary conditions

$$\operatorname{tr}_{\Gamma_D} u = u_D$$

from (3.2.3) can be modeled. To avoid technical subtleties, let the function u be continuous. Following an idea from [73], we employ on both sides the projection $\mathcal{P}^P$ onto the orthonormal basis. By equating the coefficients, one obtains

$$u_k(x) = \mathbb{E}\left[u_D(x)\Psi_k\right] \tag{3.3.20}$$

for $x \in D$. For a deterministic function $u_D = u_D(x)$, the right-hand side of (3.3.20) can be simplified as $\mathbb{E}[u_D(x)\Psi_k] = \mathbb{E}[\Psi_k]\, u_D(x)$ such that (3.3.12) and (3.3.13) provide

$$u_k(x) = \begin{cases} u_D(x) & \text{for } k = 0 \\ 0 & \text{else} \end{cases}$$

for $\{\Psi\} = \{\Psi_k^c\}$ and $\{\Psi\} = \{\Psi_k^t\}$, i.e. one can employ the usual spatial boundary discretization for u_D and allocate it to the zeroth PC mode, while all other PC coefficients vanish. For a discretization of Neumann boundary conditions, we refer to Remark 4.1.2.

Remark 3.3.4. An initial condition $u = u^0(x)$ can be treated in the same way as Dirichlet conditions in Remark 3.3.3. In particular, we have

$$u_k(x) = \begin{cases} u^0(x) & \text{for } k = 0 \\ 0 & \text{else} \end{cases}$$

if $u^0 = u^0(x)$ is deterministic and if $\{\Psi\} = \{\Psi_k^c\}$ or $\{\Psi\} = \{\Psi_k^t\}$.

In order to compute the occurring expectation values, quadrature formulas are necessary. An obvious choice in view of the involved polynomials is Gaussian quadrature with the density $\mathrm{pdf}(\cdot)$ as weighting function. In the one-dimensional space Ω_r with PC polynomials of maximal degree P_r (take $P_r = P_0$ for $\{\Psi\} = \{\Psi_k^c\}$) we apply a Gaussian quadrature with $P_r + 1$ quadrature points $\pi_1^r, \dots, \pi_{P_r+1}^r$ which are the zeros of the orthogonal polynomial of degree $P_r + 1$. The quadrature weights $\eta_{\pi_i^r}$ are given by

$$\eta_{\pi_i^r} = \int_{\Omega_r} \mathcal{L}_{r,\pi_i^r} \, \mathrm{d}\mathbb{P}_{\xi_r},$$

where $\mathcal{L}_{r,\pi_i^r}$ are the Lagrange polynomials with $\mathcal{L}_{r,\pi_i^r}(\pi_j^r) = \delta_{ij}$ as a basis of $\mathrm{Pol}^{P_r}(\Omega_r)$. Properties of Gaussian quadrature like the fact that polynomials of degree $2P_r + 1$ are computed exactly are summarized in Appendix C.

For the M-dimensional quadrature, we just take the products

$$\pi_i = (\pi_{i_1}^1, \dots, \pi_{i_r}^r), \quad \eta_{\pi_i} = \prod_{r=1}^{M} \eta_{\pi_{i_r}^r}, \tag{3.3.21}$$

and

$$\mathcal{L}_{\pi_i}(\xi(\omega)) = \prod_{r=1}^{M} \mathcal{L}_{r,\pi_{i_r}^r}(\xi_r(\omega)) \tag{3.3.22}$$

with the global index

$$i = i_1 + (P_1 + 1)(i_2 - 1) + (P_1 + 1)(P_2 + 1)(i_3 - 1) + \dots$$

We denote by $\mathcal{Q}^P$ the set of all these quadrature points π_i (omitting the index in the sequel), denote its cardinality by Π, and see that

$$\Pi := |\mathcal{Q}^P| = \prod_{r=1}^{M} (P_r + 1). \tag{3.3.23}$$

With this choice of $\mathcal{Q}^P$, the quadrature order is high enough to compute $\mathbb{E}[\Psi_k \Psi_l]$ and $\mathbb{E}[\Psi_k \Psi_l \xi_r]$ or, more general, $\mathbb{E}[uv]$ for $u, v \in \mathcal{Z}^P$ exactly. This is not the case for $\mathbb{E}[H(u)v]$ for nonlinear H but for the projected version $\mathbb{E}[\mathcal{P}^P(H(u))v]$. For that

reason, we will henceforth call this $\mathcal{Q}^P$ the *quadrature point set corresponding to* $\mathcal{Z}^P$.

With this notation, we can indicate the solution of (3.3.16)–(3.3.17). Indeed, the biorthogonal polynomials are, up to multiplicative factors, just Lagrange polynomials in Gaussian quadrature points, i.e.

$$\hat{\psi}_k^r = \frac{1}{\sqrt{\eta_{\pi_k^r}}} \mathcal{L}_{r,\pi_k^r} \tag{3.3.24}$$

and $c_k^r = \pi_k^r$, see [13, Lemma 2.1]. This will play an important role later on.

We close this subsection by defining for further use the Lagrange interpolant operator $\mathscr{I}^P : C^0(\Omega^{(M)}) \to \mathcal{Z}^P$ by

$$\mathscr{I}^P v(\xi) := \sum_{\pi \in \mathcal{Q}^P} v(\pi) \mathcal{L}_\pi(\xi), \tag{3.3.25}$$

generated by the quadrature point set $\mathcal{Q}^P$ corresponding to $\mathcal{Z}^P$. Taking the quadrature formula for the function which is constant one, one can see that

$$\sum_{\pi \in \mathcal{Q}^P} \eta_\pi = 1, \tag{3.3.26}$$

since

$$\begin{aligned} \sum_{\pi \in \mathcal{Q}^P} \eta_\pi &= \sum_{\pi \in \mathcal{Q}^P} \int_\Omega \mathcal{L}_\pi(\xi(\omega)) \, \mathrm{d}\mathbb{P}(\omega) = \int_\Omega \sum_{\pi \in \mathcal{Q}^P} \mathcal{L}_\pi(\xi(\omega)) \, \mathrm{d}\mathbb{P}(\omega) \\ &= \int_\Omega (\mathscr{I}^P \mathbf{1})(\xi(\omega)) \, \mathrm{d}\mathbb{P}(\omega) = \int_\Omega \mathbf{1} \, \mathrm{d}\mathbb{P} = |\Omega| = 1. \end{aligned}$$

Again, the generalization of the operator $\mathscr{I}^P$ for $v \in V \otimes L^2(\Omega)$ is immediate by taking the operator tensor product with the identity operator on V.

3.3.2 Finite elements in space

In this subsection, we present the space discretization by finite elements. The notation and the main ideas are the same as in Berninger [18], who follows Kornhuber [68] and Glowinski [48]. As in these references, we consider the case of a bounded, polygonal domain $D \subset \mathbb{R}^2$. However, the results presented herein also apply for polyhedral domains in higher and lower dimensions.

Let $\mathcal{T}_\jmath$ be a given partition of D into triangles $\mathrm{t} \in \mathcal{T}_\jmath$ with minimal diameter of order $\mathcal{O}(2^{-\jmath})$. The set of all $N_\jmath$ vertices of the triangles in $\mathcal{T}_\jmath$ is denoted by $\mathcal{N}_\jmath$. We assume that each triangulation $\mathcal{T}_\jmath$ is regular in the sense that the intersection of two triangles in $\mathcal{T}_\jmath$ is either empty or consists of a common edge or a common vertex.

For the convergence proofs in the following subsections, we deal with a sequence of these triangulations and assume that they possess a decreasing mesh size

$$h_\jmath := \max_{\mathrm{t} \in \mathcal{T}_\jmath} \operatorname{diam} \mathrm{t} \to 0 \quad \text{as } \jmath \to \infty. \tag{3.3.27}$$

In addition, we assume that the sequence of triangulations

$$(\mathcal{T}_\jmath)_{\jmath \geq 0} \text{ is shape regular} \tag{3.3.28}$$

in the sense that the minimal interior angle of all triangles contained in $\bigcup_{j\geq 0}\mathcal{T}_j$ is bounded from below by a positive constant.

The character of Dirichlet and Neumann boundaries should be reflected properly by the triangulation. To this end, we assume that Γ_D is closed and that each intersection point in $\Gamma_D \cap \overline{\Gamma_N}$ is contained in $\mathcal{N}_j$. The vertices on the Dirichlet boundary are denoted by $\mathcal{N}_j^D := \mathcal{N}_j \cap \Gamma_D$.

The finite element space $\mathcal{S}_j \subset H^1(D)$ is the subspace of all continuous functions in $H^1(D)$ which are linear on each triangle $\mathrm{t} \in \mathcal{T}_j$. The space $\mathcal{S}_j$ is spanned by the nodal basis

$$\Lambda_j := \{s_p^{(j)} : p \in \mathcal{N}_j\},$$

the elements $s_p^{(j)}$ of which are determined by $s_p^{(j)}(q) = \delta_{pq}$ for all $p, q \in \mathcal{N}_j$. The analogous construction in the space $H^1_{\Gamma_D}(D)$ gives rise to the finite element space $\mathcal{S}_j^D \subset H^1_{\Gamma_D}(D)$ and its nodal basis given by

$$\Lambda_j^D := \{s_p^{(j)} : p \in \mathcal{N}_j \backslash \mathcal{N}_j^D\}.$$

Together with the global polynomial space $\mathcal{Z}^P$ and the polynomial chaos functions Ψ_k introduced in (3.3.8) and (3.3.5), respectively, we define the tensor space $\mathcal{S}_j \otimes \mathcal{Z}^P \subset H^1(D) \otimes L^2(\Omega)$ by

$$\mathcal{S}_j \otimes \mathcal{Z}^P := \left\{\varphi \in H^1(D) \otimes L^2(\Omega) : \varphi \in \operatorname{span}\{s_p^{(j)}(x)\Psi_k(\xi) : p \in \mathcal{N}_j,\, k = 0, \ldots, P\}\right\}. \tag{3.3.29}$$

Recall that $\mathcal{Q}^P$ denotes the Gaussian quadrature point set corresponding to $\mathcal{Z}^P$. Then, we proceed by defining the finite dimensional analogue of the convex set $\mathcal{K}$ defined in (3.2.3) by

$$\mathcal{K}_j^P := \{v \in \mathcal{S}_j \otimes \mathcal{Z}^P : v(p,\pi) \geq u_c \,\forall p \in \mathcal{N}_j \,\forall \pi \in \mathcal{Q}^P \,\wedge\, v(p,\pi) = u_D(p,\pi) \quad \forall p \in \mathcal{N}_j^D \,\forall \pi \in \mathcal{Q}^P\}. \tag{3.3.30}$$

The set $\mathcal{K}_j^P \subset \mathcal{S}_j \otimes \mathcal{Z}^P$ is obviously convex, nonempty and closed. Observe that we have to assume that the Dirichlet boundary function u_D is continuous in each node $(p,\pi) \in \mathcal{N}_j^D \times \mathcal{Q}^P$ such that writing $u_D(p,\pi)$ makes sense. Even if this function is purely deterministic, $u_D = u_D(x)$, it has impact on the coefficients v_{ik} in the representation

$$v(x,\xi(\omega)) = \sum_{i=1}^{N_j}\sum_{k=0}^{P} v_{ik} s_{p_i}^{(j)}(x)\Psi_k(\xi(\omega)) \in \mathcal{S}_j \otimes \mathcal{Z}^P \tag{3.3.31}$$

for $\{i : p_i \in \mathcal{N}_j^D\}$, as it is described in Remark 3.3.3.

Remark 3.3.5. In general, we have $\mathcal{K}_j^P \not\subset \mathcal{K}$ since the Dirichlet boundary values in $\mathcal{K}_j^P$ differ from those in $\mathcal{K}$. Even worse, the condition $v(x,\xi(\omega)) \geq u_c$ needs not hold for $(x,\xi) \notin \mathcal{N}_j \times \mathcal{Q}^P$ because v is a polynomial in ξ. However, neither in the formulation of the problem nor in the post-processing of the solution a computation in points $\xi \notin \mathcal{Q}^P$ is needed, since the calculation of the integrals and moments is done by using these quadrature points solely. If the solution method is constructed in a way that it does not require the evaluation in such points either (more precisely: the evaluation of $H(u(x,\cdot))$ and $\Phi(u(x,\cdot))$ or the inverse Kirchhoff transformation on $u(x,\cdot)$ in such points), this fact does not pose a drawback. On the other hand, we emphasize that in spatial direction the piecewise linearity provides this obstacle condition for all $(x,\pi) \in D \times \mathcal{Q}^P$ as in the deterministic case (see [18, Remark 2.5.1]).

3.3.3 Approximation of the nonlinear functional

In this subsection, a discrete version of the functional ϕ from (3.2.12) and proofs of important properties are presented. This will allow us to formulate the fully discretized version of the convex minimization problem (3.2.26).

We seek an approximation of the nonlinear functional ϕ for functions in the subset $\mathcal{S}_\jmath \otimes \mathcal{Z}^P$. The integrals are approximated by quadrature formulas, but in a different way: the treatment of the spatial integral arises directly from the interpolation of the integrand $\Phi(v)$ in $\mathcal{S}_\jmath$, which provides positive weights of the form

$$h_p := \int_D s_p^{(\jmath)}(x)\,\mathrm{d}x, \tag{3.3.32}$$

whereas a Gaussian quadrature is taken for the expectation value with weights η_π corresponding to quadrature points $\pi \in \mathcal{Q}^P$ as introduced in Subsection 3.3.1. The discrete functional $\phi_\jmath^P : \mathcal{S}_\jmath \otimes \mathcal{Z}^P \to \mathbb{R} \cup \{+\infty\}$ then reads

$$\phi_\jmath^P(v) := \sum_{\pi \in \mathcal{Q}^P} \sum_{p \in \mathcal{N}_\jmath} \Phi(v(p,\pi)) h_p \eta_\pi \quad \forall v \in \mathcal{S}_\jmath \otimes \mathcal{Z}^P. \tag{3.3.33}$$

Observe that the evaluation $v(p,\pi)$ has according to (3.3.31) the form

$$v(p_{i_0},\pi) = \sum_{i=1}^{N_\jmath} \sum_{k=0}^{P} v_{ik} s_{p_i}^{(\jmath)}(p_{i_0}) \Psi_k(\pi) = \sum_{k=0}^{P} v_{i_0 k} \Psi_k(\pi). \tag{3.3.34}$$

This means that $\phi_\jmath^P$ is decoupled in spatial direction with regard to the nodal basis, whereas the nonlinear function Φ is still depending on all PC basis functions in each quadrature point due to their global nature. Each solution method has to take this fact into account.

The following lemma shows that $\phi_\jmath^P$ features all important properties which ϕ have.

Lemma 3.3.6. *Let H be monotonically increasing and bounded and let (3.3.23) hold. Then the functional $\phi_\jmath^P$ is convex, proper and lower semicontinuous on its domain*

$$\operatorname{dom} \phi_\jmath^P = \{v \in \mathcal{S}_\jmath \otimes \mathcal{Z}^P : v(p,\pi) \geq u_c \,\forall p \in \mathcal{N}_\jmath \,\forall \pi \in \mathcal{Q}^P\}.$$

Furthermore, it is Lipschitz continuous and satisfies

$$|\phi_\jmath^P(v)| \leq C\|v\|_{1,0} \quad \forall v \in \operatorname{dom} \phi_\jmath^P, \tag{3.3.35}$$

where the Lipschitz constant as well as the constant $C > 0$ are independent of $\jmath$ and P.

Proof. The convexity of $\phi_\jmath^P$ follows from the convexity of Φ and the fact that the weights h_p and η_π are positive. It is clearly proper and lower semicontinuous if we can prove the Lipschitz continuity.

Denoting by L the Lipschitz constant of Φ, we take u, $v \in \operatorname{dom} \phi_\jmath^P$ and have

$$\begin{aligned} |\phi_\jmath^P(u) - \phi_\jmath^P(v)| &= \left| \sum_{\pi \in \mathcal{Q}^P} \sum_{p \in \mathcal{N}_\jmath} \left(\Phi(u(p,\pi)) - \Phi(v(p,\pi))\right) h_p \eta_\pi \right| \\ &\leq L \sum_{\pi \in \mathcal{Q}^P} \sum_{p \in \mathcal{N}_\jmath} \eta_\pi |u(p,\pi) - v(p,\pi)| \int_D s_p^{(\jmath)}(x)\,\mathrm{d}x. \end{aligned}$$

Since $u(\cdot,\pi)$ and $v(\cdot,\pi)$ are linear functions on each triangle $\mathrm{t} \in \mathcal{T}_\jmath$ for all π, we can further deduce

$$\begin{aligned}|\phi_\jmath^P(u) - \phi_\jmath^P(v)| &\leq L \sum_{\pi\in\mathcal{Q}^P} \sum_{p\in\mathcal{N}_\jmath} \eta_\pi \int_{\operatorname{supp}(s_p^{(\jmath)})} |u(x,\pi) - v(x,\pi)| \,\mathrm{d}x \\ &\leq 3L \int_D \sum_{\pi\in\mathcal{Q}^P} \eta_\pi |u(x,\pi) - v(x,\pi)| \,\mathrm{d}x,\end{aligned}$$

where the last inequality is due to the fact that each triangle is contained in the support of three nodal basis functions. We apply the Cauchy–Schwarz inequality to the sum over π with (3.3.26) and exploit (3.3.23) and the high order of the Gaussian quadrature, which allows us to compute $\mathbb{E}[v^2]$ exactly, to obtain

$$\begin{aligned}|\phi_\jmath^P(u) - \phi_\jmath^P(v)| &\leq 3L \int_D \left(\sum_{\pi\in\mathcal{Q}^P} \eta_\pi \right)^{1/2} \left(\sum_{\pi\in\mathcal{Q}^P} \eta_\pi |u(x,\pi) - v(x,\pi)|^2 \right)^{1/2} \mathrm{d}x \\ &\leq 3L \int_D \left(\int_\Omega \left(u(x,\xi) - v(x,\xi)\right)^2 \,\mathrm{d}\mathbb{P} \right)^{1/2} \mathrm{d}x.\end{aligned}$$

Finally, by applying the Cauchy–Schwarz inequality in $L^2(D)$ we arrive at

$$\begin{aligned}|\phi_\jmath^P(u) - \phi_\jmath^P(v)| &\leq 3L\|1\|_{L^2(D)} \left(\int_D \int_\Omega \left(u(x,\xi) - v(x,\xi)\right)^2 \,\mathrm{d}\mathbb{P}\,\mathrm{d}x \right)^{1/2} \\ &\leq C\|u - v\|_{1,0},\end{aligned}$$

which shows the Lipschitz continuity of $\phi_\jmath^P$ and which yields (3.3.35), since $\phi_\jmath^P(0) = 0$. □

For the proof of the consistency of $\phi_\jmath^P$ in Theorem 3.3.9, some preliminaries are in order. First, we recall the definition of the projection $\mathcal{P}^P$ in (3.3.9) and introduce by $\mathcal{I}_{\mathcal{S}_\jmath}$ the piecewise linear interpolation operator $\mathcal{I}_{\mathcal{S}_\jmath} : C^0(\overline{D}) \to \mathcal{S}_\jmath$ defined by

$$\mathcal{I}_{\mathcal{S}_\jmath} v(p) := v(p) \quad \text{for all } p \in \mathcal{N}_\jmath. \tag{3.3.36}$$

The tensor product structure allows a generalization $\mathcal{I}_{\mathcal{S}_\jmath} : C^0(\overline{D}) \otimes V \to \mathcal{S}_\jmath \otimes V$ for $V \subset L^2(\Omega)$ immediately by taking the operator tensor product with the identity operator on V.

Next, we assume that the Dirichlet boundary function u_D is the trace of a function $w \in H^1(D) \otimes L^2(\Omega)$ which is uniformly continuous on Γ_D for almost all $\omega \in \Omega$, i.e.

$$u_D = \operatorname{tr}_{\Gamma_D} w \quad \text{for a } w \in \left(H^1(D) \cap C^0(\overline{D})\right) \otimes L^2(\Omega), \tag{3.3.37}$$

and that the interpolated and projected function $w_\jmath^P = \mathcal{I}_{\mathcal{S}_\jmath} \otimes \mathcal{P}^P w \in \mathcal{S}_\jmath \otimes \mathcal{Z}^P$ satisfies

$$\|w_\jmath^P - w\|_{1,0} \to 0 \quad \text{as } \jmath \to \infty,\ P \to \infty. \tag{3.3.38}$$

Condition (3.3.37) is fulfilled if, for example, $u_D(\cdot,\omega)$ is continuous a.e. on the closed boundary Γ_D, which can be seen by the Tietze extension theorem [109, Corollary B.1.6]. For condition (3.3.38), we split up $\|w_\jmath^P - w\|_{1,0} \leq \|w_\jmath^P - w^P\|_{1,0} + \|w^P - w\|_{1,0}$ for $w^P = \mathcal{P}^P w$. Now, a criterion for condition (3.3.38) can be found in [25, Theorem 16.2], which states, provided that (3.3.27) and (3.3.28) hold, requirements on the regularity of w. In our case $d = 2$, it is fulfilled if $w \in H^2(D) \otimes L^2(\Omega)$ according to the Sobolev embedding theorem (see, e.g., [22, Theorem 2.19]) and Theorem 3.3.1.

For the following convergence results, we will exploit the embedding of $\Omega^{(M)}$ into $\mathbb{R}^M$ and thus switch notation. These results can also be stated if $\Omega^{(M)}$ is unbounded. We do not follow the path of regarding truncations of $\Omega^{(M)}$ as in Remark 3.1.26, but, following an idea in [13], impose an assumption on the probability density function $\mathrm{pdf}(\cdot)$ introduced in (3.1.24) that it satisfies

$$\mathrm{pdf}_r(y_r) \leq C_{\mathrm{pdf}_r} e^{-|\delta_r y_r|^2} \tag{3.3.39}$$

for some $C_{\mathrm{pdf}_r} > 0$ and numbers δ_r which are strictly positive for unbounded Ω_r and equal to zero otherwise such that

$$\mathrm{pdf}(y) \leq C_{\mathrm{pdf}}\, e^{-\sum_{r=1}^M |\delta_r y_r|^2}.$$

Observe that we can cover all important distributions including the normal distribution by this assumption. Moreover, we introduce a positive weight σ assembled as the product of one-dimensional weights $\sigma_r(y_r) = C_{\sigma_r} e^{-|\delta_r y_r|^2/4}$ for some $C_{\sigma_r} > 0$ such that $\sigma(y) = C_\sigma e^{-\sum_{r=1}^M |\delta_r y_r|^2/4}$. Define the functional spaces

$$C^0_\sigma(\Omega^{(M)}) := \{v : v \in C^0(\Omega^{(M)}) \text{ and } \|v\|_{C^0_\sigma(\Omega^{(M)})} := \max_{y\in\Omega^{(M)}} |\sigma(y)v(y)| < \infty\}$$

and

$$C^0_{\sigma,0}(\Omega^{(M)}) := \{v : v \in C^0_\sigma(\Omega^{(M)}) \text{ and } \lim |\sigma(y)v(y)| \to 0 \text{ as } |y| \to \infty\}.$$

With this choice, the embedding $C^0_\sigma(\Omega^{(M)}) \subset L^2(\Omega^{(M)})$ is continuous. We investigate the approximation error of the interpolation $\mathscr{I}^P$ defined in (3.3.25) in the one-dimensional and in the M-dimensional case.

Lemma 3.3.7. *Let $v \in C^0_\sigma(\Omega^{(M)})$ be a continuous function for $M = 1$. Then the interpolation error satisfies*

$$\|v - \mathscr{I}^P v\|_{L^2(\Omega^{(M)})} \leq C \inf_{\bar v \in \mathrm{Pol}^P(\Omega^{(M)})} \|v - \bar v\|_{C^0_\sigma(\Omega^{(M)})} \tag{3.3.40}$$

with a constant $C > 0$ independent of P.

Proof. First, we see that the orthogonal property for Lagrange polynomials (3.3.22) in the Gaussian quadrature points

$$\int_\Omega \mathcal{L}_\pi(\xi(\omega))\mathcal{L}_\varsigma(\xi(\omega))\,\mathrm{d}\mathbb{P} = \delta_{\pi\varsigma}\eta_\pi \tag{3.3.41}$$

yields

$$\begin{aligned}\|\mathscr{I}^P(v)\|^2_{L^2(\Omega^{(M)})} &= \int_\Omega \sum_{\pi\in\mathcal{Q}^P} v^2(\pi)\mathcal{L}^2_\pi(\xi(\omega))\,\mathrm{d}\mathbb{P} \\ &\leq \max_\pi \left(v^2(\pi)\sigma^2(\pi)\right) \sum_{\pi\in\mathcal{Q}^P} \int_\Omega \frac{\mathcal{L}^2_\pi(\xi(\omega))}{\sigma^2(\pi)}\,\mathrm{d}\mathbb{P}.\end{aligned}$$

If $\Omega^{(M)}$ is bounded, we have $\sigma = C_\sigma$ and (3.3.41), whence by (3.3.26)

$$\|\mathscr{I}^P(v)\|_{L^2(\Omega^{(M)})} \leq C\|v\|_{C^0_\sigma(\Omega^{(M)})}. \tag{3.3.42}$$

In the case of unbounded $\Omega^{(M)}$, we exploit the definition of σ and condition (3.3.39), which allows us to apply the convergence result for quadrature formulas in Appendix C to obtain

$$\sum_{\pi\in\mathcal{Q}^P} \int_\Omega \frac{\mathcal{L}^2_\pi(\xi(\omega))}{\sigma^2(\pi)}\,\mathrm{d}\mathbb{P} = \sum_{\pi\in\mathcal{Q}^P} \frac{1}{\sigma^2(\pi)}\eta_\pi \to \int_\Omega \frac{1}{\sigma^2(\xi(\omega))}\,\mathrm{d}\mathbb{P} \leq \frac{C_{\mathrm{pdf}}}{C^2_\sigma}\frac{\sqrt{2\pi}}{\delta_1}$$

as $P \to \infty$, which provides (3.3.42), too.

For each $\bar{v} \in \mathrm{Pol}^P(\Omega^{(M)})$, it holds $\mathscr{I}^P \bar{v} = \bar{v}$. Thus, it follows

$$\begin{aligned}\|v - \mathscr{I}^P v\|_{L^2(\Omega^{(M)})} &\le \|v - \bar{v}\|_{L^2(\Omega^{(M)})} + \|\mathscr{I}^P(v - \bar{v})\|_{L^2(\Omega^{(M)})} \\ &\le C\|v - \bar{v}\|_{C^0_\sigma(\Omega^{(M)})}\end{aligned}$$

and consequently (3.3.40), since $\bar{v}$ was arbitrary. □

Lemma 3.3.8. *Let $v \in C^0_{\sigma,0}(\Omega^{(M)}) \subset L^2(\Omega^{(M)})$. Then*

$$\|v - \mathscr{I}^P v\|_{L^2(\Omega^{(M)})} \to 0 \quad \text{as } P \to \infty. \tag{3.3.43}$$

Proof. The definitions of $\mathscr{I}^P$ and the underlying Gaussian points π and Lagrange polynomials $\mathcal{L}_\pi$ allow us to apply Proposition 2.2.7 and (3.1.23) with $H_r = L^2(\Omega_r)$, $A_n = \mathscr{I}^P$, $A^r_n = \mathscr{I}^{P_r}_r$, where $\mathscr{I}^{P_r}_r$ is the one-dimensional Lagrange interpolation in Ω_r. Hence, we only have to show

$$\|w - \mathscr{I}^{P_r}_r w\|_{L^2(\Omega_r)} \to 0 \quad \text{as } P_r \to \infty$$

for all $w \in C^0_{\sigma_r,0}(\Omega_r)$. By Lemma 3.3.7, it is sufficient to guarantee the existence of a sequence $(p_n)_n \subset \mathrm{Pol}^n(\Omega_r)$ of polynomials such that

$$\|(p_n - w)\sigma_r\|_\infty \to 0 \quad \text{as } n \to \infty.$$

If Ω_r is bounded, this is just the Weierstraß approximation theorem (see, e.g., [67]), otherwise this holds by [79, Theorem 1.4]. □

We can now state the consistency of $\phi^P_\jmath$.

Theorem 3.3.9. *Let H be a bounded and monotonically increasing function and $v \in C^\infty(\overline{D}) \otimes C^\infty_0(\Omega^{(M)})$. If (3.3.23), (3.3.27), (3.3.28), and (3.3.39) hold, then for the interpolated function $v^P_\jmath = \mathcal{I}_{\mathcal{S}_\jmath}(v^P)$ of the projected function $v^P = \mathcal{P}^P(v)$, shortly $v^P_\jmath = \mathcal{I}_{\mathcal{S}_\jmath} \otimes \mathcal{P}^P v \in \mathcal{S}_\jmath \otimes \mathcal{Z}^P$, we have*

$$v^P_\jmath \to v \text{ in } H^1(D) \otimes L^2(\Omega) \quad \text{as } \jmath \to \infty,\ P \to \infty \tag{3.3.44}$$

and

$$\phi^P_\jmath(v^P_\jmath) \to \phi(v) \quad \text{as } \jmath \to \infty,\ P \to \infty. \tag{3.3.45}$$

If, in addition, a function w satisfies (3.3.37) and (3.3.38), then the convergence results (3.3.44) and (3.3.45) are also valid for $v = w + \tilde{v} \in w + \left(C^\infty(\overline{D}) \otimes C^\infty_0(\Omega^{(M)})\right)$ and $v^P_\jmath = w^P_\jmath + \tilde{v}^P_\jmath = \mathcal{I}_{\mathcal{S}_\jmath} \otimes \mathcal{P}^P w + \mathcal{I}_{\mathcal{S}_\jmath} \otimes \mathcal{P}^P \tilde{v}$.

Proof. We split up the convergence error as

$$\|v - v^P_\jmath\| \le \|v - v^P\| + \|v^P - v^P_\jmath\|. \tag{3.3.46}$$

The first term on the right-hand side of (3.3.46) converges as $P \to \infty$ because of Theorem 3.3.1. Since $C^\infty(\overline{D})$ is dense in all $H^k(D)$ (see [25, p. 119]), the convergence of the second term as $\jmath \to \infty$ under the assumptions (3.3.27) and (3.3.28) is well known (see [25, Theorem 16.2]). Due to (3.3.38), the same holds for $v = w + \tilde{v}$ and $v^P_\jmath = \mathcal{I}_{\mathcal{S}_\jmath} \otimes \mathcal{P}^P w + \mathcal{I}_{\mathcal{S}_\jmath} \otimes \mathcal{P}^P \tilde{v}$.

It remains to show the consistency $\phi^P_\jmath(v^P_\jmath) \to \phi(v)$. Assume first that $\phi(v) = \infty$. Then we can find an open subset $E_1 \times E_2 \subset D \times \Omega^{(M)}$ with $v(x,\xi) < u_c$ for all $(x,\xi) \in E_1 \times E_2$ and $\mathbb{P}^{(M)}(E_2) > 0$, where E_2 contains a set of the form $\prod_{r=1}^M (a_r, b_r)$.

Due to $h_\jmath \to 0$ we can find a $\jmath^*$ such that $\mathcal{N}_\jmath \cap E_1 \neq \emptyset$ for all $\jmath \geq \jmath^*$. We can also find a P^* such that for all $P > P^*$ the corresponding quadrature point set always contains a $\pi \in E_2$, see Theorem C.2. This provides $\phi_\jmath^P(v_\jmath^P) \to \infty$ as $\jmath \to \infty$, $P \to \infty$.

In the case $\phi(v) < \infty$, i.e. $v(x,\xi) = v(x,\xi(\omega)) \geq u_c$ for all $x \in D$ and $\mathbb{P}$-almost all $\omega \in \Omega$, we begin by estimating

$$\begin{aligned}
|\phi(v) - \phi_\jmath^P(v_\jmath^P)| &= \left| \int_\Omega \int_D \Phi(v(x,\xi)) \,\mathrm{d}x \,\mathrm{d}\mathbb{P} - \sum_{\pi \in \mathcal{Q}^P} \sum_{p \in \mathcal{N}_\jmath} \Phi(v_\jmath^P(p,\pi)) h_p \eta_\pi \right| \\
&\leq \left| \int_D \left(\int_\Omega \Phi(v(x,\xi)) \,\mathrm{d}\mathbb{P} - \sum_{\pi \in \mathcal{Q}^P} \Phi(v^P(x,\pi)) \eta_\pi \right) \,\mathrm{d}x \right| \\
&\quad + \left| \sum_{\pi \in \mathcal{Q}^P} \eta_\pi \left(\int_D \Phi(v^P(x,\pi)) \,\mathrm{d}x - \sum_{p \in \mathcal{N}_\jmath} \Phi(v_\jmath^P(p,\pi)) h_p \right) \right| \\
&\leq \int_D \left| \int_\Omega \left(\Phi(v(x,\xi)) - \Phi(v^P(x,\xi)) \right) \,\mathrm{d}\mathbb{P} \right| \,\mathrm{d}x \\
&\quad + \left| \int_D \int_\Omega \left(\Phi(v^P(x,\xi)) - \sum_{\pi \in \mathcal{Q}^P} \Phi(v^P(x,\pi)) \mathcal{L}_\pi(\xi) \right) \,\mathrm{d}\mathbb{P} \,\mathrm{d}x \right| \\
&\quad + \left| \sum_{\pi \in \mathcal{Q}^P} \eta_\pi \left(\int_D \Phi(v^P(x,\pi)) \,\mathrm{d}x - \sum_{p \in \mathcal{N}_\jmath} \Phi(v_\jmath^P(p,\pi)) h_p \right) \right| \\
&=: (A_1) + (A_2) + (A_3).
\end{aligned}$$

The integrand of the spatial integral in term (A_1) can be estimated using the Lipschitz continuity of Φ (with Lipschitz constant L) and the Cauchy–Schwarz inequality by

$$\begin{aligned}
\left| \int_\Omega \left(\Phi(v(x,\xi)) - \Phi(v^P(x,\xi)) \right) \,\mathrm{d}\mathbb{P} \right| &\leq L \int_\Omega \left| v(x,\xi) - v^P(x,\xi) \right| \,\mathrm{d}\mathbb{P} \\
&\leq C \| v(x,\xi) - v^P(x,\xi) \|_{L^2(\Omega)} \to 0
\end{aligned}$$

as $P \to \infty$. Since this convergence is monotone, we can use the theorem of Lebesgue (see, e.g., [109]), whence $(A_1) \to 0$.

Using the identity operator Id, we can rewrite

$$(A_2) \leq \| (\mathrm{Id} \otimes \mathscr{I}^P - \mathrm{Id} \otimes \mathrm{Id}) \Phi(v^P) \|_{L^1(D) \otimes L^1(\Omega)} \leq C \| (\mathrm{Id} \otimes \mathscr{I}^P - \mathrm{Id} \otimes \mathrm{Id}) \Phi(v^P) \|_{0,0}. \tag{3.3.47}$$

Now apply Proposition 2.2.7 with $H_1 = L^2(D)$, $H_2 = L^2(\Omega)$, $A_P^1 = \mathrm{Id}$, $A_P^2 = \mathscr{I}^P$, which provides $(A2) \to 0$ as $P \to \infty$ by means of Lemma 3.3.8 if we can show

$$\Phi(v^P) \in L^2(D) \otimes C_{\sigma,0}^0(\Omega^{(M)}). \tag{3.3.48}$$

Indeed, $v \in C^\infty(\overline{D}) \otimes C_0^\infty(\Omega^{(M)})$ implies $v^P \in C^\infty(\overline{D}) \otimes C_{\sigma,0}^0(\Omega^{(M)})$, since v^P is a polynomial in $\Omega^{(M)}$, and thus (3.3.48), since Φ is Lipschitz continuous.

For the sum in (A_3), note that $\Phi(v^P(\cdot,\pi))$ is uniformly continuous on $\overline{D}$ for each Gaussian quadrature point π, since $v(\cdot,\pi) : \overline{D} \to [u_c, \infty)$ and $v^P(\cdot,\pi) : \overline{D} \to [u_c, \infty)$ are (in case of $v = w + \tilde{u}$ this follows from (3.3.37)). Denote by $p_{1\mathrm{t}}$, $p_{2\mathrm{t}}$, $p_{3\mathrm{t}}$ the

vertices of $\mathrm{t} \in \mathcal{T}_\jmath$. Due to $\sum_{i=1}^{3} s_{p_{i\mathrm{t}}}^{(\jmath)} = 1$ on the element t, it holds

$$\begin{aligned}\left| \int_D \Phi(v^P(x,\pi))\,\mathrm{d}x - \sum_{p\in\mathcal{N}_\jmath} \Phi(v_\jmath^P(p,\pi))h_p \right| \\ \leq \sum_{\mathrm{t}\in\mathcal{T}_\jmath} \int_{\mathrm{t}} \left(\sum_{i=1}^{3} s_{p_{i\mathrm{t}}}^{(\jmath)}(x) |\Phi(v^P(x,\pi)) - \Phi(v^P(p_{i\mathrm{t}},\pi))| \right) \mathrm{d}x \\ \leq |D| \max_{\substack{|x-y|\leq h_\jmath \\ x,y\in D}} |\Phi(v^P(x,\pi)) - \Phi(v^P(y,\pi))| \to 0\end{aligned}$$

as $\jmath \to \infty$. In combination with (3.3.26) we have $(A_3) \to 0$, and the proof is complete. □

Now, we have all tools together to formulate the discrete version of the minimization problem (3.2.26). It reads

$$u_\jmath^P \in \mathcal{K}_\jmath^P : \ \mathcal{J}(u_\jmath^P) + \phi_\jmath^P(u_\jmath^P) \leq \mathcal{J}(v) + \phi_\jmath^P(v) \quad \forall v \in \mathcal{K}_\jmath^P. \tag{3.3.49}$$

We will see that we can transfer most results from the continuous to the discrete model. We begin with the most important one.

Theorem 3.3.10. *With the assumptions of Lemma 3.3.6 and with the conditions on kr and K as in Theorem 3.2.11, the discrete minimization problem (3.3.49) has a unique solution.*

Proof. This follows from Proposition 3.2.10 with $V = \mathcal{S}_\jmath \otimes \mathcal{Z}^P$, $C = \mathcal{K}_\jmath^P$ and $F = \mathcal{J} + \phi_\jmath^P$. To show the coercivity of F, we proceed as in (3.2.27) using (3.3.35) instead of (3.2.13) and obtain

$$\mathcal{J}(v) + \phi_\jmath^P(v) \geq \frac{1}{2} c_1 \|v\|_{1,0}^2 - c_2 \|v\|_{1,0} - c_3 \to \infty \tag{3.3.50}$$

as $\|v\|_{1,0} \to \infty$. The rest is clear. □

By means of Proposition 3.2.16, it is easy to see that the minimization problem (3.3.49) is equivalent to the variational inequality

$$u_\jmath^P \in \mathcal{K}_\jmath^P : \ a(u_\jmath^P, v - u_\jmath^P) - \ell(v - u_\jmath^P) + \phi_\jmath^P(v) - \phi_\jmath^P(u_\jmath^P) \geq 0 \quad \forall v \in \mathcal{K}_\jmath^P. \tag{3.3.51}$$

In the following, we want to reformulate the discrete minimization problem (3.3.49) in terms of variational inequalities and variational inclusions in consideration of the boundary conditions as it was done for the continuous problem in Subsection 3.2.3. We recall the canonical extension from Remark 3.2.19 and apply it to the functional $\phi_\jmath^P$ with regard to the convex set $\mathcal{K}_\jmath^P$. Furthermore, we recall the definition of the translation operator introduced in (3.2.31).

We now choose a fixed $w_\jmath^P \in \mathcal{S}_\jmath \otimes \mathcal{Z}^P$ with

$$w_\jmath^P(p,\pi) = u_D(p,\pi) \quad \forall p \in \mathcal{N}_\jmath^D \, \forall \pi \in \mathcal{Q}^P$$

and set $u_\jmath^P = w_\jmath^P + \tilde{u}_\jmath^P$. The translated discrete convex set analogously to (3.2.32) is then given by

$$\mathcal{K}_{\jmath,\,\Gamma_D}^P := \mathcal{K}_\jmath^P - w_\jmath^P = \{ v \in \mathcal{S}_\jmath^D \otimes \mathcal{Z}^P : v(p,\pi) \geq u_c - w_\jmath^P(p,\pi) \, \forall p \in \mathcal{N}_\jmath \, \forall \pi \in \mathcal{Q}^P \}. \tag{3.3.52}$$

Again, as in Remark 3.3.5, we have in general $\mathcal{K}^P_{\jmath,\Gamma_D} \not\subset \mathcal{K}_{\Gamma_D}$.

Proceeding as in the derivation of Propositions 3.2.20, 3.2.21 and 3.2.27, we can now state their discrete counterparts.

Proposition 3.3.11. *The minimization problem (3.3.49) is equivalent to*

$$\tilde{u}^P_\jmath \in \mathcal{S}^D_\jmath \otimes \mathcal{Z}^P : \ \mathcal{J}_{w^P_\jmath}(\tilde{u}^P_\jmath) + (\phi^P_\jmath)_{w^P_\jmath}(\tilde{u}^P_\jmath) \le \mathcal{J}_{w^P_\jmath}(v) + (\phi^P_\jmath)_{w^P_\jmath}(v) \quad \forall v \in \mathcal{S}^D_\jmath \otimes \mathcal{Z}^P \tag{3.3.53}$$

and equivalent to the variational inequality

$$\begin{aligned}\tilde{u}^P_\jmath \in \mathcal{S}^D_\jmath \otimes \mathcal{Z}^P : \ & a(w^P_\jmath + \tilde{u}^P_\jmath, v - \tilde{u}^P_\jmath) - \ell(v - \tilde{u}^P_\jmath) \\ & + \phi^P_\jmath(w^P_\jmath + v) - \phi^P_\jmath(w^P_\jmath + \tilde{u}^P_\jmath) \ge 0 \quad \forall v \in \mathcal{S}^D_\jmath \otimes \mathcal{Z}^P\end{aligned} \tag{3.3.54}$$

and equivalent to the variational inclusion

$$\tilde{u}^P_\jmath \in \mathcal{S}^D_\jmath \otimes \mathcal{Z}^P : \ 0 \in a(w^P_\jmath + \tilde{u}^P_\jmath, \cdot) - \ell(\cdot) + \partial\phi^P_\jmath(w^P_\jmath + \tilde{u}^P_\jmath) \tag{3.3.55}$$

in $\left(\mathcal{S}^D_\jmath \otimes \mathcal{Z}^P\right)'$*, always in the sense that the solution* $u^P_\jmath$ *of (3.3.49) equals* $w^P_\jmath + \tilde{u}^P_\jmath$*.*

The assertion in Proposition 3.3.11 remains true if one replaces the space $\mathcal{S}^D_\jmath \otimes \mathcal{Z}^P$ with the set $\mathcal{K}^P_{\jmath,\Gamma_D}$ in (3.3.53) or (3.3.54).

Remark 3.3.12. Note that we can prove in the same way as in Proposition 3.2.22 an analogous well-posedness result for the discrete problem (3.3.49).

Remark 3.3.13. Since we can interchange the sums and the subdifferential according to [33, Prop. I.5.6], the subdifferential $\partial\phi^P_\jmath(v_0) \subset \left(\mathcal{S}_\jmath \otimes \mathcal{Z}^P\right)'$ can be expressed in view of (3.2.44) directly as

$$\partial\phi^P_\jmath(v_0)(v) = \sum_{\pi \in \mathcal{Q}^P} \sum_{p \in \mathcal{N}_\jmath} \partial\Phi(v_0(p,\pi))v(p,\pi)h_p\eta_\pi \quad \forall v \in \mathcal{S}_\jmath \otimes \mathcal{Z}^P \tag{3.3.56}$$

with $\operatorname{dom}\partial\phi^P_\jmath = \operatorname{dom}\phi^P_\jmath$. Note that the difference revealed in (3.2.51) does no longer play a role in the discrete world.

3.3.4 A convergence result for the limit cases

In the last subsection, we derived the discretized problem (3.3.49) with the unique solution $u^P_\jmath \in \mathcal{K}^P_\jmath$. The next obvious step should be the proof that

$$u^P_\jmath \to u \quad \text{in } H^1(D) \otimes L^2(\Omega) \quad \text{as } \jmath \to \infty,\ P \to \infty, \tag{3.3.57}$$

where u is the unique solution of problem (3.2.26). This is true for the deterministic Richards equation (1.2.13) with solution $\hat{u} \in \hat{\mathcal{K}}$ and the solution $\hat{u}_\jmath \in \hat{\mathcal{K}}_\jmath$ of the problem discretized with linear finite elements in $\mathcal{S}_\jmath$, i.e.

$$\hat{u}_\jmath \to \hat{u} \quad \text{in } H^1(D) \quad \text{as } \jmath \to \infty,$$

see [18, Theorem 2.5.9]. Analyzing the proofs of this theorem or of similar convergence results (e.g. [48, Theorem I.6.2], [68, Theorem 1.13]), we can break them down into two major components:

(i) There exist a set $\mathscr{M}$ which is dense in $H^1(D)$ and a mapping $r_\jmath : \mathscr{M} \to \mathcal{S}_\jmath$ for which $\lim_{\jmath\to\infty} \hat{\phi}_\jmath(r_\jmath\hat{v}) = \hat{\phi}(\hat{v})$ for all $\hat{v} \in \mathscr{M}$.

(ii) If $\hat{v}_\jmath \in \mathcal{S}_\jmath$ for all $\jmath$ and the sequence converges weakly to $\hat{v} \in H^1(D)$ as $\jmath \to \infty$, then

$$\liminf_{\jmath\to\infty} \hat{\phi}_\jmath(\hat{v}_\jmath) \geq \hat{\phi}(\hat{v}).$$

We try to transfer these two points into our setting. Concerning the consistency condition (i), this is already done in Theorem 3.3.9. For the stability condition (ii), a corresponding condition for the stochastic Richards equation would be

(ii') If $v_\jmath^P \in \mathcal{S}_\jmath \otimes \mathcal{Z}^P$ for all $\jmath$ and P and the sequence converges weakly to a $v \in H^1(D) \otimes L^2(\Omega)$ as $\jmath \to \infty$, $P \to \infty$, then

$$\liminf_{\substack{\jmath\to\infty \\ P\to\infty}} \phi_\jmath^P(v_\jmath^P) \geq \phi(v).$$

Condition (ii) can only be shown by an interplay of the convexity of Φ, the piecewise linearity of $\hat{v}_\jmath$ and the corresponding quadrature formula for $\hat{\phi}$, which provides

$$\hat{\phi}_\jmath(\hat{v}_\jmath) \geq \hat{\phi}(\hat{v}_\jmath) \quad \forall \hat{v}_\jmath \in \mathcal{S}_\jmath \tag{3.3.58}$$

and, together with the weak lower semicontinuity of $\hat{\phi}$,

$$\liminf_{\jmath\to\infty} \hat{\phi}_\jmath(\hat{v}_\jmath) \geq \liminf_{\jmath\to\infty} \hat{\phi}(\hat{v}_\jmath) \geq \hat{\phi}(\hat{v}).$$

However, the condition in the stochastic setting corresponding to (3.3.58)

$$\phi_\jmath^P(v_\jmath^P) \geq \phi(v_\jmath^P) \quad \forall v_\jmath^P \in \mathcal{S}_\jmath \otimes \mathcal{Z}^P$$

is false in general for $P > 1$, as can be seen by simple examples. Thus, it is not clear under which assumptions (ii') is valid or a proof for (3.3.57) can be given.

It is, however, possible to show the convergence (3.3.57) in some special cases, namely for the limit cases from Remark 3.2.32. We carry out the proof within a greater context by developing a convergence theory for stochastic obstacle problems as will be defined in the following.

Consider the stochastic obstacle problem

$$u \in \bar{\mathcal{K}} : \quad a(u, v-u) \geq \ell(v-u) \quad \forall v \in \bar{\mathcal{K}} \tag{3.3.59}$$

on the convex set

$$\bar{\mathcal{K}} := \left\{ v \in H_0^1(D) \otimes L^2(\Omega^{(M)}) : v \geq \Psi \text{ a.e. in } D \times \Omega^{(M)} \right\} \tag{3.3.60}$$

with an obstacle function $\Psi \in (H^1(D) \cap C^0(\overline{D})) \otimes (L^2(\Omega^{(M)}) \cap C^0(\Omega^{(M)}))$ which satisfies $\Psi(\cdot, \xi) \leq 0$ in a neighborhood of ∂D for all $\xi \in \Omega^{(M)}$ and $\Psi(x, \cdot) \leq 0$ outside a compact set $\bar{C}_1 \subset \Omega^{(M)}$ for all $x \in D$. Note that we take homogeneous Dirichlet conditions for sake of simplicity. Recall from (3.1.22) and Remark 3.1.23 that $(\Omega^{(M)}, \mathbb{P}^{(M)})$ and $(\Omega, \mathbb{P})$ can be identified in L^2 sense. For the following results however, this background via the Karhunen–Loève expansion is not necessary and we only assume for the rest of this section that $(\Omega^{(M)}, \mathrm{Bor}(\Omega^{(M)}), \mathbb{P}^{(M)})$ is a probability space with $\Omega^{(M)} \subset \mathbb{R}^M$, where $\ell(\cdot)$ is a linear continuous functional on $V = H_0^1(D) \otimes L^2(\Omega^{(M)})$ and $a(\cdot, \cdot)$ is a continuous and coercive bilinear form on $V \times V$.

Theorem 3.3.14. *The stochastic obstacle problem (3.3.59) has a unique solution.*

Proof. We apply Proposition 3.2.10 with $V = H_0^1(D) \otimes L^2(\Omega^{(M)})$, $C = \bar{\mathcal{K}}$ and $F(v) = \frac{1}{2}a(v,v) - \ell(v)$ and only have to check that the conditions on F and $\bar{\mathcal{K}}$ are satisfied. As in the proof of Theorem 3.2.11, we detect that F is strictly convex, lower semicontinuous, proper and coercive.

Furthermore, it is clear that $\bar{\mathcal{K}}$ is convex. To show the closedness of $\bar{\mathcal{K}}$, one can proceed as in Proposition 2.3.4 by replacing u_c with Ψ. Finally, $\bar{\mathcal{K}}$ is nonempty because $\Psi^+ := \max(0, \Psi) \in \bar{\mathcal{K}}$ due to the conditions on Ψ. □

Proceeding as in Subsections 3.3.1 and 3.3.2, the discretized problem now reads

$$u_\jmath^P \in \bar{\mathcal{K}}_\jmath^P : \quad a(u_\jmath^P, v - u_\jmath^P) \geq \ell(v - u_\jmath^P) \quad \forall v \in \bar{\mathcal{K}}_\jmath^P \tag{3.3.61}$$

on the convex set

$$\bar{\mathcal{K}}_\jmath^P := \{v \in \mathcal{S}_\jmath \otimes \mathcal{Z}^P : v(p,\pi) \geq \Psi(p,\pi)\, \forall p \in \mathcal{N}_\jmath \, \forall \pi \in \mathcal{Q}^P\}. \tag{3.3.62}$$

We want to show that the discretized solutions $u_\jmath^P$ converge to the solution $u \in \bar{\mathcal{K}}$ as $\jmath \to \infty$ and $P \to \infty$. First, we state the following lemma. Recall the definition of $\mathscr{D}$ from (2.2.6).

Lemma 3.3.15. *Under the above assumptions on $\bar{\mathcal{K}}$, the set $\mathscr{D} \cap \bar{\mathcal{K}}$ is dense in $\bar{\mathcal{K}}$.*

Proof. The proof proceeds in the same way as the proof of Lemma II.2.4 in [48], just by performing it in our tensor space instead of $H_0^1(D)$. We sketch it for completeness.

Let $v \in \bar{\mathcal{K}}$. Lemma 2.2.5 implies the existence of a sequence $(\tilde{v}_n)_n \subset \mathscr{D}$ with $\|\tilde{v}_n - v\|_{1,0} \to 0$. Defining

$$v_n := \max(\Psi, \tilde{v}_n),$$

the functions v_n also converge strongly to v, since $v \in \bar{\mathcal{K}}$. Moreover, the conditions on $\bar{\mathcal{K}}$ imply that v_n is a sequence in the set

$$\mathscr{C} := \left\{ v \in \bar{\mathcal{K}} \cap \left(C^0(\overline{D}) \otimes C^0(\overline{\Omega^{(M)}}) \right) : v \text{ has compact support in } D \times \Omega^{(M)} \right\}. \tag{3.3.63}$$

Therefore, it is sufficient to prove the existence of a sequence $(v_n)_n \subset \mathscr{D} \cap \bar{\mathcal{K}}$ with

$$\|v_n - v\|_{1,0} \to 0$$

for an arbitrary $v \in \mathscr{C}$. To this end, we approximate $v \in \mathscr{C}$ as in the proof of Lemma 2.2.5 by the sum $v_N = \sum_{i=1}^N v_i^D v_i^\Omega$ of functions $v_i^D \in H_0^1(D) \cap C^0(\overline{D})$, $v_i^\Omega \in L^2(\Omega^{(M)}) \cap C^0(\overline{\Omega^{(M)}})$ with compact support, respectively, such that

$$\|v_N - v\|_{1,0} < \varepsilon/2,$$

and take a sequence of mollifiers $(\varphi_n^D) \in C_0^\infty(\mathbb{R}^d)$ and $(\varphi_n^\Omega) \in C_0^\infty(\mathbb{R}^M)$ with decreasing support and

$$\bigcap_{n=1}^{\infty} \operatorname{supp}(\varphi_n^D) \times \operatorname{supp}(\varphi_n^\Omega) = \{0\}. \tag{3.3.64}$$

Let

$$\tilde{v}_n := v_N * (\varphi_n^D \otimes \varphi_n^\Omega) := \sum_{i=1}^{N} \int_{\mathbb{R}^d} \varphi_n^D(x-x') v_i^D(x') \,\mathrm{d}x' \int_{\mathbb{R}^M} \varphi_n^\Omega(\xi-\xi') v_i^\Omega(\xi') \,\mathrm{d}\mathbb{P}^{(M)}(\xi'),$$

then $\tilde{v}_n \in C_0^\infty(\mathbb{R}^d) \otimes C_0^\infty(\mathbb{R}^M)$ with

$$\mathrm{supp}(\tilde{v}_n) \subset \mathrm{supp}(v_N) + \mathrm{supp}(\varphi_n^D) \times \mathrm{supp}(\varphi_n^\Omega)$$

and $\|v_N - \tilde{v}_n\|_{1,0} < \varepsilon/2$ if n is large enough. Thus, $\tilde{v}_n \to v$ in $H^1(\mathbb{R}^d) \otimes L^2(\mathbb{R}^M)$ and, due to (3.3.64) and the bounded supports, even in $L^\infty(\mathbb{R}^d) \otimes L^\infty(\mathbb{R}^M)$. This is still valid if $\tilde{v}_n$ is restricted to $D \times \Omega^{(M)}$ such that $\tilde{v}_n \in \mathscr{D}$.

Define $D_\delta := \{x \in D : \mathrm{dist}(x, \partial D) < \delta\}$. Take a compact set $\bar{C}_2^n \subset \Omega^{(M)}$ which is a superset of $\bar{C}_1$ and the support of $\tilde{v}_n$ in $\Omega^{(M)}$-direction and a $\delta > 0$ such that $\Psi(x,\xi) \le 0$ and $v = 0$ in $D_\delta \times (\Omega^{(M)} \backslash \bar{C}_2^n)$. Due to the convergence of $\tilde{v}_n$ in $L^\infty(\mathbb{R}^d) \otimes L^\infty(\mathbb{R}^M)$, there exists for all $\varepsilon > 0$ an $n_0 = n_0(\varepsilon)$ such that

$$v(x,\xi) - \varepsilon \le \tilde{v}_n(x,\xi) \le v(x,\xi) + \varepsilon \quad \forall (x,\xi) \in (D \backslash D_{\delta/2}) \times \bar{C}_2^n \tag{3.3.65}$$

and

$$\tilde{v}_n(x,\xi) = 0 \ge \Psi(x,\xi) \quad \forall (x,\xi) \in D_{\delta/2} \times (\Omega^{(M)} \backslash \bar{C}_2^n) \tag{3.3.66}$$

for all $n > n_0$. Take a function $\theta_n \in \mathscr{D}$ with $\theta_n \ge 0$ in $D \times \Omega^{(M)}$ and $\theta_n(x,\xi) = 1$ for all $(x,\xi) \in (\overline{D} \backslash D_{\delta/2}) \times \bar{C}_2^n$. Finally, define $v_n^\varepsilon := \tilde{v}_n + \varepsilon\theta_n$. Then, $v_n^\varepsilon \in \mathscr{D}$ with $v_n^\varepsilon(x,\xi) \ge \Psi(x,\xi)$ for all $(x,\xi) \in D \times \Omega^{(M)}$ by (3.3.65) and (3.3.66), and it is

$$\|v_n^\varepsilon - v\|_{1,0} \to 0 \quad \text{as } \varepsilon \to 0 \text{ and } n \to \infty,\ n \ge n_0(\varepsilon).$$

This concludes the proof. □

We now turn to the convergence proof.

Theorem 3.3.16. *With the above assumptions on Ψ and the conditions (3.3.23), (3.3.27), and (3.3.28) to the discretization and (3.3.39), we have*

$$u_\jmath^P \to u \text{ in } H_0^1(D) \otimes L^2(\Omega^{(M)}) \quad \text{as } \jmath \to \infty,\ P \to \infty, \tag{3.3.67}$$

where $u_\jmath^P$ is the solution of (3.3.61) and u is the solution of (3.3.59).

Proof. At the beginning, we show the following two approximation results:

(K1) If $(v_\jmath^P)_{\jmath,P}$ is such that $v_\jmath^P \in \bar{\mathcal{K}}_\jmath^P$ for all $\jmath$ and P and that it converges weakly to v as $\jmath \to \infty$ and $P \to \infty$, then $v \in \bar{\mathcal{K}}$.

(K2) There exist a set $\mathscr{M}$ which is dense in $\bar{\mathcal{K}}$ and an operator $r_\jmath^P : \mathscr{M} \to \bar{\mathcal{K}}_\jmath^P$ such that $\|r_\jmath^P v - v\|_{1,0} \to 0$ as $\jmath \to \infty$ and $P \to \infty$ for all $v \in \mathscr{M}$.

Let us first prove (K1). Consider a function $\varphi = \sum_{i=1}^N \varphi_i^D \varphi_i^\Omega \in \mathscr{D}$ with $\varphi \ge 0$, where $\varphi_i^D \in C_0^\infty(D)$ and $\varphi_i^\Omega \in C_0^\infty(\Omega^{(M)})$, and define $\varphi_\jmath^P$ by

$$\varphi_\jmath^P := \sum_{\pi \in \mathcal{Q}^P} \sum_{\mathrm{t} \in \mathcal{T}_\jmath} \varphi(G_\mathrm{t}, \pi) \mathbf{1}_\mathrm{t} \mathcal{L}_\pi = \mathscr{I}^P \left(\sum_{\mathrm{t} \in \mathcal{T}_\jmath} \varphi(G_\mathrm{t}, \cdot) \mathbf{1}_\mathrm{t} \right). \tag{3.3.68}$$

Here, G_t is the centroid of the triangle $\mathrm{t} \in \mathcal{T}_\jmath$ with vertices $p_{1\mathrm{t}}$, $p_{2\mathrm{t}}$, $p_{3\mathrm{t}}$. The triangle inequality and the definition of the tensor scalar product provide

$$\begin{aligned}\|\varphi - \varphi_\jmath^P\|_{0,0} \leq & \left\| \sum_{i=1}^N \left(\varphi_i^D - \sum_{\mathrm{t}\in\mathcal{T}_\jmath} \varphi_i^D(G_\mathrm{t})\mathbf{1}_\mathrm{t} \right) \varphi_i^\Omega \right\|_{0,0} \\ & + \left\| \sum_{i=1}^N \left(\sum_{\mathrm{t}\in\mathcal{T}_\jmath} \varphi_i^D(G_\mathrm{t})\mathbf{1}_\mathrm{t} \right) \left(\varphi_i^\Omega - \sum_{\pi\in\mathcal{Q}^P} \varphi_i^\Omega(\pi)\mathcal{L}_\pi \right) \right\|_{0,0} \\ \leq C \sum_{i=1}^N & \left\| \varphi_i^D - \sum_{\mathrm{t}\in\mathcal{T}_\jmath} \varphi_i^D(G_\mathrm{t})\mathbf{1}_\mathrm{t} \right\|_{L^2(D)} + C \sum_{i=1}^N \left\| \varphi_i^\Omega - \sum_{\pi\in\mathcal{Q}^P} \varphi_i^\Omega(\pi)\mathcal{L}_\pi \right\|_{L^2(\Omega^{(M)})},\end{aligned}$$

since $\|\varphi_i^\Omega\|_{L_2(\Omega^{(M)})}$ and $\left\|\sum_{\mathrm{t}\in\mathcal{T}_\jmath} \varphi_i^D(G_\mathrm{t})\mathbf{1}_\mathrm{t}\right\|_{L_2(D)}$ are bounded (the latter because it is convergent). Now,

$$\left\| \varphi_i^D - \sum_{\mathrm{t}\in\mathcal{T}_\jmath} \varphi_i^D(G_\mathrm{t})\mathbf{1}_\mathrm{t} \right\|_{L^2(D)} \to 0 \quad \text{as } \jmath \to \infty$$

because of the uniform continuity of φ_i^D (we even have convergence in $L^\infty(D)$) and

$$\left\| \varphi_i^\Omega - \sum_{\pi\in\mathcal{Q}^P} \varphi_i^\Omega(\pi)\mathcal{L}_\pi \right\|_{L^2(\Omega^{(M)})} = \left\|\varphi_i^\Omega - \mathscr{I}^P\varphi_i^\Omega\right\|_{L^2(\Omega^{(M)})} \to 0 \quad \text{as } P \to \infty$$

with Lemma 3.3.8, since $\varphi_i^\Omega \in C_0^\infty(\Omega^{(M)})$. Altogether, we have

$$\|\varphi - \varphi_\jmath^P\|_{0,0} \to 0 \quad \text{as } \jmath \to \infty,\ P \to \infty. \tag{3.3.69}$$

For the approximation of Ψ by $\Psi_\jmath^P$, we take piecewise linear interpolation in D and Lagrange interpolation in $\Omega^{(M)}$, i.e. $\Psi_\jmath^P = \mathcal{I}_{\mathcal{S}_\jmath} \otimes \mathscr{I}^P\Psi$. In particular, we have

$$\Psi(p,\pi) = \Psi_\jmath^P(p,\pi) \quad \forall p \in \mathcal{N}_\jmath\, \forall \pi \in \mathcal{Q}^P.$$

One can approximate Ψ by the function $\Psi_N = \sum_{i=1}^N \Psi_i^D\Psi_i^\Omega$, where $\Psi_i^D \in C^\infty(\overline{D})$ and $\Psi_i^\Omega \in C_0^\infty(\Omega^{(M)})$, since $C^\infty(\overline{D})$ is dense in $H^1(D)$, analogously to Lemma 2.2.5. By definition of the operator tensor product in (2.2.8), $\mathcal{I}_{\mathcal{S}_\jmath} \otimes \mathscr{I}^P\Psi_N$ is an approximation of $\Psi_\jmath^P$. Moreover, we can reduce as above

$$\|\mathcal{I}_{\mathcal{S}_\jmath}\otimes\mathscr{I}^P\Psi_N - \Psi_N\|_{0,0} \leq C\sum_{i=1}^N \left\|\Psi_i^D - \mathcal{I}_{\mathcal{S}_\jmath}\Psi_i^D\right\|_{L^2(D)} + C\sum_{i=1}^N \left\|\Psi_i^\Omega - \mathscr{I}^P\Psi_i^\Omega\right\|_{L^2(\Omega^{(M)})},$$

where the convergence $\left\|\Psi_i^D - \mathcal{I}_{\mathcal{S}_\jmath}\Psi_i^D\right\|_{L^2(D)} \to 0$ as $\jmath \to \infty$ is clear [25, Theorem 16.2] and the convergence

$$\left\|\Psi_i^\Omega - \mathscr{I}^P\Psi_i^\Omega\right\|_{L^2(\Omega^{(M)})} \to 0 \quad \text{as } P \to \infty$$

is provided by Lemma 3.3.8.

Hence, we have

$$\|\Psi_\jmath^P - \Psi\|_{0,0} \to 0 \quad \text{as } \jmath \to \infty,\ P \to \infty. \tag{3.3.70}$$

Now, let $(v_\jmath^P)_{\jmath,P} \subset \bar{\mathcal{K}}_\jmath^P$ be a sequence converging weakly to $v \in H_0^1(D) \otimes L^2(\Omega^{(M)})$ as $\jmath \to \infty$, $P \to \infty$. Then, (3.3.69) and (3.3.70) imply

$$\lim_{\substack{\jmath\to\infty \\ P\to\infty}} (v_\jmath^P - \Psi_\jmath^P, \varphi_\jmath^P)_{0,0} = (v - \Psi, \varphi)_{0,0}. \tag{3.3.71}$$

Inserting (3.3.68), this can be rewritten as

$$(v_\jmath^P - \Psi_\jmath^P, \varphi_\jmath^P)_{0,0} = \int_{\Omega^{(M)}} \int_D (v_\jmath^P - \Psi_\jmath^P) \left(\sum_{\pi\in\mathcal{Q}^P} \sum_{\mathrm{t}\in\mathcal{T}_\jmath} \varphi(G_\mathrm{t}, \pi) \mathbf{1}_\mathrm{t} \mathcal{L}_\pi \right) \mathrm{d}x \, \mathrm{d}\mathbb{P}^{(M)}.$$

We now exploit the exactness of the Gaussian quadrature for functions in $\bar{\mathcal{K}}_\jmath^P$, the identity $\mathcal{L}_\pi(\varsigma) = \delta_{\pi\varsigma}$ for $\pi, \varsigma \in \mathcal{Q}^P$ and the exact quadrature formula for linear FE functions

$$\int_\mathrm{t} w \, \mathrm{d}x = \frac{|\mathrm{t}|}{3} \sum_{i=1}^3 w(p_{i\mathrm{t}})$$

(see [48, p. 34]) and obtain by the definition of $\bar{\mathcal{K}}_\jmath^P$

$$\begin{aligned}
(v_\jmath^P - \Psi_\jmath^P, \varphi_\jmath^P)_{0,0} &= \sum_{\pi\in\mathcal{Q}^P} \eta_\pi \left(\sum_{\mathrm{t}\in\mathcal{T}_\jmath} \varphi(G_\mathrm{t}, \pi) \int_\mathrm{t} \left(v_\jmath^P(x,\pi) - \Psi_\jmath^P(x,\pi) \right) \mathrm{d}x \right) \\
&= \sum_{\pi\in\mathcal{Q}^P} \eta_\pi \left(\sum_{\mathrm{t}\in\mathcal{T}_\jmath} \varphi(G_\mathrm{t}, \pi) \frac{|\mathrm{t}|}{3} \sum_{i=1}^3 \left(v_\jmath^P(p_{i\mathrm{t}}, \pi) - \Psi_\jmath^P(p_{i\mathrm{t}}, \pi) \right) \right) \\
&\geq 0
\end{aligned}$$

for all $\varphi \in \mathscr{D}$ with $\varphi \geq 0$. Letting $\jmath \to \infty$ and $P \to \infty$ provides by (3.3.71)

$$\int_{\Omega^{(M)}} \int_D (v - \Psi) \varphi \, \mathrm{d}x \, \mathrm{d}\mathbb{P}^{(M)} \geq 0 \quad \forall \varphi \in \mathscr{D} \text{ with } \varphi \geq 0,$$

which implies $v \geq \Psi$ a.e. in $D \times \Omega^{(M)}$. Thus, (K1) is proven.

We now turn to (K2). In view of Lemma 3.3.15, we take $\mathscr{M} = \mathscr{D} \cap \bar{\mathcal{K}}$ and again $r_\jmath^P = \mathcal{I}_{\mathcal{S}_\jmath} \otimes \mathscr{I}^P$. Obviously, it is $r_\jmath^P v \in \bar{\mathcal{K}}_\jmath^P$ for all $v \in \mathscr{M}$, since $r_\jmath^P v(p,\pi) = v(p,\pi) \geq \Psi(p,\pi)$ for all $p \in \mathcal{N}_\jmath$ and $\pi \in \mathcal{Q}^P$ by construction.

We still have to show $\|r_\jmath^P v - v\|_{1,0} \to 0$ as $j \to \infty$, $P \to \infty$. With $v = \sum_{i=1}^N v_i^D v_i^\Omega$ and $r_\jmath^P v = \sum_{i=1}^N (\mathcal{I}_{\mathcal{S}_\jmath} v_i^D)(\mathscr{I}^P v_i^\Omega)$, we can estimate as above

$$\|r_\jmath^P v - v\|_{1,0} \leq C \sum_{i=1}^N \left\| v_i^D - \mathcal{I}_{\mathcal{S}_\jmath} v_i^D \right\|_{H^1(D)} + C \sum_{i=1}^N \left\| v_i^\Omega - \mathscr{I}^P v_i^\Omega \right\|_{L^2(\Omega^{(M)})},$$

where the convergence $\left\| v_i^D - \mathcal{I}_{\mathcal{S}_\jmath} v_i^D \right\|_{H^1(D)} \to 0$ as $\jmath \to \infty$ is again provided by [25, Theorem 16.2] and the convergence $\left\| v_i^\Omega - \mathscr{I}^P v_i^\Omega \right\|_{L^2(\Omega^{(M)})} \to 0$ as $P \to \infty$ is again ensured by Lemma 3.3.8, since $v_i^\Omega \in C_0^\infty(\Omega^{(M)})$. This validates (K2).

This preliminary work done, we can turn to the convergence. We follow [48, Theorem I.5.2] and divide the proof of (3.3.67) into three parts: first show the boundedness of $(u_\jmath^P)_{\jmath,P}$, then the weak convergence, and finally the strong convergence of this sequence.

We will now show that there exist constants C_1 and C_2 independent of $\jmath$ and P such that

$$\|u_\jmath^P\|_{1,0}^2 \leq C_1 \|u_\jmath^P\|_{1,0} + C_2 \tag{3.3.72}$$

for all $\jmath$ and P. By (3.3.61), we have

$$a(u_\jmath^P, u_\jmath^P) \leq a(u_\jmath^P, v_\jmath^P) - \ell(v_\jmath^P - u_\jmath^P) \quad \forall v_\jmath^P \in \bar{\mathcal{K}}_\jmath^P,$$

whence by means of the coercivity (3.2.7)

$$c\|u_\jmath^P\|_{1,0}^2 \leq C\|u_\jmath^P\|_{1,0}\|v_\jmath^P\|_{1,0} + \|\ell\|(\|u_\jmath^P\|_{1,0} + \|v_\jmath^P\|_{1,0}) \quad \forall v_\jmath^P \in \bar{\mathcal{K}}_\jmath^P. \tag{3.3.73}$$

Let $v \in \mathscr{M}$ and $v_\jmath^P = r_\jmath^P v \in \bar{\mathcal{K}}_\jmath^P$. Since $\|r_\jmath^P v - v\|_{1,0} \to 0$ due to (K2), we know that $\|v_\jmath^P\|_{1,0}$ is uniformly bounded by a constant C_3. Hence, (3.3.73) can be written as

$$\|u_\jmath^P\|_{1,0}^2 \leq \frac{1}{c}\left(CC_3 + \|\ell\|\right)\|u_\jmath^P\|_{1,0} + \|\ell\|C_3\big).$$

Thus, we have (3.3.72), which implies $\|u_\jmath^P\|_{1,0} \leq C_4$ for all $\jmath$ and P.

As second step, we prove that $u_\jmath^P$ converges weakly to u. Since it was just shown that $u_\jmath^P$ is uniformly bounded, we can exploit that $H_0^1(D) \otimes L^2(\Omega^{(M)})$ is a Hilbert space and state the existence of a subsequence $(u_{\jmath_i}^{P_i})$ which converges weakly to a $u^* \in H_0^1(D) \otimes L^2(\Omega^{(M)})$. By (K1), we have $u^* \in \bar{\mathcal{K}}$. Now show that u^* is a solution of (3.3.59). For $v \in \mathscr{M}$ and $v_{\jmath_i}^{P_i} = r_{\jmath_i}^{P_i} v$, we can state as above

$$a(u_{\jmath_i}^{P_i}, u_{\jmath_i}^{P_i}) \leq a(u_{\jmath_i}^{P_i}, r_{\jmath_i}^{P_i} v) - \ell(r_{\jmath_i}^{P_i} v - u_{\jmath_i}^{P_i}). \tag{3.3.74}$$

Since $r_{\jmath_i}^{P_i} v$ converges strongly to v and $u_{\jmath_i}^{P_i}$ converges weakly to u^* as $\jmath_i \to \infty$, $P_i \to \infty$, taking the limit in (3.3.74) provides

$$\liminf_{\substack{\jmath_i \to \infty \\ P_i \to \infty}} a(u_{\jmath_i}^{P_i}, u_{\jmath_i}^{P_i}) \leq a(u^*, v) - \ell(v - u^*) \quad \forall v \in \mathscr{M}. \tag{3.3.75}$$

On the other hand, it is

$$0 \leq a(u_{\jmath_i}^{P_i} - u^*, u_{\jmath_i}^{P_i} - u^*) = a(u_{\jmath_i}^{P_i}, u_{\jmath_i}^{P_i}) - a(u_{\jmath_i}^{P_i}, u^*) - a(u^*, u_{\jmath_i}^{P_i}) + a(u^*, u^*),$$

which can be rewritten as

$$a(u_{\jmath_i}^{P_i}, u^*) + a(u^*, u_{\jmath_i}^{P_i}) - a(u^*, u^*) \leq a(u_{\jmath_i}^{P_i}, u_{\jmath_i}^{P_i}).$$

Taking the limit, we obtain

$$a(u^*, u^*) \leq \liminf_{\substack{\jmath_i \to \infty \\ P_i \to \infty}} a(u_{\jmath_i}^{P_i}, u_{\jmath_i}^{P_i}). \tag{3.3.76}$$

Combining (3.3.75) and (3.3.76), we have $u^* \in \bar{\mathcal{K}}$ with

$$a(u^*, v - u^*) \geq \ell(v - u^*) \quad \forall v \in \mathscr{M}.$$

From the density of $\mathscr{M}$ and the continuity of $a(\cdot,\cdot)$ and $\ell(\cdot)$, we can conclude

$$a(u^*, v - u^*) \geq \ell(v - u^*) \quad \forall v \in \bar{\mathcal{K}}. \tag{3.3.77}$$

Since the solution of (3.3.59) is unique according to Theorem 3.3.14, we obtain $u^* = u$. Hence, u is the only cluster point of $(u_\jmath^P)$ in the weak topology. Consequently, the whole sequence converges weakly to u.

As last step, we show that $u_\jmath^P$ even converges strongly to u. With $v \in \mathscr{M}$, we can derive analogously to (3.3.74) that

$$a(u_\jmath^P, u_\jmath^P) \leq a(u_\jmath^P, r_\jmath^P v) - \ell(r_\jmath^P v - u_\jmath^P) \tag{3.3.78}$$

with $r_\jmath^P v \in \bar{\mathcal{K}}$. By coercivity, we have moreover

$$0 \le c\|u_\jmath^P - u\|_{1,0}^2 \le a(u_\jmath^P - u, u_\jmath^P - u) = a(u_\jmath^P, u_\jmath^P) - a(u, u_\jmath^P) - a(u_\jmath^P, u) + a(u,u). \tag{3.3.79}$$

Since $u_\jmath^P$ converges weakly to u and $r_\jmath^P v$ converges strongly to v by (K2) as $\jmath \to \infty$ and $P \to \infty$, we can combine (3.3.78) and (3.3.79) and take the limit to obtain

$$0 \le c \liminf_{\substack{\jmath\to\infty \\ P\to\infty}} \|u_\jmath^P - u\|_{1,0}^2 \le c \limsup_{\substack{\jmath\to\infty \\ P\to\infty}} \|u_\jmath^P - u\|_{1,0}^2 \le a(u, v-u) - \ell(v-u) \tag{3.3.80}$$

for all $v \in \mathscr{M}$. By density and continuity, (3.3.80) also holds for all $v \in \bar{\mathcal{K}}$. Taking $v = u$ in (3.3.80), we obtain

$$\lim_{\substack{\jmath\to\infty \\ P\to\infty}} \|u_\jmath^P - u\|_{1,0} = 0,$$

as stated. □

We apply this for the limit cases of the Richards equation.

Theorem 3.3.17. *Let the discretization satisfy (3.3.23), (3.3.27), (3.3.28), and let (3.3.39) hold with $\Gamma_N = \emptyset$. Moreover, let the function w be an element of $(H^1(D) \cap C^0(\overline{D})) \otimes (L^2(\Omega^{(M)}) \cap C^0(\Omega^{(M)}))$ and satisfy (3.3.37)–(3.3.38). If $H = H_0$ or $H = H_\infty$ as defined in (1.1.18) and (1.1.20), respectively, then we have*

$$u_\jmath^P \to u \quad \text{in } H^1(D) \otimes L^2(\Omega) \quad \text{as } \jmath \to \infty,\ P \to \infty, \tag{3.3.81}$$

where u is the unique solution of problem (3.2.26) and $u_\jmath^P \in \mathcal{K}_\jmath^P$ is the unique solution of the discretized problem (3.3.49).

Proof. Recall from Remark 3.2.32 that (3.2.26) can be written as a stochastic obstacle problem

$$\tilde{u} \in \mathcal{K}_{\Gamma_D} :\ a(\tilde{u}, v - \tilde{u}) - \tilde{\ell}(v - \tilde{u}) \ge 0 \quad \forall v \in \mathcal{K}_{\Gamma_D}$$

with $u = w + \tilde{u}$ and

$$\mathcal{K}_{\Gamma_D} = \{v \in H^1_{\Gamma_D}(D) \otimes L^2(\Omega) : v \ge u_c - w\}$$

from (3.2.32). Setting $\Psi = u_c - w$, (3.3.81) now follows from Theorem 3.3.16. □

3.4 Numerical experiments

In this section, we turn to the error made by the discretization presented in this chapter. In the following, we will disregard the error from the Karhunen–Loève expansion and the time discretization and solely concentrate on the error from the PC and FE approximation.

At the beginning, let us recall known error estimates for the linear case. Consider $H \equiv 0$ with homogeneous Dirichlet conditions and $\Gamma_N = \emptyset$, which leads to the linear diffusion equation

$$u \in H^1_0(D) \otimes L^2(\Omega) :\ a(u, v) = \ell(v) \quad \forall v \in H^1_0(D) \otimes L^2(\Omega), \tag{3.4.1}$$

confer Remark 3.2.14. This problem was examined in depth in Babuška et al. [11, 12] including a theoretical analysis of the discretization error. Their result reads as follows.

Theorem 3.4.1. *Let u be the solution of (3.4.1) with a functional $\ell(v) = \mathbb{E}[\int_D f v \,\mathrm{d}x]$ with a continuous and bounded $f \in L^2(D) \otimes L^2(\Omega)$ expanded as in (3.1.34) and let $\alpha \in (0,1)$. Assume that the KL expansion (3.1.11) for K is strongly uniformly coercive in the sense that there exists a constant $\tilde{C} > 0$ such that*

$$\min_{(x,\omega)\in\overline{D}\times\Omega,} \left(K(x,\omega) - \sqrt{\lambda_{r_0}} g_{r_0}(x)\xi_{r_0}(\omega)\right) - \sqrt{\lambda_{r_0}}\|g_{r_0}\|_{L^\infty(D)}\|\xi_{r_0}\|_{L^\infty(\Omega)} \geq \tilde{C}$$

for all $r_0 = 1,\dots,M$ and differentiable on D in the sense that $\bar{K}$, $g_r \in C^1(\overline{D})$ for all $r = 1,\dots,M$. Furthermore, let Assumption 3.1.21 hold with bounded intervals $\Omega_r = \xi_r(\Omega)$ and bounded density functions $\mathrm{pdf}_r(\cdot)$. Then, u is analytic with respect to ξ and there exist a constant $C > 0$, independent of $\jmath$ and P, and a constant $0 < \beta < 1$, only depending on α and C, such that

$$\inf_{v\in\mathcal{S}_\jmath\otimes\mathcal{Z}^P} \|u - v\|_{1,0} \leq C\left(h_\jmath + \frac{1}{\alpha}\sum_{r=1}^{M}\beta^{P_r+1}\right) \tag{3.4.2}$$

and

$$\inf_{v\in\mathcal{S}_\jmath\otimes\mathcal{Z}^P} \|\mathbb{E}[u] - \mathbb{E}[v]\|_{L^2(D)} \leq C\left(h_j^2 + \frac{1}{\alpha}\sum_{r=1}^{M}\beta^{2P_r+2}\right). \tag{3.4.3}$$

We will see in the following that our numerical experiments suggest similar error bounds for the stochastic Richards equation. To this end, we set up two test cases in space dimensions $d = 1$ and $d = 2$ and solve them with the methods described in Chapter 4. Another numerical test with a stochastic obstacle problem showing similar results was carried out in [37].

3.4.1 Results for one space dimension

We start with a very simple problem in one space and one stochastic dimension with a smooth solution u. Nevertheless, it allows to draw some interesting conclusions which will be supported by the results in Subsection 3.4.2.

We choose $\eta = 10^{-3}\,[kg/ms]$, $\rho = 10^3\,[kg/m^3]$, $g = 10\,[m/s]$, $\mathfrak{n} = 1$ as parameters in a sandy soil on a domain $D = (-10, 10)$. Then, we utilize Brooks–Corey functions with $\lambda = 2/3$, $\theta_m = 0.21$, $\theta_M = 0.95$, $p_b = -1$ as in the example in Section 1.1. The generalized saturation $H(u)$ is given by (1.1.15) (see Figure 1.5) with critical generalized pressure $u_c = -4/3$. The uniformly distributed permeability is given by the Karhunen–Loève expansion (3.1.11) with $M = 1$, $\xi_1 \propto \mathscr{U}(-1,1)$, where the expectation of K is modeled by $\bar{K} = 1.9 \cdot (0.95 - x/200)$ and the eigenvalue λ_1 and eigenfunction g_1 are known analytically for the exponential covariance kernel

$$V_K(x,y) = \exp(-|x-y|/20)$$

according to Appendix A. Multiplied with an appropriate scaling factor, the permeability function satisfies

$$K(x,\omega) \in [8.88\cdot 10^{-13}, 2.92\cdot 10^{-12}].$$

The exact solution is given by

$$u(t,x,\omega) = -1.1 + 0.01x + 10^{-3}t(x^2 - 100)\exp(\xi_1(\omega)),$$

and we select the function f and corresponding Dirichlet boundary conditions such that $u(1,x,\omega)$ solves (3.2.2) when performing a single time step with $\tau = 1$. Note

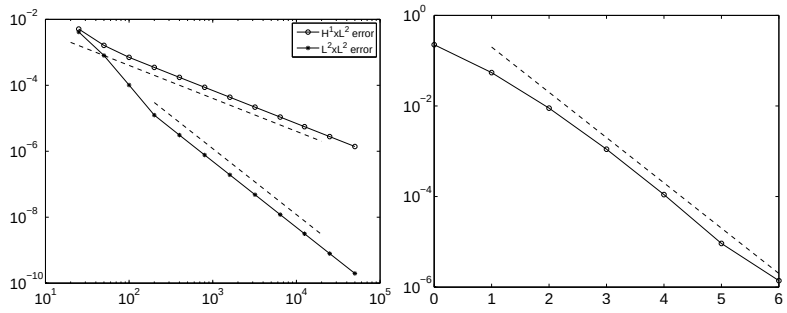


Figure 3.3: Discretization error over the number of spatial unknowns N_J for decreasing mesh size h_J and fixed polynomial degree $P_0 = 6$ (left) and over increasing polynomial degree $P_0 = 0, \ldots, 6$, for a fixed grid with $N_J = 50\,000$ unknowns (right).

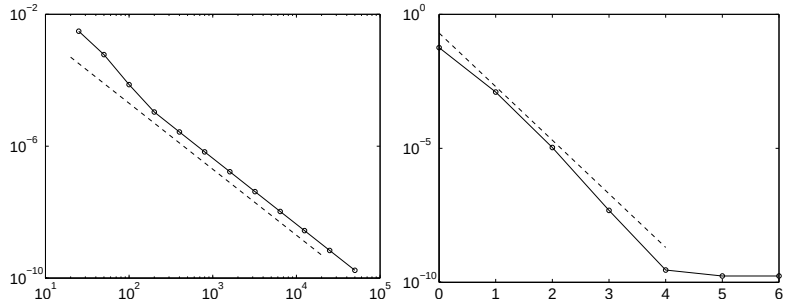


Figure 3.4: L^2 error of the expectation value over the number of spatial unknowns N_J for decreasing mesh size h_J and fixed polynomial degree $P_0 = 6$ (left) and over increasing polynomial degree $P_0 = 0, \ldots, 6$, for a fixed grid with $N_J = 50\,000$ unknowns (right).

that u is smooth in x and ξ_1. We discretize the problem as described in Section 3.3. In spatial direction, we use an equidistant grid with N_J nodes and mesh size $h_J = 20/(N_J - 1)$, and in $\Omega^{(M)} = [-1, 1]$, we take a PC basis of one-dimensional Legendre polynomials (see Appendix B.2) with maximal degree P_0.

We start with the question of the dependence of the discretization error on the mesh size h_J and on the polynomial degree P_0. First, we fix the polynomial degree $P_0 = 6$ and decrease the mesh size h_J. As shown in the left picture of Figure 3.3, the error $\|u - u_J^P\|_{1,0}$ (upper solid line with markers $\circ$) now decreases with order $\mathcal{O}(h_j)$ (upper dashed line) while the L^2 error $\|u - u_J^P\|_{0,0}$ (lower solid line with markers $*$) even behaves like $\mathcal{O}(h_J^2)$ (lower dashed line). In the other setting on a fixed spatial grid with $N_J = 50\,000$, the right picture of Figure 3.3 shows an exponential decay of the error $\|u - u_J^P\|_{1,0}$ (solid line) of order $\mathcal{O}(\beta^{P_r+1})$ with $\beta = 0.1$ (dashed line).

Let us perform the same test scenario in order to investigate the error of the expectation value $\mathbb{E}[u]$. For fixed $P_0 = 6$ and decreasing mesh size h_J, the left picture of Figure 3.4 indicates that the error $\|\mathbb{E}[u] - \mathbb{E}[u_J^P]\|_{L^2(D)}$ behaves like $\mathcal{O}(h_J^2)$ (dashed line), while we obtain an exponential decay of order $\mathcal{O}(\beta^{2P_r+2})$ with $\beta = 0.1$ (dashed line) for a fixed spatial grid with $N_J = 50\,000$ and increasing polynomial degree $P_0 = 0, \ldots, 6$ in the right picture. Note that for $P_0 > 4$ a larger PC basis no longer reduces the overall error since the spatial error dominates the term $\|\mathbb{E}[u] - \mathbb{E}[u_J^P]\|_{L^2(D)}$.

These experimental results suggest that the theoretical error estimates in (3.4.2)

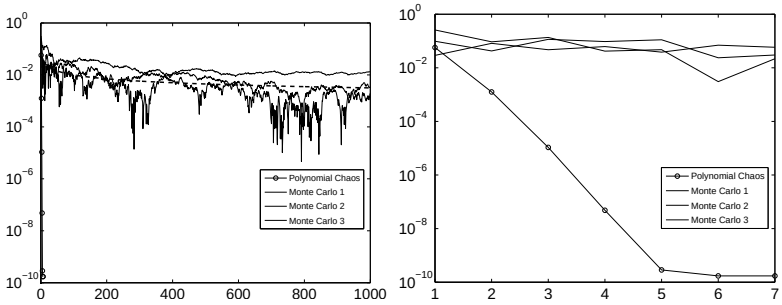


Figure 3.5: L^2 error of the expectation values over the number N_{MC} of deterministic solves: Monte Carlo runs (solid lines), $\mathcal{O}(1/\sqrt{N_{\mathrm{MC}}})$ (dashed line), and polynomial chaos (solid line with markers $\circ$).

and (3.4.3) in Theorem 3.4.1 for the linear case can possibly be extended to our case of the stochastic Richards equation.

The last investigation is dedicated to the comparison between the polynomial chaos and the Monte Carlo method. The Monte Carlo samples were drawn in $\Omega^{(M)}$, i.e. after KL approximation, as will be described in Subsection 4.1.3. We now start three Monte Carlo runs, each of them consisting of N_{MC} samples, and approximate the expectation $\mathbb{E}[u]$ by the sample mean $\bar{u}_J$ from (4.1.27). The error $\|\mathbb{E}[u]-\bar{u}_J\|_{L^2(D)}$ is shown in Figure 3.5 for all three Monte Carlo runs, where the left picture shows the results for $N_{\mathrm{MC}} = 1000$ and the right picture displays the same zoomed to the interval $[1, 7]$. The dashed line in the left picture remarks the expected $\mathcal{O}(1/\sqrt{N_{\mathrm{MC}}})$ behavior as will be deduced in Subsection 4.1.3. Moreover, we plot again the error $\|\mathbb{E}[u]-\mathbb{E}[u_J^P]\|_{L^2(D)}$ from Figure 3.4 but now over the number of (stochastic) degrees of freedom. This comparison assumes that each (stochastic) degree of freedom, which corresponds to a PC basis function, is equivalent to one deterministic solve of the space-discretized Richards equation. As we will show in Section 4.2, this assumption is true for all PC bases mentioned in Subsection 3.3.1 (but not for the classical PC basis $\{\Psi_k^c\}$ for $M > 1$). Figure 3.5 displays a much higher efficiency of the polynomial chaos approach in comparison with Monte Carlo with a factor of about 10^8 for $P_0 = 4$. This is mainly a consequence of the smoothness of u, which is exploited by polynomial chaos and not by Monte Carlo, and of the low stochastic dimension $M = 1$.

3.4.2 Results for two space dimensions

We turn to a two-dimensional domain $D \subset \mathbb{R}^2$ and consider the Richards equation (1.1.11) without gravity, i.e. $\nabla z = 0$. This does not affect the generality of our computations since this term only contributes to the right-hand side in the time-discrete weak formulation (3.2.2).

Take $\lambda = 0.694$ in the Brooks–Corey functions, which corresponds to a sandy soil. Moreover, set $\theta_m = 0.0458$, $\theta_M = 1$, $p_b = -1$, which yields $u_c \approx -1.32446$ by (1.1.14), and the function $H(u)$ according to (1.1.15). Furthermore, choose $D = (-1, 1) \times (-1, 1)$ and a lognormal permeability with exponential covariance. More precisely, let $K = 3.9248 \cdot 10^{-12} \exp(\tilde{K})$, where $\tilde{K}$ is approximated by a Karhunen–Loève expansion

$$\tilde{K} = \mathbb{E}[\tilde{K}] + \sum_{r=1}^{M} \sqrt{\lambda_r} g_r(x) \xi_r(\omega)$$

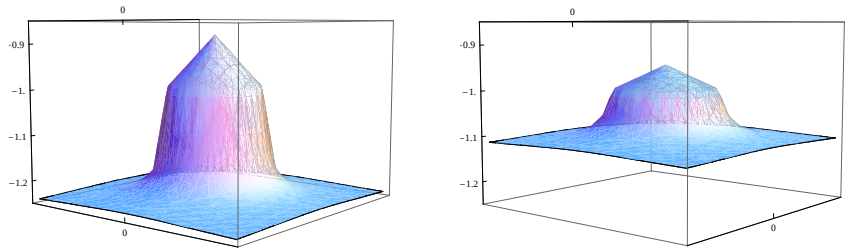


Figure 3.6: The exact solution $u(1, x, \omega)$ for $(\xi_1(\omega), \xi_2(\omega)) = (0, 0)$ (left) and $(\xi_1(\omega), \xi_2(\omega)) = (0.5, -2.6)$ (right).

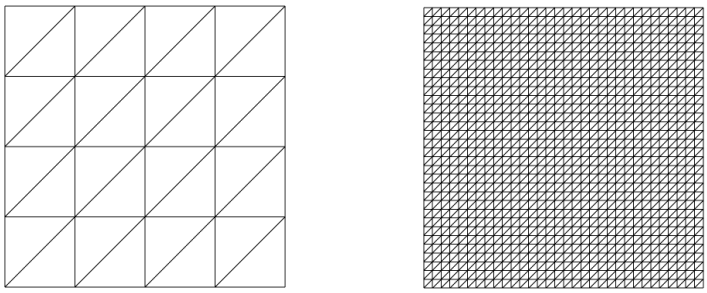


Figure 3.7: The triangulations for $\jmath = 1$ and $\jmath = 4$.

with $M = 2$ and $\xi_1, \xi_2 \propto \mathcal{N}(0, 1)$, cf. Remark 3.1.27. The eigenvalues λ_r and eigenfunctions g_r are known analytically for the separable exponential covariance kernel

$$V_{\tilde{K}}(x, y) = \exp(-|x - y|_1/40)$$

according to Appendix A. We define $\mathbb{E}[\tilde{K}] = 1.4 \cdot (1 - 0.1|x|_2^2)$ and can compute that the permeability function satisfies

$$\mathbb{P}\left(K(x, \omega) \in [2.54 \cdot 10^{-12}, 1.22 \cdot 10^{-11}]\right) \geq 0.9946.$$

We define the exact solution as

$$u(t, x, \omega) = \begin{cases} -1 + 0.05(r - |x|)\cdot \\ \left(\xi_1^4(\omega) - (3 - \xi_1(\omega))^2 - 3\xi_1(\omega) + 15\right)\cos^2(\xi_2(\omega)\pi/2), & |x| < r \\ -1 - \left(2|\xi_1(\omega) + \xi_2(\omega)| + 4\right)^{-1} + \left(100\left(|x| - 0.16 - 0.2t\right)\right)^{-1}, & |x| \geq r \end{cases}$$

with $r = r(t, \xi_1, \xi_2) = 2\left(0.1 + 0.1t + 0.01|\xi_1 + \xi_2|\right)$ and select the function f and corresponding stochastic Dirichlet boundary functions such that $u(1, x, \omega)$ solves (3.2.2) when performing a single time step with $\tau = 1$. The solution $u(1, x, \omega)$ is illustrated in Figure 3.6 at two different realizations. Note that r determines the boundary between the saturated and the unsaturated regime and that it varies with ω. The solution u is not differentiable at this boundary and beyond this boundary the pressure is decreasing extremely rapidly, which is a typical behavior for such problems, cf. the computations in [18].

The problem is discretized as described in Section 3.3. In spatial direction, we use a sequence of triangulations $\mathcal{T}_\jmath$ with mesh size $h_\jmath = 2^{-\jmath}$ as obtained by successive

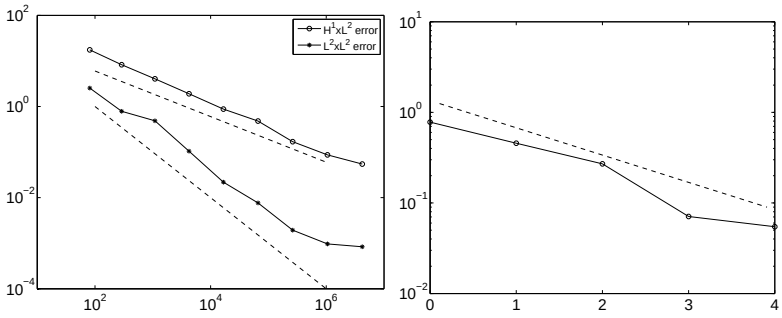


Figure 3.8: Discretization error over the number of spatial unknowns $N_\jmath$ for decreasing mesh size $h_\jmath$, $\jmath = 2, \dots, 10$, and fixed polynomial degree $P_r = 4$ (left) and over increasing polynomial degree $P_r = 0, \dots, 4$, for fixed mesh size $h = h_{10}$ (right).

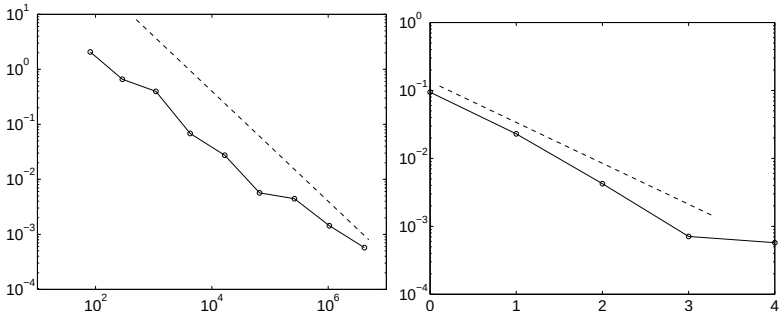


Figure 3.9: L^2 error of the expectation value over the number of spatial unknowns $N_\jmath$ for decreasing mesh size $h_\jmath$, $\jmath = 2, \dots, 10$, and fixed polynomial degree $P_r = 4$ (left) and over increasing polynomial degree $P_r = 0, \dots, 4$, for fixed mesh size $h = h_{10}$ (right).

uniform refinement of an initial triangulation $\mathcal{T}_0$ which comes from one uniform refinement step applied to a partition of D into two congruent triangles, see Figure 3.7. In $\Omega^{(M)} = \mathbb{R}^2$ equipped with the density functions of standard normal distributions, we take a tensor product PC basis with maximal degree P_1 and P_2, where the polynomials Ψ_k from (3.3.5) are products of one-dimensional Hermite polynomials, see Appendix B.1.

We start again with the question of how the discretization error depends on the spatial mesh size $h_\jmath$ and on the polynomial degree P_r. As in the previous example, the left picture of Figure 3.8 shows that for fixed polynomial degree $P_1 = P_2 = 4$ and decreasing mesh size $h_\jmath$, $\jmath = 2, \dots, 10$, the error $\|u - u_\jmath^P\|_{1,0}$ (upper solid line with markers $\circ$) decreases with order $\mathcal{O}(h_\jmath)$ (upper dashed line) while the L^2 error $\|u - u_\jmath^P\|_{0,0}$ (lower solid line with markers $*$) behaves like $\mathcal{O}(h_\jmath^2)$ (lower dashed line). Now, we fix the mesh size $h = h_{10}$ and increase the polynomial degree $P_r = 0, \dots, 4$ for $r = 1, 2$, where we always take $P_1 = P_2$. Here, the right picture of Figure 3.8 shows an exponential decay of the error $\|u - u_\jmath^P\|_{1,0}$. The dashed auxiliary line in this picture indicates the order $\mathcal{O}(\beta^{P_r+1})$ for $\beta = 0.5$.

The error of the expectation value $\mathbb{E}[u]$ is investigated in a similar way. As shown in the left picture of Figure 3.9, the error $\|\mathbb{E}[u] - \mathbb{E}[u_\jmath^P]\|_{L^2(D)}$ behaves like $\mathcal{O}(h_\jmath^2)$ (dashed line) for decreasing mesh size $h_\jmath$ and fixed $P_1 = P_2 = 4$, while the right picture indicates an exponential decay of order $\mathcal{O}(\beta^{2P_r+2})$ with $\beta = 0.5$ (dashed line) for fixed $h = h_{10}$. For $P_1 = P_2 \geq 3$, the spatial error starts to dominate

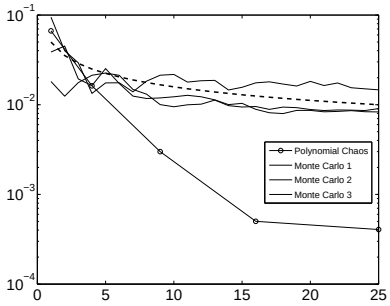


Figure 3.10: L^2 error of the expectation values over the number N_{MC} of deterministic solves: Monte Carlo runs (solid lines), $\mathcal{O}(1/\sqrt{N_{\mathrm{MC}}})$ (dashed line), and polynomial chaos (solid line with markers $\circ$).

again. These experimental results also suggest that the discretization error for the stochastic Richards equation has similar estimates as given in Theorem 3.4.1 for the linear case.

Finally, let us compare the polynomial chaos approach with the Monte Carlo method. We fix the mesh size $h = h_{10}$ and start three runs of the Monte Carlo method. The errors of the sample mean (4.1.27) as approximation of $\mathbb{E}[u]$ in $\|\cdot\|_{L^2(D)}$ norm over the number N_{MC} of deterministic solves are shown as upper solid lines in Figure 3.10, where the dashed line remarks the expected $\mathcal{O}(1/\sqrt{N_{\mathrm{MC}}})$ behavior as will be deduced in Subsection 4.1.3. We compare it with the error of the polynomial chaos approach (lower line with markers $\circ$), where we assume again that each stochastic degree of freedom is equivalent to one deterministic solve. Similar to the example with $d = 1$, we observe a much higher efficiency of the polynomial chaos approach. Here, the factor is about 20 for $P_1 = P_2 = 3$, which is however worse than the corresponding result in Subsection 3.4.1 for two reasons: first, the spatial error prevents a higher accuracy, and secondly, the convergence rate of the polynomial chaos method depends on the dimension M, which is not the case for Monte Carlo. The results demonstrate again the advantage of the polynomial chaos approach.

Chapter 4

Numerical solution

In this chapter, we present our approach for the numerical treatment of the discrete minimization problem

$$u_J^P \in \mathcal{K}_J^P : \ \mathcal{J}(u_J^P) + \phi_J^P(u_J^P) \leq \mathcal{J}(v) + \phi_J^P(v) \quad \forall v \in \mathcal{K}_J^P \tag{4.0.1}$$

introduced in (3.3.49), where $\mathcal{J}$, ϕ_J^P and $\mathcal{K}_J^P$ are defined in (3.2.21), (3.3.33) and (3.3.30), respectively. We start in Section 4.1 by giving an overview over the most commonly used methods for stochastic partial differential equations which feature the stochastic input noise in a parameterizable form as the Karhunen–Loève expansion in Section 3.1. Our new approach is presented in Section 4.2. It is based on a successive minimization in direction of the nodal basis functions in form of a Block Gauß–Seidel method, where, in each block, an inner minimization in stochastic direction is performed in accordance with the form of the stochastic basis functions. The outer iteration in spatial direction is extended to a monotone multigrid approach in order to obtain a more efficient method, which is done in Section 4.3. Moreover, we present some first numerical results concerning convergence rates for the previously described methods. We conclude this chapter by describing some post-processing steps for the calculation of moments and probabilities from the computed solution in Section 4.4.

4.1 Common methods for SPDEs

As in Remark 3.2.14, we assume that Φ and H are continuous and defined on the whole real line and that, for sake of notation, $u_D \equiv 0$. In that case, the discrete version of problem (3.2.28) reads

$$u_J^P \in \mathcal{S}_J^D \otimes \mathcal{Z}^P : a(u_J^P, v) - \ell(v) + \sum_{\pi \in \mathcal{Q}^P} \sum_{p \in \mathcal{N}_J} H(u_J^P(p,\pi)) v(p,\pi) h_p \eta_\pi = 0 \quad \forall v \in \mathcal{S}_J^D \otimes \mathcal{Z}^P \tag{4.1.1}$$

according to Proposition 3.3.11 and Remark 3.3.13. Recall that the permeability function K in the definition (3.2.4) of the bilinear form $a(\cdot,\cdot)$ is approximated by a truncated KL expansion of the form (3.1.11).

In this section, we discuss the application of the most commonly used methods in this context to (4.1.1), viz. the stochastic Galerkin approach, the stochastic collocation approach and the Monte Carlo approach. By doing this, we also set some notation for the following sections.

4.1.1 Stochastic Galerkin method

The stochastic Galerkin approach has gained much attention in the last few years since the work of Ghanem and Spanos [47], who called it "stochastic finite elements method". In the broader sense, this term denotes the discretization in Section 3.3, which employs a standard (finite element) approximation in space and an approximation with global polynomials (PC) in the stochastic domain. We, however, will denote by this term the method of projecting the terms on the PC basis in order to compute their coefficients. This is the standard method for all kind of linear stochastic PDEs (cf. [11, 29, 32, 38, 64, 75, 101, 113]) and stochastic ODEs (cf. [107, 114]), but is also used for a variety of nonlinear problems (cf. [63, 73, 82]).

Denoting by $\{s_i^{(j)}\} = \{s_{p_i}^{(j)}\} \subset \Lambda_j^D$ the nodal basis in $\mathcal{S}_j^D$ with cardinality $N_j = |\Lambda_j^D|$ and by $\{\Psi_k\}$ the PC basis in $\mathcal{Z}^P$, the solution u_j^P of (4.1.1) has the representation

$$u_j^P(x, \xi(\omega)) = \sum_{i=1}^{N_j} \sum_{k=0}^{P} u_{ik} s_i^{(j)}(x) \Psi_k(\xi(\omega)), \tag{4.1.2}$$

and we aim at computing the coefficients u_{ik}. Therefor, we insert (4.1.2) and test functions of the form

$$v(x, \xi(\omega)) = s_j^{(j)}(x) \Psi_l(\xi(\omega))$$

into the weak formulation (4.1.1).

For the bilinear form $a(\cdot, \cdot)$, we obtain

$$\begin{aligned}
a(u_j^P, v) &= \tau_n \mathbb{E}\left[\int_D K \nabla u_j^P \nabla v \,\mathrm{d}x\right] \\
&= \tau_n \mathbb{E}\left[\int_D \left(\bar{K}(x) + \sum_{r=1}^{M} \sqrt{\lambda_r} g_r(x) \xi_r\right) \right. \\
&\qquad \left. \nabla \left(\sum_{i=1}^{N_j} \sum_{k=0}^{P} u_{ik} s_i^{(j)}(x) \Psi_k(\xi)\right) \nabla \left(s_j^{(j)}(x) \Psi_l(\xi)\right) \,\mathrm{d}x\right] \qquad (4.1.3) \\
&= \tau_n \sum_{i=1}^{N_j} \sum_{k=0}^{P} u_{ik} \left(\left(\int_D \bar{K}(x) \nabla s_i^{(j)}(x) \nabla s_j^{(j)}(x) \,\mathrm{d}x\right) \mathbb{E}[\Psi_k \Psi_l] \right. \\
&\qquad \left. + \sum_{r=1}^{M} \sqrt{\lambda_r} \left(\int_D g_r(x) \nabla s_i^{(j)}(x) \nabla s_j^{(j)}(x) \,\mathrm{d}x\right) \mathbb{E}[\xi_r \Psi_k \Psi_l] \right),
\end{aligned}$$

and the right-hand side reads

$$\begin{aligned}
\ell(v) &= \mathbb{E}\left[\int_D H(u_{n-1}) v \,\mathrm{d}x\right] + \tau_n \mathbb{E}\left[\int_D K \, kr(H(u_{n-1})) \rho g e_z \nabla v \,\mathrm{d}x\right] \\
&= \int_\Omega \int_D H\left(u_{n-1}(x, \xi)\right) s_j^{(j)}(x) \Psi_l(\xi) \,\mathrm{d}x \,\mathrm{d}\mathbb{P} \quad + \qquad (4.1.4) \\
&\tau_n \int_\Omega \int_D \left(\bar{K}(x) + \sum_{r=1}^{M} \sqrt{\lambda_r} g_r(x) \xi_r\right) kr\left(H\left(u_{n-1}(x, \xi)\right)\right) \rho g e_z \nabla s_j^{(j)}(x) \Psi_l(\xi) \,\mathrm{d}x \,\mathrm{d}\mathbb{P},
\end{aligned}$$

where the argument $u_{n-1}(x, \xi)$ has the form

$$u_{n-1}(x, \xi(\omega)) = \sum_{i=1}^{N_j} \sum_{k=0}^{P} (u_{n-1})_{ik} s_i^{(j)}(x) \Psi_k(\xi(\omega)).$$

Remark 4.1.1. Since the function H in (4.1.4) couples the variables x and ω, we need a simultaneous quadrature in D and Ω in order to calculate $\ell(v)$. We compute the spatial integral separately on each triangle $\mathrm{t} \in \mathcal{T}_j$ and apply there a usual quadrature scheme (see [25, Section 25]), where, at each quadrature point, a Gaussian quadrature with points $\mathcal{Q}^P$ is employed for the integral over Ω (see Subsection 3.3.1).

Remark 4.1.2. For the treatment of Neumann boundary conditions, the approach is still working. If $\ell(v)$ contains further terms, e.g.

$$\mathbb{E}\left[\int_D f(x,\omega)v(x,\omega)\,\mathrm{d}x\right] \quad \text{or} \quad \mathbb{E}\left[\int_{\Gamma_N} f_N(x,\omega)v(x,\omega)\,\mathrm{d}\sigma\right],$$

the procedure is just the same as in (4.1.4).

Looking at (4.1.3), we note the benefit arising from the tensor product structure of $H^1(D) \otimes L^2(\Omega)$ and from the property of the KL expansion of separating the functions g_r depending on $x \in D$ and ξ_r depending on $\omega \in \Omega$. The spatial integrals and the expectation values can therefore be calculated independently of each other. This gives rise to the block matrix $\mathsf{A} \in \mathbb{R}^{N_J(P+1)\times N_J(P+1)}$ with

$$\mathsf{A} = ([\mathsf{A}_{ij}]_{kl})_{i,j,k,l} \quad \text{for } i,\, j = 1,\dots,N_J \text{ and } k,\, l = 0,\dots,P, \tag{4.1.5}$$

which consists of N_J^2 blocks $\mathsf{A}_{ij} \in \mathbb{R}^{(P+1)\times(P+1)}$. The entry $[\mathsf{A}_{ij}]_{kl}$ is given by

$$\begin{aligned}[\mathsf{A}_{ij}]_{kl} = \tau_n \Bigg(&\left(\int_D \bar{K}(x)\nabla s_i^{(J)}(x)\nabla s_j^{(J)}(x)\,\mathrm{d}x\right)\mathbb{E}[\Psi_k\Psi_l] \\ &+ \sum_{r=1}^{M}\sqrt{\lambda_r}\left(\int_D g_r(x)\nabla s_i^{(J)}(x)\nabla s_j^{(J)}(x)\,\mathrm{d}x\right)\mathbb{E}[\xi_r\Psi_k\Psi_l]\Bigg),\end{aligned} \tag{4.1.6}$$

where the spatial integrals are approximated by means of a quadrature formula. The stochastic integrals, however, are known exactly: on the one hand, it holds $\mathbb{E}[\Psi_k\Psi_l] = \delta_{kl}$ due to (3.3.7), on the other hand, we have by (3.1.25)

$$\begin{aligned}\mathbb{E}[\xi_r\Psi_k\Psi_l] &= \int_\Omega \xi_r(\omega)\Psi_k(\xi(\omega))\Psi_l(\xi(\omega))\,\mathrm{d}\mathbb{P}(\omega) \\ &= \int_{\Omega_M}\cdots\int_{\Omega_1} y_r\Psi_k(y_1,\dots,y_M)\Psi_l(y_1,\dots,y_M)\mathrm{pdf}_1(y_1)\,\mathrm{d}y_1\cdots\mathrm{pdf}_M(y_M)\,\mathrm{d}y_M \\ &= \left(\int_{\Omega_r} y_r\psi_k^r(y_r)\psi_l^r(y_r)\mathrm{pdf}_r(y_r)\,\mathrm{d}y_r\right)\prod_{\substack{s=1,\dots,M\\ s\neq r}}\int_{\Omega_s}\psi_k^s(y_s)\psi_l^s(y_s)\mathrm{pdf}_s(y_s)\,\mathrm{d}y_s\end{aligned}$$

with $\Psi_k = \prod_{s=1}^M \psi_k^s$ according to (3.3.5). The integrals over Ω_s for $s \neq r$ are again either zero or one, whereas the integral over Ω_r in the parentheses can be calculated by an exact Gaussian quadrature with points $\mathcal{Q}^P$ (the integrand is a polynomial of degree at most $2P_r + 1$). Note that it is sufficient to generate only once the one-dimensional integrals and to store them for the rest of the time.

Remark 4.1.3. A short way to represent the stiffness matrix A is to use the matrix Kronecker product. The matrix A is then written shortly as

$$\mathsf{A} = \mathsf{E}_0 \otimes \mathsf{F}_0 + \sum_{r=1}^{M} \mathsf{E}_r \otimes \mathsf{F}_r,$$

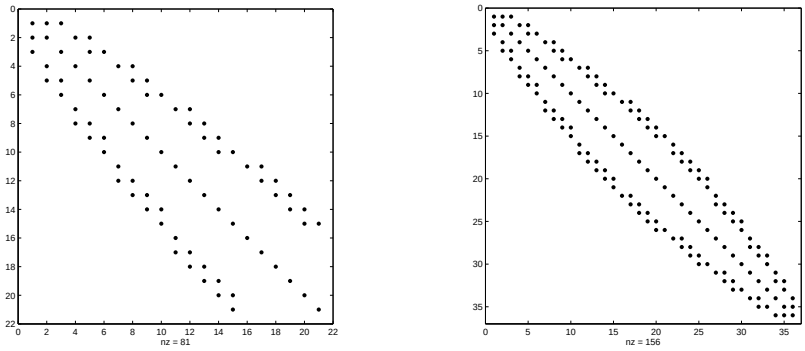


Figure 4.1: Sparsity pattern of each nonzero matrix block A_{ij} for $P_0 = P_r = 4$ in case of classical PC (left, $P + 1 = 21$) and tensor product PC (right, $P + 1 = 36$).

where the matrices $\mathsf{E}_r \in \mathbb{R}^{N_j \times N_j}$ are built up by the spatial integrals and the matrices $\mathsf{F}_r \in \mathbb{R}^{(P+1)\times(P+1)}$ are built up by the stochastic integrals for all $0 \leq r \leq M$. This underlines the aforementioned fact that both integrals can be calculated independently of each other. This has the advantage that in case of basis changes in $\mathcal{S}_j$, for example due to a restriction or prolongation in multigrid approaches or due to adaptive refinements of the grid, it is only necessary to update the matrices E_r and to rebuild A by the Kronecker product; and vice versa for changes in $\mathcal{Z}^P$.

Definition 4.1.4. We denote by $a_j^P(\cdot,\cdot)$ the approximation of the bilinear form $a(\cdot,\cdot)$ by quadrature in the spatial domain D as described above and by $\ell_j^P(\cdot)$ the approximation of the linear form $\ell(\cdot)$ by quadrature as explained in Remark 4.1.1.

These considerations provide the sparsity structure of the stiffness matrix A. Regarding A just consisting of the block matrices $(\mathsf{A}_{ij})_{i,j}$, it possesses the same pattern as in the deterministic finite element context, i.e.

$$\mathsf{A}_{ij} \neq 0 \Rightarrow p_i,\, p_j \in \mathrm{t} \text{ for a } \mathrm{t} \in \mathcal{T}_j.$$

The pattern arising from the expectation values is always the same within each matrix block $\mathsf{A}_{ij} \neq 0$, but it is depending on the used PC basis functions. For the biorthogonal basis $\{\Psi_k\} = \{\Psi_k^b\}$, each block is diagonal. For $\{\Psi_k\} = \{\Psi_k^c\}$ or $\{\Psi_k\} = \{\Psi_k^t\}$, we have sparse blocks; a typical pattern is displayed in Figure 4.1.

Remark 4.1.5. In Remark 3.1.20, we discussed the case of non-independent ξ_r. The suggestion of applying a nonlinear transformation G to a Gaussian expansion as in (3.1.20) causes difficulties. Since the functions $g_{Y,r}$ and ξ_r are coupled via G, one cannot separate the spatial and stochastic integrals as in (4.1.3). A simultaneous quadrature over D and Ω is necessary, and the diagonal or sparse structure of the blocks A_{ij} is destroyed.

For the representation of the variables ξ_r in a PC expansion as in (3.1.21), the separation of the integrals persists, but the resulting matrix blocks are dense (cf. the sparsity patterns depicted in [32]), since expectation values $\mathbb{E}[\Psi_k \Psi_l \Psi_m]$ instead of $\mathbb{E}[\xi_r \Psi_k \Psi_l]$ are obtained in (4.1.6).

Furthermore, we define the block vectors $\mathsf{u} \in \mathbb{R}^{N_j(P+1)}$ and $\mathsf{b} \in \mathbb{R}^{N_j(P+1)}$. Both

vectors consist of N_j blocks $\mathsf{u}_i \in \mathbb{R}^{P+1}$ and $\mathsf{b}_i \in \mathbb{R}^{P+1}$, respectively, according to

$$\mathsf{u} = \begin{pmatrix} \vdots \\ \hline \mathsf{u}_i \\ \hline \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ \hline u_{i0} \\ \vdots \\ \hline u_{iP} \\ \hline \vdots \end{pmatrix} \quad \text{and} \quad \mathsf{b} = \begin{pmatrix} \vdots \\ \hline \mathsf{b}_i \\ \hline \vdots \end{pmatrix}, \tag{4.1.7}$$

where the entries of b are

$$[\mathsf{b}_i]_k = \ell_j^P(s_i^{(j)}\Psi_k). \tag{4.1.8}$$

In the linear case, i.e. for $H \equiv 0$, the problem (4.1.1) can be written as linear system

$$\mathsf{u} \in \mathbb{R}^{N_j(P+1)} : \quad \mathsf{A}\mathsf{u} = \mathsf{b}. \tag{4.1.9}$$

The matrix A is obviously symmetric. Moreover, it is positive definite, if the approximate bilinear form $a_j^P(\cdot,\cdot)$ is coercive, i.e. if there exists a $c > 0$ such that

$$a_j^P(v,v) \geq c\|v\|_{1,0}^2 \quad \forall v \in \mathcal{S}_j \otimes \mathcal{Z}^P. \tag{4.1.10}$$

This follows by [25, Theorem 27.1] from the coercivity condition (3.2.7) if the quadrature scheme used in Definition 4.1.4 is exact for constant functions. Regarding the discretization error in the linear case, there exist theoretical estimates from [11], which we already cited in Theorem 3.4.1.

We return to the nonlinear case $H \not\equiv 0$. Applying the stochastic Galerkin approach to the sum in (4.1.1), we obtain

$$\begin{aligned} &\sum_{\pi\in\mathcal{Q}^P}\sum_{p\in\mathcal{N}_j} H(u_j^P(p,\pi))v(p,\pi)h_p\eta_\pi \\ &\qquad = \sum_{\pi\in\mathcal{Q}^P}\sum_{p\in\mathcal{N}_j} H\left(\sum_{i=1}^{N_j}\sum_{k=0}^{P} u_{ik}s_i^{(j)}(p)\Psi_k(\pi)\right) s_j^{(j)}(p)\Psi_l(\pi)h_p\eta_\pi \\ &\qquad\qquad = \sum_{\pi\in\mathcal{Q}^P} H\left(\sum_{k=0}^{P} u_{p_jk}\Psi_k(\pi)\right)\Psi_l(\pi)h_{p_j}\eta_\pi \end{aligned} \tag{4.1.11}$$

for all $j = 1,\dots,N_j$ and $l = 0,\dots,P$. This means that the nonlinear function H couples all entries of the vector block u_j, and it is far from obvious how to compute the coefficients u_{ik} of the solution. In Section 4.2, we present an approach to overcome this difficulty.

Remark 4.1.6. Even if H is a "nice" function like a polynomial of low order, we obtain expressions which are hard to deal with; e.g. for $H(z) = z^2$ the term (4.1.11) reads after reordering

$$\sum_{\pi\in\mathcal{Q}^P}\sum_{k=0}^{P}\sum_{m=0}^{P} u_{p_jk}u_{p_jm}\Psi_k(\pi)\Psi_m(\pi)\Psi_l(\pi)h_{p_j}\eta_\pi.$$

The sum $\sum_{\pi\in\mathcal{Q}^P}\Psi_k(\pi)\Psi_m(\pi)\Psi_l(\pi)\eta_\pi$ as approximation of $\mathbb{E}[\Psi_k\Psi_l\Psi_m]$ is in general not zero for $k \neq m$ and the Gaussian quadrature is no longer exact. Debusschere et al. [30] propose a reprojection onto $\mathcal{Z}^P$ with accepting an approximation error for operations on PC functions, but it is clear that this approach is limited (cf. [81] for an illustrative example).

4.1.2 Stochastic collocation

An alternative to the stochastic Galerkin approach in Subsection 4.1.1 is the stochastic collocation approach, which became popular in the mid of the first decade in this century by the works of Xiu and Hesthaven [112] and Babuška et al. [13], although its idea was already used earlier, e.g. as "Nonintrusive spectral projection" in [74]. It combines the advantages of Monte Carlo methods and the stochastic Galerkin method. On the one hand, it shows similar convergence results as the latter having its high resolution by exploiting the interpolation properties of multivariate polynomials. On the other hand, it is straightforward to implement like Monte Carlo algorithms, since it requires only the solution of corresponding deterministic problems at each collocation point. In this subsection, we describe the method and give a short overview over known convergence results with an emphasis on the choice of collocation points.

We imply again the assumptions on H as in Remark 3.2.14 which led to the variational equality (3.2.28). In view of (3.1.25), we can regard this equality as a problem in D in dependence of an M-dimensional parameter $y \in \Omega^{(M)}$ and consider the solution u as a function $u : \Omega^{(M)} \to H^1_{\Gamma_D}(D)$, where we use the notation $u(y)$ to highlight this dependence. Equation (3.2.28) is then equivalent to the problem of finding a function $u : \Omega^{(M)} \to H^1_{\Gamma_D}(D)$ such that

$$\partial_v \phi^y(u(y)) + a^y(u(y), v) - \ell^y(v) = 0 \quad \forall v \in H^1_{\Gamma_D}(D), \quad \text{a.e. in } \Omega^{(M)} \tag{4.1.12}$$

with

$$\phi^y(u(y)) := \int_D \Phi(u(y))\,\mathrm{d}x \quad \text{and} \quad \partial_v \phi^y(u(y)) = \int_D H(u(y))v\,\mathrm{d}x, \tag{4.1.13}$$

$$a^y(u(y), v) := \tau_n \int_D K(\cdot, y)\nabla u(y)\nabla v\,\mathrm{d}x, \tag{4.1.14}$$

$$\ell^y(v) := \int_D H(u_{n-1}(y))v\,\mathrm{d}x + \tau_n \int_D K(\cdot, y)kr(H(u_{n-1}(y)))\rho g e_z \nabla v\,\mathrm{d}x. \tag{4.1.15}$$

We first introduce a semi-discretization in D with finite elements as in Subsection 3.3.2 and obtain by projecting equation (4.1.12) onto the subspace $\mathcal{S}^D_J$ and approximating the functional ϕ^y analogously to (3.3.33) and (3.3.56) an approximation $u_J : \Omega^{(M)} \to \mathcal{S}^D_J$ for each $y \in \Omega^{(M)}$ as the solution of

$$\sum_{p \in \mathcal{N}_J} H(u_J(p, y))v(p)h_p + a^y(u_J(y), v) - \ell^y(v) = 0 \quad \forall v \in \mathcal{S}^D_J. \tag{4.1.16}$$

The next step consists in choosing a collocation point set $\mathcal{C} = \{y^{(k)}\}_{k=1,\dots,N_\mathcal{C}}$ in $\Omega^{(M)}$. This is the crucial part of this approach and we will specify it below. We now solve the problem (4.1.16) in each collocation point $y = y^{(k)}$ and denote the solution by $u_J(\cdot, y^{(k)}) \in \mathcal{S}^D_J$. Note that the whole problem is naturally decoupled in this approach and that we work with $N_\mathcal{C}$ deterministic problems.

The last step is to take the Lagrange polynomials $\mathcal{L}_{y^{(k)}}$ for each collocation point and obtain the approximation

$$u^\mathcal{C}_J(x, y) = \sum_{k=1}^{N_\mathcal{C}} u_J(x, y^{(k)})\mathcal{L}_{y^{(k)}}(y) \tag{4.1.17}$$

as a function in $\mathcal{S}^D_J \otimes \operatorname{span}\{\mathcal{L}_{y^{(1)}}, \dots, \mathcal{L}_{y^{(N_\mathcal{C})}}\}$.

It remains to specify the set $\mathcal{C}$. The minimal requirement is that the Lagrange interpolation is feasible, i.e. that there exists a unique polynomial p such that $p(y_k) = f(y_k)$, $1 \le k \le N_{\mathcal{C}}$, for any $f : \Omega^{(M)} \to \mathbb{R}$. This property is called *poisedness of* $\mathcal{C}$ and is nontrivial in the multivariate case $M > 1$ (see [94] and the references therein). The first natural idea is to use univariate point sets $\mathcal{C}_r$ with $N_{\mathcal{C}_r}$ collocation points in each dimension and to build tensor product polynomials. This is done by first associating to a vector of indices $(k_1, \dots, k_M)$ a global index

$$k = k_1 + N_{\mathcal{C}_1}(k_2 - 1) + N_{\mathcal{C}_1} N_{\mathcal{C}_2}(k_3 - 1) + \dots$$

for a point $y^{(k)} = (y^{(1,k_1)}, \dots, y^{(M,k_M)}) \in \Omega^{(M)}$ and calculating $\mathcal{L}_{y^{(k)}}$ as the product

$$\mathcal{L}_{y^{(k)}}(y) = \prod_{r=1}^{M} \mathcal{L}_{r,k_r}(y_r)$$

of one-dimensional Lagrange polynomials $\mathcal{L}_{r,j}$ with $\mathcal{L}_{r,j}(y^{(r,l)}) = \delta_{jl}$ for $j, l = 1, \dots, N_{\mathcal{C}_r}$ representing a basis of the space $\mathrm{Pol}^{N_{\mathcal{C}_r}-1}(\Omega_r)$. An obvious choice for the collocation points is then taking the zeros of the corresponding orthogonal polynomials. With $N_{\mathcal{C}_r} = P_r + 1$ for all $1 \le r \le M$, we arrive at

$$\mathrm{span}\{\mathcal{L}_{y^{(1)}}, \dots, \mathcal{L}_{y^{(N_{\mathcal{C}})}}\} = \mathcal{Z}^P \tag{4.1.18}$$

with $\mathcal{Z}^P$ from (3.3.14). In this way, this stochastic collocation method can be seen as stochastic Galerkin method with biorthogonal polynomials $\{\Psi_k^b\}$, which was first discovered in [13]. A drawback of this idea consists again in the "curse of dimensionality" because of

$$N_{\mathcal{C}} = \prod_{r=1}^{M} N_{\mathcal{C}_r}, \tag{4.1.19}$$

see columns with label $|\{\Psi_k^t\}|$ in Table 3.3. This initiated the investigation of sparse grid schemes like Stroud cubature (used in [112]) or constructions by the Smolyak algorithm (examined in [87] and [112]).

We cite some known results about the error for the linear case $H \equiv 0$. For tensor product polynomials with the zeros of orthogonal polynomials as collocation points, it is in view of (4.1.18) not surprising that one can prove the following analogue to Theorem 3.4.1.

Theorem 4.1.7 ([13])**.** *Under the assumptions of Theorem 3.4.1 there exist positive constants α_r and C, independent of $\jmath$ and $N_{\mathcal{C}}$, such that*

$$\|u - u_\jmath^{\mathcal{C}}\|_{1,0} \le C \left(h_\jmath + \sum_{r=1}^{M} e^{-\alpha_r P_r} \right). \tag{4.1.20}$$

In case of unbounded Ω_r, the convergence rate in (4.1.20) deteriorates slightly and can be estimated as

$$\|u - u_\jmath^{\mathcal{C}}\|_{1,0} \le C \left(h_\jmath + \sum_{r=1}^{M} \sqrt{P_r} e^{-\alpha_r \sqrt{P_r}} \right)$$

under some further assumptions, see [13] for details.

For sparse grids with a Smolyak formula based on Clenshaw–Curtis abcissas, Nobile et al. [87] proved the following error of the solution $u_\jmath^{\mathrm{Sm}}$.

Theorem 4.1.8. *Let $u \in H_0^1(D) \otimes C^0(\Omega^{(M)})$. Then, under the assumptions of Theorem 3.4.1, there exist positive constants α and C, independent of $\jmath$ and the number N_{Sm} of collocation points, such that*

$$\|u - u_\jmath^{\mathrm{Sm}}\|_{H^1(D)\otimes L^\infty(\Omega^{(M)})} \leq C \left(h_\jmath + N_{\mathrm{Sm}}^{-\frac{\alpha}{1+\log(2M)}} \right). \tag{4.1.21}$$

It thus features only an algebraic instead of an exponential error decay in stochastic direction, but on a distinctly smaller point set than (4.1.19). Further research indicates that sparse grids are preferable for large M, while the tensor product collocation is superior for long correlation lengths and anisotropic products, i.e. if the numbers P_r vary for different $1 \leq r \leq M$. On top of that, both methods clearly outperform the Monte Carlo method. For more details, we refer to [87, Section 5].

To close this subsection, we underline the fact that this collocation approach can be applied in our nonlinear case as seen above and that it can be generalized straightforward to variational inequalities and minimization problems in which we are interested. Indeed, if we define for all $y \in \Omega^{(M)}$ the convex set

$$\mathcal{K}_\jmath(y) := \{v(y) \in \mathcal{S}_\jmath : v(p,y) \geq u_c \, \forall p \in \mathcal{N}_\jmath \wedge v(p,y) = u_D(p,y) \, \forall p \in \mathcal{N}_\jmath^D\} \tag{4.1.22}$$

and the functional

$$\mathcal{J}^y(v) := \frac{1}{2} a^y(v,v) - \ell^y(v) \quad \forall v \in H^1(D) \tag{4.1.23}$$

as well as

$$\phi_\jmath^y(v) := \sum_{p \in \mathcal{N}_\jmath} \Phi(v(p,y)) h_p \quad \forall v(\cdot, y) \in \mathcal{S}_\jmath \tag{4.1.24}$$

as discretization of ϕ^y defined in (4.1.13), then the semidiscretization of (3.2.26) reads

$$u_\jmath : \Omega^{(M)} \to \mathcal{K}_\jmath(y) : \quad \mathcal{J}^y(u_\jmath(y)) + \phi_\jmath^y(u_\jmath(y)) \leq \mathcal{J}^y(v) + \phi_\jmath^y(v) \quad \forall v \in \mathcal{K}_\jmath(y). \tag{4.1.25}$$

For the derivation of (4.1.25), we refer to the considerations in Section 3.3 and also the discretization in the deterministic case in [18, Section 2.5].

We continue this point in Subsection 4.2.2, where we show that our solution method can be regarded in special cases as a collocation.

4.1.3 Monte Carlo method

For sake of completeness, we outline the Monte Carlo method as applied to the stochastic Richards equation and compare it with the methods presented in the last two sections. The method was developed in the 1940s [85] and can be applied in a huge variety of fields due to its simplicity. For its use in hydrological context, we refer as example to [39] and [44] and the references therein.

In contrast to Remark 2.3.7, we now apply this method to problem (2.3.8) after approximation of K by the Karhunen–Loève expansion as done in Section 3.1. Note that the Monte Carlo method does not allow to compute the approximate solution $u_\jmath^P \in \mathcal{S}_\jmath \otimes \mathcal{Z}^P$ of (4.0.1), i.e. a representation as a function on $D \times \Omega^{(M)}$, but only approximations of the moments of u as discrete functions on $\mathcal{S}_\jmath \subset H^1(D)$, i.e. for functions which correspond to $\mathbb{E}[(u_\jmath^P)^n]$ for $n \geq 1$. We will concentrate

in the following on the approximation of $\mathbb{E}[u]$ although the computation of higher moments (as will be defined in Section 4.4) is immediate.

We sample independent and identically distributed (i.i.d.) realizations in the probability space $L^2(\Omega^{(M)}, \mathrm{Bor}(\Omega^{(M)}), \mathbb{P}^{(M)})$ from (3.1.22) and denote the samples by y_n, $n = 1, \dots, N_{\mathrm{MC}}$. Using the discretizations and notations from Subsection 4.1.2, we now have to solve N_{MC} independent problems

$$u_\jmath(y_n) \in \mathcal{K}_\jmath(y_n): \quad \mathcal{J}^{y_n}(u_\jmath(y_n)) + \phi_\jmath^{y_n}(u_\jmath(y_n)) \leq \mathcal{J}^{y_n}(v) + \phi_\jmath^{y_n}(v) \quad \forall v \in \mathcal{K}_\jmath(y_n) \tag{4.1.26}$$

for all $n = 1, \dots, N_{\mathrm{MC}}$. We collect the solutions $u_\jmath(y_n) \in \mathcal{S}_\jmath$ and compute the *sample mean*

$$\bar{u}_\jmath := \frac{1}{N_{\mathrm{MC}}} \sum_{n=1}^{N_{\mathrm{MC}}} u_\jmath(y_n). \tag{4.1.27}$$

By the strong law of large numbers [78, Section 16], the sample mean $\bar{u}_\jmath$ converges almost surely to $\mathbb{E}[u]$ provided that the finite element discretization converges, cf. Remark 2.3.7. This is a result of the splitting

$$\mathbb{E}[u] - \bar{u}_\jmath = (\mathbb{E}[u] - \mathbb{E}[u_\jmath]) + (\mathbb{E}[u_\jmath] - \bar{u}_\jmath) =: E_\jmath + E_{\mathrm{MC}}, \tag{4.1.28}$$

where $u_\jmath(y) \in \mathcal{S}_\jmath$ is the finite element approximation of $u(y) \in \hat{\mathcal{K}}$.

As in the previous subsections, we first consider the linear case with homogeneous Dirichlet boundary conditions and $H \equiv 0$, which leads to the problem of finding a $u : D \times \Omega^{(M)} \to \mathbb{R}$ with

$$u(y) \in H_0^1(D): \quad a^y(u(y), v) = \ell^y(v) \quad \forall v \in H_0^1(D) \tag{4.1.29}$$

for almost all $y \in \Omega^{(M)}$. Babuška et al. [11] proved the following result.

Theorem 4.1.9. *Let u be the solution of (4.1.29) with $\ell^y(v) = \int_D f(y)v \,\mathrm{d}x$ for a continuous and bounded function $f \in L^2(D) \otimes L^2(\Omega)$. Let $K \in L^2(D) \otimes L^2(\Omega)$ with*

$$\mathbb{P}\left(\omega \in \Omega : K(\cdot, \omega) \in C^1(\overline{D}) \quad \text{and} \quad \max_{x \in \overline{D}} |\nabla K(x, \omega)| < C_1\right) = 1$$

for a constant C_1. Assume that there exists a constant $C_2 > 0$, independent of N_{MC} and $\jmath$, such that

$$N_{\mathrm{MC}} \cdot \mathbb{E}\left[\|\mathbb{E}[u_\jmath] - \bar{u}_\jmath\|_{L^2(D)}^2\right] \leq C_2 \quad \text{for all } N_{\mathrm{MC}} \text{ and } \jmath.$$

Then for any given $\varepsilon > 0$ there exists a constant $C > 0$ which is independent of ε, N_{MC} and $\jmath$ such that

$$\|\mathbb{E}[u] - \bar{u}_\jmath\|_{L^2(D)} \leq C\left(h_\jmath^2 + \frac{\varepsilon}{\sqrt{N_{\mathrm{MC}}}}\right) \tag{4.1.30}$$

with a probability greater than $1 - C_2/\varepsilon^2$.

Comparing this result with (3.4.3) in Theorem 3.4.1 clearly indicates that the stochastic Galerkin (as well as the stochastic collocation) approach outperforms the Monte Carlo method for the linear case and moderate dimensions M. For more information, see the discussions and numerical experiments in [12, Sections 8 and 9].

We return to our nonlinear variational inequality and consider the Monte Carlo error E_{MC} from (4.1.28). Note that this error is a random variable because $\bar{u}_J$ is random. Thus, we consider

$$\mathbb{E}[E_{\mathrm{MC}}^2] = \mathbb{E}[u_J]^2 - \frac{2}{N_{\mathrm{MC}}}\mathbb{E}[u_J]\mathbb{E}\left[\sum_{n=1}^{N_{\mathrm{MC}}} u_J(y_n)\right] + \frac{1}{N_{\mathrm{MC}}^2}\mathbb{E}\left[\left(\sum_{n=1}^{N_{\mathrm{MC}}} u_J(y_n)\right)^2\right].$$

We exploit that the realizations are i.i.d. and deduce

$$\begin{aligned}\mathbb{E}[E_{\mathrm{MC}}^2] &= \mathbb{E}[u_J]^2 - 2\mathbb{E}[u_J]^2 + \frac{1}{N_{\mathrm{MC}}^2}\left(\left(\sum_{n=1}^{N_{\mathrm{MC}}} \mathbb{E}[u_J]\right)^2 + \sum_{n=1}^{N_{\mathrm{MC}}}\left(\mathbb{E}[u_J^2] - \mathbb{E}[u_J]^2\right)\right)\\ &= \frac{N_{\mathrm{MC}}}{N_{\mathrm{MC}}^2}\left(\mathbb{E}[u_J^2] - \mathbb{E}[u_J]^2\right) = \frac{1}{N_{\mathrm{MC}}}\operatorname{Var}[u_J],\end{aligned}$$

which means that the expected L^2 error

$$N_{\mathrm{MC}} \cdot \mathbb{E}\left[\|E_{\mathrm{MC}}\|_{L^2(D)}^2\right] = \mathbb{E}\left[\|u_J\|_{L^2(D)}^2\right] - \|\mathbb{E}[u_J]\|_{L^2(D)}^2$$

of the Monte Carlo method is only depending on the variance of u_J.

For further analysis, we need the theorem of Berry–Esseen [35, Section XVI.5].

Theorem 4.1.10. *Let $X_1, X_2, \ldots$ be i.i.d. random variables with finite moments $|\mathbb{E}[X_1]| < \infty$, $\sigma^2 := \mathbb{E}\left[(X_1 - \mathbb{E}[X_1])^2\right] \in (0,\infty)$ and $\bar{\mu}_3 := \mathbb{E}\left[|X_1 - \mathbb{E}[X_1]|^3\right] < \infty$ on $(\Omega, \mathcal{E}, \mathbb{P})$. Denote by*

$$\Phi_{0,1} : \ x \mapsto \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-t^2/2}\,\mathrm{d}t$$

the cumulative distribution function of the standard normal distribution $\mathcal{N}(0,1)$ and consider the random variable

$$S_n^* := \frac{X_1 + \ldots + X_n - n\mathbb{E}[X_1]}{\sigma\sqrt{n}}.$$

Then, the convergence of the distribution of S_n^* to $\mathcal{N}(0,1)$ according to the central limit theorem can be estimated by

$$\sup_{x\in\mathbb{R}} |\mathbb{P}(S_n^* \le x) - \Phi_{0,1}(x)| \le \frac{0.8\bar{\mu}_3}{\sigma^3\sqrt{n}}. \tag{4.1.31}$$

We apply this theorem to our problem and detect that the right-hand side of (4.1.31) is nearly zero if

$$N_{\mathrm{MC}} \gg \frac{\mathbb{E}\left[|u_J - \mathbb{E}[u_J]|^3\right]^2}{\mathbb{E}\left[(u_J - \mathbb{E}[u_J])^2\right]^3}.$$

In this case, one obtains

$$\mathbb{P}^{(M)}\left(|\mathbb{E}[u_J] - \bar{u}_J| \le x_0\frac{\sigma_J}{\sqrt{N_{\mathrm{MC}}}}\right) \approx \Phi_{0,1}(x_0) - \Phi_{0,1}(-x_0) = 2\Phi_{0,1}(x_0) - 1 \tag{4.1.32}$$

for an arbitrary $x_0 > 0$ with $\sigma_J^2 := \mathbb{E}\left[(u_J - \mathbb{E}[u_J])^2\right]$. By this way, we constructed a confidence interval with level x_0 for the estimation of the function $\mathbb{E}[u_J]$. Moreover, by replacing σ_J^2 with the sample variance

$$s_J^2 := \frac{1}{N_{\mathrm{MC}} - 1}\sum_{n=1}^{N_{\mathrm{MC}}} (u_J(y_n) - \bar{u}_J)^2,$$

we obtain a termination condition for the Monte Carlo method. Finally note that (4.1.32) is only depending on the variance and on the number N_{MC} of Monte Carlo iterations and not on the dimension M of the sample space.

4.2 Nonlinear Block Gauß–Seidel and minimization

In this section, we present the main ingredients of our method to solve the discrete convex minimization problem (4.0.1). An outer iteration employs a division of the space $\mathcal{S}_J$ and is a successive minimization in direction of the single nodal basis functions, which gives rise to a nonlinear Block Gauß–Seidel relaxation and is a straight generalization of the method for the deterministic Richards equation in [18]. In contrast to the deterministic approach, we do not arrive at one-dimensional problems but at a minimization within the space $\mathcal{Z}^P$. We examine several possibilities to solve the inner minimization. In case of tensor product PC functions $\{\Psi_k\} = \{\Psi_k^t\}$, we present a transformation of the inner minimization, which results in decoupled one-dimensional minimization problems, and we show its connection to stochastic collocation. We emphasize the nature of this approach which turns out to be robust with respect to the slope of H and which remains feasible in case of piecewise smooth Φ and in the limit cases (1.1.18) and (1.1.20).

4.2.1 Nonlinear Block Gauß–Seidel

We start with a nonlinear Block Gauß–Seidel method in $\mathcal{S}_J \otimes \mathcal{Z}^P$. This method will work as a smoother for the multigrid method which will be presented in Section 4.3. As defined in Subsection 3.3.2, we look at nodes $p \in \mathcal{N}_J$, where $\mathcal{N}_J$ has the cardinality $N_J = |\mathcal{N}_J|$, and regard the corresponding nodal basis functions $s_p^{(J)} \in \Lambda_J$. Then, we introduce for $i = 1, \dots, N_J$ the splitting

$$\mathcal{S}_J \otimes \mathcal{Z}^P = \sum_{i=1}^{N_J} V_i \quad \text{with } V_i = \operatorname{span}\{s_{p_i}^{(J)}\} \otimes \mathcal{Z}^P \tag{4.2.1}$$

of $\mathcal{S}_J \otimes \mathcal{Z}^P$ into N_J subspaces of dimension $P+1$. Analogously, we define for $i = 1, \dots, N_J$ the splitting

$$\mathcal{K}_J^P = \sum_{i=1}^{N_J} \mathcal{K}_{J,i}^P$$

of the convex set $\mathcal{K}_J^P$ into N_J subsets

$$\mathcal{K}_{J,i}^P := \mathcal{K}_J^P \cap V_i.$$

Observe that this division of $\mathcal{K}_J^P$ is only possible due to the special structure of this set and the fact that

$$s_{p_i}^{(J)}(p_j) = \delta_{ij}.$$

Moreover, the element $v \in \mathcal{K}_J^P$ can be written as $v(x, \xi(\omega)) = \sum_{i=1}^{N_J} v_i(x, \xi(\omega))$ with $v_i \in \mathcal{K}_{J,i}^P$, and the functional ϕ_J^P is decoupled in spatial direction with

$$\phi_J^P(v) = \sum_{i=1}^{N_J} \sum_{\pi \in \mathcal{Q}^P} \Phi\left(v_i(p_i, \pi)\right) h_{p_i} \eta_\pi = \sum_{i=1}^{N_J} \sum_{\pi \in \mathcal{Q}^P} \Phi\left(\sum_{k=0}^{P} v_{ik} \Psi_k(\pi)\right) h_{p_i} \eta_\pi \tag{4.2.2}$$

according to (3.3.33) and (3.3.34).

Then, starting with a given iterate $w_0^\nu = (u_J^P)^\nu \in \mathcal{K}_J^P$, we compute a sequence of intermediate iterates $w_i^\nu = w_{i-1}^\nu + \bar{v}_i^\nu$, $i = 1, \dots, N_J$, by solving a sequence of convex

minimization problems of finding corrections

$$\begin{aligned}\bar{v}_i^\nu \in V_i \ \text{ with } \ w_{i-1}^\nu + \bar{v}_i^\nu \in \mathcal{K}_J^P : \ & \mathcal{J}(w_{i-1}^\nu + \bar{v}_i^\nu) + \phi_J^P(w_{i-1}^\nu + \bar{v}_i^\nu) \\ & \leq \mathcal{J}(w_{i-1}^\nu + v) + \phi_J^P(w_{i-1}^\nu + v) \quad \forall v \in V_i \ \text{ with } \ w_{i-1}^\nu + v \in \mathcal{K}_J^P \end{aligned} \tag{4.2.3}$$

and define the next iterate by

$$\mathcal{M}_J^P\left((u_J^P)^\nu\right) := (u_J^P)^{\nu+1} = w_{N_J}^\nu = (u_J^P)^\nu + \sum_{i=1}^{N_J} \bar{v}_i^\nu. \tag{4.2.4}$$

By construction, we have monotonically decreasing energy

$$\mathcal{J}(w_i^\nu) + \phi_J^P(w_i^\nu) \leq \mathcal{J}(w_{i-1}^\nu) + \phi_J^P(w_{i-1}^\nu). \tag{4.2.5}$$

In light of Theorem 3.3.10, each subproblem (4.2.3) is uniquely solvable, which means that equality holds in (4.2.5) if, and only if, $w_i^\nu = w_{i-1}^\nu$. This leads to

$$\mathcal{J}(\mathcal{M}_J^P(w)) + \phi_J^P(\mathcal{M}_J^P(w)) = \mathcal{J}(w) + \phi_J^P(w) \quad \Leftrightarrow \quad \mathcal{M}_J^P(w) = w \tag{4.2.6}$$

as a characterization of the fixed points of $\mathcal{M}_J^P$. Furthermore, we assume that the iteration operator $\mathcal{M}_J^P$ is continuous, i.e.

$$w_n^\nu \to w \quad \Rightarrow \quad \mathcal{M}_J^P(w_n^\nu) \to \mathcal{M}_J^P(w) \tag{4.2.7}$$

as $n \to \infty$. In Remark 4.2.6, it will be shown that condition (4.2.7) is actually satisfied by the iteration operator. Finally, we note that the minimization problems (4.2.3) can be rewritten as the variational inequalities

$$\begin{aligned}\bar{v}_i^\nu \in V_i \ \text{ with } \ w_{i-1}^\nu + \bar{v}_i^\nu \in \mathcal{K}_J^P : \ & a(w_{i-1}^\nu + \bar{v}_i^\nu, v - \bar{v}_i^\nu) - \ell(v - \bar{v}_i^\nu) \\ & + \phi_J^P(w_{i-1}^\nu + v) - \phi_J^P(w_{i-1}^\nu + \bar{v}_i^\nu) \geq 0 \quad \forall v \in V_i \ \text{ with } \ w_{i-1}^\nu + v \in \mathcal{K}_J^P \end{aligned} \tag{4.2.8}$$

according to Proposition 3.3.11.

The following global convergence theorem is a generalization of a result for the case of one-dimensional V_i which can be found in [48, Theorem V.3.1] or [68, Theorem 2.1]. We adapt the proof of the latter to our block case. However, we remark that the proof in [48] goes without condition (4.2.7) and can be generalized to the block case in the same way as it is done here.

Theorem 4.2.1. *We assume the conditions given in Theorem 3.3.10. Then, for any initial iterate $(u_J^P)^0 \in \mathcal{K}_J^P$, the sequence of iterates $\left((u_J^P)^\nu\right)_{\nu \geq 0}$ provided by the nonlinear Block Gauß–Seidel relaxation method (4.2.4) converges to the solution u_J^P of the discrete problem (4.0.1).*

Proof. For sake of notation, we use the abbreviation $F_J^P = \mathcal{J} + \phi_J^P$. The proof is then divided into three steps. First, we show that the sequence of iterates $\left((u_J^P)^\nu\right)_{\nu \geq 0}$ is bounded. This can be seen by contradiction, since $\|(u_J^P)^\nu\| \to \infty$ would imply $F_J^P\left((u_J^P)^\nu\right) \to \infty$ by (3.3.50), but it holds

$$F_J^P\left((u_J^P)^\nu\right) \leq F_J^P\left((u_J^P)^0\right) < \infty$$

for all $\nu \geq 0$.

Since $\left((u_J^P)^\nu\right)_{\nu\geq 0}$ is bounded in the finite-dimensional space $\mathcal{S}_J \otimes \mathcal{Z}^P$, there exists a convergent subsequence $\left((u_J^P)^{\nu_k}\right)_{k\geq 0}$ with

$$(u_J^P)^{\nu_k} \to u^* \in \mathcal{S}_J \otimes \mathcal{Z}^P$$

as $k \to \infty$. We have $u^* \in \mathcal{K}_J^P$, because $(u_J^P)^{\nu_k} \in \mathcal{K}_J^P$ for all $k \geq 0$ and $\mathcal{K}_J^P$ is closed. We now prove as second step that u^* must be a fixed point of $\mathcal{M}_J^P$. The monotonicity (4.2.5) implies

$$F_J^P\left((u_J^P)^{\nu_{k+1}}\right) \leq F_J^P\left((u_J^P)^{\nu_k+1}\right) = F_J^P\left(\mathcal{M}_J^P\left((u_J^P)^{\nu_k}\right)\right) \leq F_J^P\left((u_J^P)^{\nu_k}\right)$$

for all $k \geq 0$. The continuity of $\mathcal{M}_J^P$ and F_J^P on $\mathcal{K}_J^P$ yields

$$F_J^P\left(\mathcal{M}_J^P(u^*)\right) = F_J^P(u^*),$$

and we conclude from (4.2.6) that $\mathcal{M}_J^P(u^*) = u^*$.

Finally, we show that all fixed points u^* of $\mathcal{M}_J^P$ are equal to the solution u_J^P of the problem (4.0.1). In this step, we make use of the structure of our functional ϕ_J^P. In view of (4.2.8), each local variational inequality can be written as

$$\begin{aligned}\bar{v}_i^\nu \in V_i \text{ with } w_{i-1}^\nu + \bar{v}_i^\nu \in \mathcal{K}_J^P: \; & a(w_{i-1}^\nu + \bar{v}_i^\nu, v - \bar{v}_i^\nu) - \ell(v - \bar{v}_i^\nu) \\ + \sum_{\pi\in\mathcal{Q}^P} \Phi\left((w_{i-1}^\nu + v)(p_i,\pi)\right) h_{p_i}\eta_\pi & - \sum_{\pi\in\mathcal{Q}^P} \Phi\left((w_{i-1}^\nu + \bar{v}_i^\nu)(p_i,\pi)\right) h_{p_i}\eta_\pi \geq 0 \\ & \forall v \in V_i \text{ with } w_{i-1}^\nu + v \in \mathcal{K}_J^P \quad (4.2.9)\end{aligned}$$

for all $i = 1, \ldots, N_J$. For the fixed point u^*, all local corrections $\bar{v}_i^\nu$ of u^* must be zero, which means that the inequality in (4.2.9) reads

$$a(u^*, v) - \ell(v) + \sum_{\pi\in\mathcal{Q}^P} \Phi\left((u^* + v)(p_i,\pi)\right) h_{p_i}\eta_\pi - \sum_{\pi\in\mathcal{Q}^P} \Phi\left(u^*(p_i,\pi)\right) h_{p_i}\eta_\pi \geq 0. \quad (4.2.10)$$

Now consider some arbitrary but fixed $\tilde{v} \in \mathcal{S}_J \otimes \mathcal{Z}^P$ with representation (3.3.31) and insert the interpolation

$$v = I_{V_i}(\tilde{v} - u^*) := \sum_{k=0}^P \tilde{v}_{ik} s_{p_i}^{(J)} \Psi_k - \sum_{k=0}^P u_{ik}^* s_{p_i}^{(J)} \Psi_k$$

in (4.2.10) to obtain

$$\begin{aligned}& a(u^*, I_{V_i}(\tilde{v} - u^*)) - \ell(I_{V_i}(\tilde{v} - u^*)) + \\ & \sum_{\pi\in\mathcal{Q}^P} \Phi\left(\sum_{k=0}^P \tilde{v}_{ik} s_{p_i}^{(J)}(p_i)\Psi_k(\pi)\right) h_{p_i}\eta_\pi - \sum_{\pi\in\mathcal{Q}^P} \Phi\left(\sum_{k=0}^P u_{ik}^* s_{p_i}^{(J)}(p_i)\Psi_k(\pi)\right) h_{p_i}\eta_\pi \geq 0.\end{aligned}$$

Adding up all these local inequalities for $i = 1, \ldots, N_J$, this results in

$$a(u^*, \tilde{v} - u^*) - \ell(\tilde{v} - u^*) + \phi_J^P(\tilde{v}) - \phi_J^P(u^*) \geq 0$$

according to (4.2.2). Since $\tilde{v} \in \mathcal{S}_J \otimes \mathcal{Z}^P$ was arbitrary, we see by (3.3.51) that u^* is equal to the unique solution u_J^P.

We have shown that each convergent subsequence of $\left((u_J^P)^\nu\right)_{\nu\geq 0}$ converges to u_J^P. Hence, the whole sequence must converge to u_J^P, and the proof is complete. □

Remark 4.2.2. Note that we did not assert global convergence in Theorem 4.2.1 for all initial iterates $(u_J^P)^0 \in \mathcal{S}_J \otimes \mathcal{Z}^P$. In fact, for $\nu = 0$ and $i = 1$ the functional ϕ_J^P in (4.2.3) is equal to ∞ if the condition $(u_J^P)^0(p_j, \pi) \geq u_c$ is not satisfied at points p_j with $j > i = 1$. However, if one takes successively the local problems (4.2.9) instead of (4.2.3), then one arrives at $(u_J^P)^1 \in \mathcal{K}_J^P$ even for $(u_J^P)^0 \notin \mathcal{K}_J^P$. In this sense, we can define the iteration operator $\mathcal{M}_J^P : \mathcal{S}_J \otimes \mathcal{Z}^P \to \mathcal{K}_J^P$ and obtain global convergence on the whole space $\mathcal{S}_J \otimes \mathcal{Z}^P$.

Remark 4.2.3. In Theorem 4.2.1, we did not distinguish between nodes $p_i \in \mathcal{N}_J^D$ and $p_i \in \mathcal{N}_J \backslash \mathcal{N}_J^D$. Indeed, for $p_i \in \mathcal{N}_J^D$ we set $w_0^\nu(p_i, \cdot) = \mathcal{P}^P(u_D(p_i, \cdot))$ according to Remark 3.3.3 and have all corrections $\bar{v}_i^\nu \in V_i$ equal to zero for $\nu \geq 0$.

Now we turn to the practical realization of the correction problems (4.2.3). Due to Remark 4.2.3, we henceforth only consider points

$$p_i \in \mathcal{N}_J \backslash \mathcal{N}_J^D.$$

We switch to matrix notation with a stiffness matrix A defined in (4.1.5) and block vectors u and b introduced in (4.1.7). Denote by $\mathsf{w} \in \mathbb{R}^{N_J(P+1)}$ and $\bar{\mathsf{v}} \in \mathbb{R}^{(P+1)}$ the coefficient vectors of w_{i-1}^ν and $\bar{v}_i^\nu$, respectively, and by $\mathsf{v} \in \mathbb{R}^{N_J(P+1)}$ the long block vector with $\mathsf{v}_i = \bar{\mathsf{v}}$ and $\mathsf{v}_j = 0$ for $i \neq j$. Then, we can rewrite (4.2.3) as

$$\begin{aligned} \operatorname*{arg\,min}_{\bar{\mathsf{v}} \in \mathbb{R}^{P+1}} \frac{1}{2}(\mathsf{w}+\mathsf{v})^T \mathsf{A}(\mathsf{w}+\mathsf{v}) - (\mathsf{w}+\mathsf{v})^T \mathsf{b} + \sum_{i=1}^{N_J} \sum_{\pi \in \mathcal{Q}^P} \Phi\left(\sum_{k=0}^{P} (w_{ik} + v_{ik}) \Psi_k(\pi) \right) h_{p_i} \eta_\pi \\ \text{subject to } \sum_{i=1}^{N_J} \sum_{k=0}^{P} (w_{ik} + v_{ik}) s_{p_i}^{(J)} \Psi_k \in \mathcal{K}_J^P. \end{aligned}$$

We eliminate all constant terms which do not contribute to the minimization and obtain equivalently with $\bar{\mathsf{A}} := \mathsf{A}_{ii}$

$$\begin{aligned} \operatorname*{arg\,min}_{\bar{\mathsf{v}} \in \mathbb{R}^{P+1}} \frac{1}{2}\bar{\mathsf{v}}^T \bar{\mathsf{A}} \bar{\mathsf{v}} + \bar{\mathsf{v}}^T([\mathsf{A}\mathsf{w}]_i - \mathsf{b}_i) + \sum_{\pi \in \mathcal{Q}^P} \Phi\left(\sum_{k=0}^{P} ([\mathsf{w}_i + \bar{\mathsf{v}}]_k) \Psi_k(\pi) \right) h_{p_i} \eta_\pi \\ \text{subject to } \sum_{k=0}^{P} [\mathsf{w}_i + \bar{\mathsf{v}}]_k s_{p_i}^{(J)} \Psi_k \in \mathcal{K}_{J,i}^P. \quad (4.2.11) \end{aligned}$$

We now introduce the local evaluation matrix $\mathsf{B} \in \mathbb{R}^{\Pi \times (P+1)}$ with entries

$$\mathsf{B}_{\pi k} = \sqrt{\eta_\pi} \Psi_k(\pi) \quad \text{for } \pi \in \mathcal{Q}^P,\ k = 0, \ldots, P. \quad (4.2.12)$$

This matrix will play an important role in the further considerations. For now, it simplifies (4.2.11) in view of the definition of $\mathcal{K}_{J,i}^P$ as

$$\begin{aligned} \operatorname*{arg\,min}_{\bar{\mathsf{v}} \in \mathbb{R}^{P+1}} \frac{1}{2}\bar{\mathsf{v}}^T \bar{\mathsf{A}} \bar{\mathsf{v}} + \bar{\mathsf{v}}^T([\mathsf{A}\mathsf{w}]_i - \mathsf{b}_i) + \sum_{\pi \in \mathcal{Q}^P} \Phi\left(\frac{1}{\sqrt{\eta_\pi}} [\mathsf{B}(\mathsf{w}_i + \bar{\mathsf{v}})]_\pi \right) h_{p_i} \eta_\pi \\ \text{subject to } \frac{1}{\sqrt{\eta_\pi}} [\mathsf{B}(\mathsf{w}_i + \bar{\mathsf{v}})]_\pi \geq u_c \quad \forall \pi \in \mathcal{Q}^P. \quad (4.2.13) \end{aligned}$$

In the following two subsections, we solve this minimization problem depending on the used stochastic basis functions.

4.2.2 Transformation of the tensor product PC

Throughout this subsection, we will concentrate on the tensor product PC basis $\{\Psi_k\} = \{\Psi_k^t\}$ with (3.3.23). Then, the matrix B defined in (4.2.12) is quadratic (we continue however for sake of comprehensibility using indices π and k to distinguish between the quadrature points and the PC modes) and has the following properties.

Proposition 4.2.4. *Let (3.3.23) hold and* $\{\Psi_k\} = \{\Psi_k^t\}$.

a) B *is an orthogonal matrix with* $\mathsf{B}^T\mathsf{B} = \mathsf{B}\mathsf{B}^T = \mathrm{I}$.

b) The rows $\mathsf{c}_\pi \in \mathbb{R}^{P+1}$ *of* B *are the eigenvectors of* $\bar{\mathsf{A}}$, *i.e.* $\bar{\mathsf{A}}\mathsf{c}_\pi = \mu_\pi \mathsf{c}_\pi$.

Proof. The entries of $\mathsf{B}^T\mathsf{B}$ are given by

$$[\mathsf{B}^T\mathsf{B}]_{kl} = \sum_{\pi\in\mathcal{Q}^P} \sqrt{\eta_\pi}\sqrt{\eta_\pi}\Psi_k(\pi)\Psi_l(\pi) = \int_\Omega \Psi_k\Psi_l \,\mathrm{d}\mathbb{P} = \delta_{kl},$$

since the Gaussian quadrature is exact. This yields in particular $\mathsf{B}\mathsf{B}^T = \mathrm{I}$, i.e.

$$[\mathsf{B}\mathsf{B}^T]_{\pi\varsigma} = \sum_{l=0}^{P} \Psi_l(\pi)\Psi_l(\varsigma)\sqrt{\eta_\pi}\sqrt{\eta_\varsigma} = \delta_{\pi\varsigma}. \tag{4.2.14}$$

For part b), we define by $\tilde{K}$ and $\tilde{g}_r$ the values of

$$\tau_n \int_{\mathrm{supp}(s_{p_i}^{(j)})} \bar{K}(x) \left(\nabla s_{p_i}^{(j)}(x)\right)^2 \,\mathrm{d}x$$

and

$$\tau_n\sqrt{\lambda_r} \int_{\mathrm{supp}(s_{p_i}^{(j)})} g_r(x) \left(\nabla s_{p_i}^{(j)}(x)\right)^2 \,\mathrm{d}x$$

for $r = 1,\dots,M$, respectively, approximated by the quadrature formula used in (4.1.6). Note that $\nabla s_{p_i}^{(j)}$ is constant. Then the entries of the diagonal block $\bar{\mathsf{A}} = \mathsf{A}_{ii}$ read

$$\bar{\mathsf{A}}_{kl} = \tilde{K}\mathbb{E}[\Psi_k\Psi_l] + \sum_{r=1}^{M} \tilde{g}_r \mathbb{E}[\xi_r\Psi_k\Psi_l]. \tag{4.2.15}$$

We use again the exactness of the quadrature to obtain

$$\begin{aligned}
[\bar{\mathsf{A}}\mathsf{c}_\pi]_k &= \sum_{l=0}^{P} \tilde{K}\left(\sum_{\varsigma\in\mathcal{Q}^P} \Psi_k(\varsigma)\Psi_l(\varsigma)\eta_\varsigma\right)\sqrt{\eta_\pi}\Psi_l(\pi) \\
&\quad + \sum_{l=0}^{P}\sum_{r=1}^{M} \tilde{g}_r\left(\sum_{\varsigma\in\mathcal{Q}^P} \varsigma\Psi_k(\varsigma)\Psi_l(\varsigma)\eta_\varsigma\right)\sqrt{\eta_\pi}\Psi_l(\pi) \\
&= \sum_{\varsigma\in\mathcal{Q}^P} \tilde{K}\Psi_k(\varsigma)\sqrt{\eta_\varsigma}\left(\sum_{l=0}^{P}\Psi_l(\varsigma)\Psi_l(\pi)\sqrt{\eta_\varsigma}\sqrt{\eta_\pi}\right) \\
&\quad + \sum_{\varsigma\in\mathcal{Q}^P}\sum_{r=1}^{M} \tilde{g}_r\varsigma\Psi_k(\varsigma)\sqrt{\eta_\varsigma}\left(\sum_{l=0}^{P}\Psi_l(\varsigma)\Psi_l(\pi)\sqrt{\eta_\varsigma}\sqrt{\eta_\pi}\right).
\end{aligned}$$

We apply (4.2.14) and get

$$\begin{aligned}[\bar{\mathsf{A}}\mathsf{c}_\pi]_k &= \left(\tilde{K} + \sum_{r=1}^{M} \tilde{g}_r \pi\right) \Psi_k(\pi)\sqrt{\eta_\pi} \\ &= \mu_\pi [\mathsf{c}_\pi]_k,\end{aligned}$$

where μ_π denotes the approximation of $K(p_i, \pi)$ in the parentheses. □

Remark 4.2.5. Proposition 4.2.4 is also valid if we use an exponential KL expansion from (3.1.33). In this case, the block entry $\bar{\mathsf{A}}_{kl}$ from (4.2.15) now reads

$$\bar{\mathsf{A}}_{kl} = \sum_{\varsigma \in \mathcal{Q}^P} \tilde{K}_\varsigma \Psi_k(\varsigma)\Psi_l(\varsigma)\eta_\varsigma,$$

where $\tilde{K}_\varsigma$ is the value of

$$\int_{\mathrm{supp}(s_{p_i}^{(J)})} \exp\left(\bar{K}(x) + \sum_{r=1}^{M} \sqrt{\lambda_r} g_r(x)\varsigma\right) \left(\nabla s_{p_i}^{(J)}(x)\right)^2 \,\mathrm{d}x$$

approximated by the quadrature formula used in (4.1.6). Then, the same calculation as in Proposition 4.2.4 shows $\bar{\mathsf{A}}\mathsf{c}_\pi = \mu_\pi \mathsf{c}_\pi$ with $\mu_\pi = \tilde{K}_\pi$.

Note that

$$\mathsf{B}^{-1} = \mathsf{B}^T \quad \text{and} \quad \mathsf{B}^{-T} = \mathsf{B} \tag{4.2.16}$$

follows from part a) in Proposition 4.2.4. Thus, we have found an eigenvalue decomposition of $\bar{\mathsf{A}} = \mathsf{A}_{ii}$, which reads

$$\mathsf{B}^{-T}\bar{\mathsf{A}}\mathsf{B}^{-1} = \mathsf{B}\bar{\mathsf{A}}\mathsf{B}^T = \mathrm{diag}(\mu_1, \ldots, \mu_\Pi) =: \mathsf{D}_i. \tag{4.2.17}$$

We introduce the new vector $\mathfrak{w} \in \mathbb{R}^\Pi$ which represents the evaluation of the new iterate at the points (p_i, π) for $\pi \in \mathcal{Q}^P$; more precisely, it is

$$\mathfrak{w}_\pi = [\mathsf{B}(\mathsf{w}_i + \bar{\mathsf{v}})]_\pi = \sqrt{\eta_\pi} \sum_{k=0}^{P} [\mathsf{w}_i + \bar{\mathsf{v}}]_k \Psi_k(\pi). \tag{4.2.18}$$

In the new variable $\mathfrak{w}$, the minimization (4.2.13) reads

$$\begin{aligned}\operatorname*{arg\,min}_{\mathfrak{w}\in\mathbb{R}^\Pi} \frac{1}{2}(\mathsf{B}^{-1}\mathfrak{w} - \mathsf{w}_i)^T \bar{\mathsf{A}}(\mathsf{B}^{-1}\mathfrak{w} - \mathsf{w}_i) + (\mathsf{B}^{-1}\mathfrak{w} - \mathsf{w}_i)^T([\mathsf{A}\mathsf{w}]_i - \mathsf{b}_i) + \\ \sum_{\pi\in\mathcal{Q}^P} \Phi\left(\frac{1}{\sqrt{\eta_\pi}}\mathfrak{w}_\pi\right) h_{p_i}\eta_\pi \qquad \text{subject to } \frac{1}{\sqrt{\eta_\pi}}\mathfrak{w}_\pi \geq u_c \quad \forall \pi \in \mathcal{Q}^P.\end{aligned}$$

Again eliminating all constant terms and reordering leads to

$$\begin{aligned}\operatorname*{arg\,min}_{\mathfrak{w}\in\mathbb{R}^\Pi} \frac{1}{2}\mathfrak{w}^T(\mathsf{B}^{-T}\bar{\mathsf{A}}\mathsf{B}^{-1})\mathfrak{w} + \mathfrak{w}^T\mathsf{B}^{-T}([\mathsf{A}\mathsf{w}]_i - \mathsf{b}_i - \bar{\mathsf{A}}\mathsf{w}_i) + \sum_{\pi\in\mathcal{Q}^P} \Phi\left(\frac{1}{\sqrt{\eta_\pi}}\mathfrak{w}_\pi\right) h_{p_i}\eta_\pi \\ \text{subject to } \frac{1}{\sqrt{\eta_\pi}}\mathfrak{w}_\pi \geq u_c \quad \forall \pi \in \mathcal{Q}^P,\end{aligned}$$

which can be rewritten by virtue of (4.2.16) and (4.2.17) as

$$\begin{aligned}\operatorname*{arg\,min}_{\mathfrak{w}\in\mathbb{R}^\Pi} \frac{1}{2}\mathfrak{w}^T\mathsf{D}_i\mathfrak{w} + \mathfrak{w}^T\mathsf{r} + \sum_{\pi\in\mathcal{Q}^P} \Phi\left(\frac{1}{\sqrt{\eta_\pi}}\mathfrak{w}_\pi\right) h_{p_i}\eta_\pi \\ \text{subject to } \frac{1}{\sqrt{\eta_\pi}}\mathfrak{w}_\pi \geq u_c \quad \forall \pi \in \mathcal{Q}^P\end{aligned} \tag{4.2.19}$$

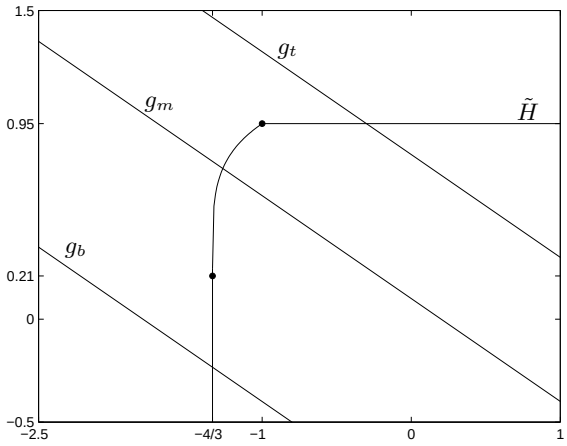


Figure 4.2: Possible intersections of $\tilde{H}$ and $G_{p_i,\pi}$

with

$$\mathsf{r} := \mathsf{B}([\mathsf{Aw}]_i - \mathsf{b}_i - \bar{\mathsf{A}}\mathsf{w}_i). \tag{4.2.20}$$

By (4.2.19), we achieved our aim of decoupling within the block at each spatial node p_i. Recalling the multifunction $\tilde{H} = \partial\Phi$ defined in (3.2.43), the entries of $\mathfrak{w}$ are the solutions of $\Pi = P + 1$ scalar inclusions

$$0 \in \mu_\pi \mathfrak{w}_\pi + \mathsf{r}_\pi + \frac{h_{p_i}\eta_\pi}{\sqrt{\eta_\pi}} \tilde{H}\left(\frac{1}{\sqrt{\eta_\pi}}\mathfrak{w}_\pi\right) \tag{4.2.21}$$

for all $\pi \in \mathcal{Q}^P$ or

$$0 \in \mu_\pi \sqrt{\eta_\pi} y_\pi + \mathsf{r}_\pi + h_{p_i}\sqrt{\eta_\pi}\tilde{H}(y_\pi), \tag{4.2.22}$$

where we used the rescaling

$$y_\pi = \frac{1}{\sqrt{\eta_\pi}}\mathfrak{w}_\pi \tag{4.2.23}$$

and where the constraints in (4.2.19) just reduce to

$$y_\pi \geq u_c \tag{4.2.24}$$

for all $\pi \in \mathcal{Q}^P$. We can interpret y_π as the intersection point of the real linear function

$$G_{p_i,\pi} : x \mapsto -\frac{\mu_\pi}{h_{p_i}} x - \frac{\mathsf{r}_\pi}{h_{p_i}\sqrt{\eta_\pi}}$$

with the multifunction $\tilde{H}$ such that (4.2.22) can be written as

$$y_\pi \in \mathbb{R} : \;\; G_{p_i,\pi}(y_\pi) = \tilde{H}(y_\pi). \tag{4.2.25}$$

Observe that the linear functions $G_{p_i,\pi}$ are strictly decreasing since both μ_π and h_{p_i} are positive numbers. Thus, equation (4.2.25) always has a unique solution.

As an example, we choose H as defined in the setting leading to Figure 1.5 in Section 1.1 using Brooks–Corey parameters. As it can be seen in Figure 4.2, we have to distinguish three cases:

(1) If $G_{p_i,\pi}(u_c) \leq \theta_m$ (see bottom line g_b), then set $y_\pi = u_c$ due to (4.2.24).

(2) If $G_{p_i,\pi}(-1) \geq \theta_M$ (see top line g_t), then

$$y_\pi = -\frac{1}{\mu_\pi \sqrt{\eta_\pi}} \left(\mathsf{r}_\pi + \sqrt{\eta_\pi} h_{p_i} \theta_M \right).$$

(3) Otherwise (see middle line g_m), we have $u_c < y_\pi < -1$. We solve this numerically (up to machine precision) with the bisection method.

Remark 4.2.6. Using (4.2.16), (4.2.18), (4.2.23) and the definition

$$\mathsf{H} := \operatorname{diag}(\eta_1, \ldots, \eta_\Pi)$$

of the scaling matrix $\mathsf{H} \in \mathbb{R}^{\Pi \times \Pi}$, we retransform the vector $\mathsf{y} = (y_1, \ldots, y_\Pi)$ according to

$$\bar{\mathsf{v}} = \mathsf{B}^T \mathsf{H}^{\frac{1}{2}} \mathsf{y} - \mathsf{w}_i.$$

This shows that the correction $\bar{\mathsf{v}}$ depends continuously on w and y. Furthermore, y also depends continuously on w by the definition of r and Proposition 3.2.22, because, using the antiderivative of $G_{p_i,\pi}$, (4.2.25) can be written as a one-dimensional minimization problem

$$\arg\min_y \frac{1}{2} \frac{\mu_\pi}{h_{p_i}} y^2 + \frac{\mathsf{r}_\pi}{h_{p_i} \sqrt{\eta_\pi}} y + \Phi(y).$$

Altogether, $\bar{\mathsf{v}}$ depends continuously on w and the operator $\mathcal{M}_J^P$ thus depends continuously on w_i^ν, as claimed in (4.2.7).

We now resume the collocation idea from Subsection 4.1.2. Taking the zeros of the orthogonal polynomials of order $P_r + 1$ in each dimension $1 \leq r \leq M$ and using the tensor product approach, we have

$$\mathcal{C} = \mathcal{Q}^P \quad \text{with} \quad N_\mathcal{C} = \Pi. \tag{4.2.26}$$

We thus switch the notation for the collocation points from $y^{(k)}$ to π and apply the usual nonlinear Gauß–Seidel method (see [48]) with subspaces $V_i = \operatorname{span}\{s_{p_i}^{(j)}\}$ to the minimization problem (4.1.25). Each correction step analogously to (4.2.3) in order to solve (4.1.25) then reads

$$\begin{aligned} &\bar{v}_i^\nu(\pi) \in V_i \ \text{ with } \ w_{i-1}^\nu(\pi) + \bar{v}_i^\nu(\pi) \in \mathcal{K}_j(\pi) : \\ &\mathcal{J}^\pi(w_{i-1}^\nu(\pi) + \bar{v}_i^\nu(\pi)) + \phi_j^\pi(w_{i-1}^\nu(\pi) + \bar{v}_i^\nu(\pi)) \leq \mathcal{J}^\pi(w_{i-1}^\nu(\pi) + v) + \phi_j^\pi(w_{i-1}^\nu(\pi) + v) \\ &\qquad\qquad \forall v \in V_i \ \text{ with } \ w_{i-1}^\nu(\pi) + v \in \mathcal{K}_j(\pi) \end{aligned} \tag{4.2.27}$$

with intermediate iterates $w_i^\nu(\pi) = w_{i-1}^\nu(\pi) + \bar{v}_i^\nu(\pi)$ and an iteration

$$\mathcal{M}_j^\pi(w_0^\nu(\pi)) = w_{N_j}^\nu(\pi) = w_0^\nu(\pi) + \sum_{i=1}^{N_j} \bar{v}_i^\nu(\pi). \tag{4.2.28}$$

We define the stiffness matrix $\mathsf{A}^\pi = (a_{ij}^\pi)_{i,j} \in \mathbb{R}^{N_j \times N_j}$ with

$$a_{ij}^\pi = a^\pi \left(s_i^{(j)}(x), s_j^{(j)}(x) \right) \tag{4.2.29}$$

and the vector $\mathsf{b}^\pi = (b_i^\pi)_i \in \mathbb{R}^{N_j}$ with

$$b_i^\pi = \ell^\pi \left(s_i^{(j)}(x) \right) \tag{4.2.30}$$

in analogy to (4.1.5) and (4.1.8), respectively. To evaluate the spatial integrals in (4.2.29) and (4.2.30), we use the same quadrature formulas as in (4.1.6) and Remark 4.1.1, respectively. Furthermore, we denote by $\mathsf{w}^\pi = (w_j^\pi)_j \in \mathbb{R}^{N_J}$ and $\mathsf{v}^\pi = (v_j^\pi)_j \in \mathbb{R}^{N_J}$ the coefficient vectors of $w_{i-1}^\nu(\pi)$ and $\bar{v}_i^\nu(\pi)$, respectively. Then, with ideas similar to those which led to (4.2.13) and (4.2.21), the correction step (4.2.27) can be rewritten as a scalar inclusion

$$0 \in a_{ii}^\pi v_i^\pi - b_i^\pi + [\mathsf{A}^\pi \mathsf{w}^\pi]_i + h_{p_i} \tilde{H}(w_i^\pi + v_i^\pi) \tag{4.2.31}$$

for each $\pi \in \mathcal{Q}^P$, see [18, pp. 104–105] for a strict derivation. The constraint has now the form

$$w_i^\pi + v_i^\pi \geq u_c. \tag{4.2.32}$$

It turns out to be convenient to work with a scaled form $\check{\mathsf{B}} \in \mathbb{R}^{\Pi \times (P+1)}$ of the matrix B by taking $\check{\mathsf{B}} := \mathsf{H}^{-\frac{1}{2}}\mathsf{B}$ or, equivalently,

$$\check{\mathsf{B}}_{\pi k} = \Psi_k(\pi).$$

Then, we can state the following result.

Proposition 4.2.7. *With the notations and assumptions of this subsection, one Gauß–Seidel iteration (4.2.28) for all $\pi \in \mathcal{Q}^P = \mathcal{C}$ is equivalent to one Block Gauß–Seidel iteration (4.2.4).*

Proof. We want to reformulate the inclusions in (4.2.21). First, we show

$$\mu_\pi = a_{ii}^\pi. \tag{4.2.33}$$

Using the representation (4.2.15), the orthonormality of the basis $\{\Psi_k\}$, the exactness of the quadrature and the identity (4.2.14), it holds

$$\begin{aligned}
\mu_\pi &= [\mathsf{B}\bar{\mathsf{A}}\mathsf{B}^T]_{\pi\pi} \\
&= \sum_{l=0}^{P}\sum_{k=0}^{P}\left(\tilde{K}\mathbb{E}[\Psi_k\Psi_l] + \sum_{r=1}^{M}\tilde{g}_r\left(\sum_{\varsigma\in\mathcal{Q}^P}\varsigma\Psi_k(\varsigma)\Psi_l(\varsigma)\eta_\varsigma\right)\right)\sqrt{\eta_\pi}\sqrt{\eta_\pi}\Psi_k(\pi)\Psi_l(\pi) \\
&= \tilde{K} + \sum_{r=1}^{M}\tilde{g}_r\sum_{\varsigma\in\mathcal{Q}^P}\varsigma\left(\sum_{k=0}^{P}\Psi_k(\pi)\Psi_k(\varsigma)\sqrt{\eta_\pi}\sqrt{\eta_\varsigma}\right)\left(\sum_{l=0}^{P}\Psi_l(\pi)\Psi_l(\varsigma)\sqrt{\eta_\pi}\sqrt{\eta_\varsigma}\right) \\
&= \tilde{K} + \sum_{r=1}^{M}\tilde{g}_r\pi = a_{ii}^\pi.
\end{aligned}$$

In the same way, one can show

$$[\mathsf{B}\mathsf{b}_i]_\pi = \sqrt{\eta_\pi}b_i^\pi. \tag{4.2.34}$$

Using (4.2.33) and (4.2.34), the scalar inclusion (4.2.21) can be rewritten in view of (4.2.18) and (4.2.20) as

$$0 \in a_{ii}^\pi[\mathsf{B}\bar{\mathsf{v}}]_\pi - \sqrt{\eta_\pi}b_i^\pi + \mu_\pi[\mathsf{B}\mathsf{w}_i]_\pi - [\mathsf{B}\bar{\mathsf{A}}\mathsf{w}_i]_\pi + [\mathsf{B}[\mathsf{A}\mathsf{w}]_i]_\pi + h_{p_i}\sqrt{\eta_\pi}\tilde{H}\left(\left[\check{\mathsf{B}}(\mathsf{w}_i + \bar{\mathsf{v}})\right]_\pi\right). \tag{4.2.35}$$

The third and fourth term on the right-hand side of (4.2.35) cancel, because (4.2.17) provides

$$\mu_\pi[\mathsf{B}\mathsf{w}_i]_\pi = [\mathsf{D}_i\mathsf{B}\mathsf{w}_i]_\pi = [\mathsf{B}\bar{\mathsf{A}}\mathsf{w}_i]_\pi.$$

Finally, observe that the diagonalization in (4.2.17) also works if A_{ij} is not a block on the diagonal. With the same arguments as in the derivation of (4.2.33), we therefore obtain

$$[\mathsf{B}[\mathsf{Aw}]_i]_\pi \;=\; \left[\mathsf{B}\sum_{j=1}^{N_\jmath}\mathsf{A}_{ij}\mathsf{w}_j\right]_\pi \;=\; \left[\sum_{j=1}^{N_\jmath}(\mathsf{BA}_{ij}\mathsf{B}^T)\mathsf{Bw}_j\right]_\pi \;=\; \sum_{j=1}^{N_\jmath} a_{ij}^\pi[\mathsf{Bw}_j]_\pi.$$

Altogether, (4.2.35) can be reformulated as

$$0 \in \sqrt{\eta_\pi}\left(a_{ii}^\pi[\check{\mathsf{B}}\mathsf{v}_i]_\pi - b_i^\pi + \sum_{j=1}^{N_\jmath} a_{ij}^\pi[\check{\mathsf{B}}\mathsf{w}_j]_\pi + h_{p_i}\tilde{H}\left(\left[\check{\mathsf{B}}\mathsf{w}_i\right]_\pi + \left[\check{\mathsf{B}}\bar{\mathsf{v}}\right]_\pi\right)\right). \quad (4.2.36)$$

Comparing (4.2.36) with the inclusion (4.2.31) of the collocation approach, we detect that we have $[\check{\mathsf{B}}\mathsf{v}_i]_\pi = v_i^\pi$ if $[\check{\mathsf{B}}\mathsf{w}_j]_\pi = w_j^\pi$ for all $1 \le j \le N_\jmath$, taking into account that the constraints in (4.2.13) and (4.2.32) also correspond in this case. Since the computation in (4.2.36) is realized for all $\pi \in \mathcal{Q}^P$ independently of each other, we furthermore conclude

$$\check{\mathsf{B}}\mathsf{w}_j = (w_j^\pi)_{\pi\in\mathcal{Q}^P} \quad \forall\, 1 \le j \le N_\jmath \quad \Rightarrow \quad \check{\mathsf{B}}\mathsf{v}_i = (v_i^\pi)_{\pi\in\mathcal{Q}^P}.$$

This means that the intermediate iterates $w_i^\nu(\pi) = w_{i-1}^\nu(\pi) + \bar{v}_i^\nu(\pi)$ for all $\pi \in \mathcal{Q}^P$ from (4.2.27) and the intermediate iterates $w_i^\nu = w_{i-1}^\nu + \bar{v}_i^\nu$ from (4.2.3) correspond via the transformation of its coefficient vectors by $\check{\mathsf{B}}$, if $w_{i-1}^\nu(\pi)$ and w_{i-1}^ν do. Thus, the iterations defined in (4.2.28) for all $\pi \in \mathcal{Q}^P$ and (4.2.4) correspond, as claimed. □

We can use these results for gaining an insight into the relationship between the stochastic Galerkin approach and stochastic collocation. Applying the former to (4.0.1) results in

$$\begin{aligned}\operatorname*{arg\,min}_{\mathsf{v}\in\mathbb{R}^{\bar{N}_\jmath(P+1)}} \frac{1}{2}\mathsf{v}^T\mathsf{Av} - \mathsf{v}^T\mathsf{b} + \sum_{p_i\in\mathcal{N}_\jmath\setminus\mathcal{N}_\jmath^D}\sum_{\pi\in\mathcal{Q}^P}\Phi\left(\frac{1}{\sqrt{\eta_\pi}}\left[\mathsf{Bv}_i\right]_\pi\right)h_{p_i}\eta_\pi \\ \text{subject to } v(p_i,\pi) = \frac{1}{\sqrt{\eta_\pi}}\left[\mathsf{Bv}_i\right]_\pi \ge u_c \quad \forall p_i \in \mathcal{N}_\jmath\setminus\mathcal{N}_\jmath^D\,\forall\pi\in\mathcal{Q}^P, \quad (4.2.37)\end{aligned}$$

see the derivation of (4.2.13), while the latter can be written as

$$\begin{aligned}\operatorname*{arg\,min}_{\mathsf{v}^\pi\in\mathbb{R}^{\bar{N}_\jmath}} \frac{1}{2}(\mathsf{v}^\pi)^T\mathsf{A}^\pi\mathsf{v}^\pi - (\mathsf{v}^\pi)^T\mathsf{b}^\pi + \sum_{p_i\in\mathcal{N}_\jmath\setminus\mathcal{N}_\jmath^D}\Phi\left(v_i^\pi\right)h_{p_i} \\ \text{subject to } v^\pi(p_i) = v_i^\pi \ge u_c \quad \forall p_i\in\mathcal{N}_\jmath\setminus\mathcal{N}_\jmath^D \quad (4.2.38)\end{aligned}$$

for all $\pi \in \mathcal{Q}^P$, where $\bar{N}_\jmath$ is the number of nodes in $\mathcal{N}_\jmath\setminus\mathcal{N}_\jmath^D$. In our setting, both methods are indeed equal.

Theorem 4.2.8. *Consider the convex minimization problem*

$$u \in \mathcal{K}: \;\; \mathcal{J}(u) + \phi(u) \le \mathcal{J}(v) + \phi(v) \quad \forall v \in \mathcal{K} \quad (4.2.39)$$

as in (3.2.26). Let (SG) be the stochastic Galerkin approach for problem (4.2.39) with $\{\Psi_k\} = \{\Psi_k^t\}$ and (3.3.23) on the discretized problem (4.0.1). Let (SC) be the stochastic collocation approach for problem (4.2.39) with (4.2.26) on the discretized problem (4.1.25).

If the same quadrature formulas are used in the computation of the spatial integrals in (4.1.6) and (4.2.29) on the one hand and the same quadrature formulas are used for the spatial integrals in Remark 4.1.1 and in (4.2.30) on the other hand, then (SG) and (SC) are equivalent.

Proof. Let $\underline{\mathsf{B}} \in \mathbb{R}^{\bar{N}_\jmath \Pi \times \bar{N}_\jmath (P+1)}$ be the block diagonal matrix, where the entries $\underline{\mathsf{B}}_{ii} = \mathsf{B} \in \mathbb{R}^{\Pi \times (P+1)}$ are defined in (4.2.12). With the transformation $\mathfrak{v} = \underline{\mathsf{B}}\mathsf{v}$, the minimization (4.2.37) can be written as

$$\begin{aligned} \underset{\mathfrak{v} \in \mathbb{R}^{\bar{N}_\jmath \Pi}}{\arg\min} \frac{1}{2} \mathfrak{v}^T \left(\underline{\mathsf{B}} \mathsf{A} \underline{\mathsf{B}}^T \right) \mathfrak{v} - \mathfrak{v}^T \left(\underline{\mathsf{B}} \mathsf{b} \right) + \sum_{p_i \in \mathcal{N}_\jmath \backslash \mathcal{N}_\jmath^D} \sum_{\pi \in \mathcal{Q}^P} \Phi \left(\frac{1}{\sqrt{\eta_\pi}} \mathfrak{v}_{i\pi} \right) h_{p_i} \eta_\pi \\ \text{subject to } \frac{1}{\sqrt{\eta_\pi}} \mathfrak{v}_{i\pi} \geq u_c \quad \forall p_i \in \mathcal{N}_\jmath \backslash \mathcal{N}_\jmath^D \, \forall \pi \in \mathcal{Q}^P. \end{aligned} \tag{4.2.40}$$

As shown in the proof of Proposition 4.2.7, it holds

$$\left[\left[\underline{\mathsf{B}} \mathsf{A} \underline{\mathsf{B}}^T \right]_{\pi\varsigma} \right]_{ij} = \left[\left[\underline{\mathsf{B}} \mathsf{A} \underline{\mathsf{B}}^T \right]_{ij} \right]_{\pi\varsigma} = a_{ij}^\pi \delta_{\pi\varsigma} \tag{4.2.41}$$

and

$$\left[\left[\underline{\mathsf{B}} \mathsf{b} \right]_\pi \right]_i = \left[\left[\underline{\mathsf{B}} \mathsf{b} \right]_i \right]_\pi = \sqrt{\eta_\pi} b_i^\pi. \tag{4.2.42}$$

Observe that we rearranged the block matrices and block vectors on the left-hand side of (4.2.41) and (4.2.42) having block indices π, ς and inner indices i, j. The rearranged matrix $\underline{\mathsf{B}} \mathsf{A} \underline{\mathsf{B}}^T$ is now block diagonal, and we can thus reformulate (4.2.40) as finding

$$\begin{aligned} \underset{\mathfrak{v}_\pi \in \mathbb{R}^{\bar{N}_\jmath}}{\arg\min} \frac{1}{2} \mathfrak{v}_\pi^T \mathsf{A}^\pi \mathfrak{v}_\pi - \sqrt{\eta_\pi} \mathfrak{v}_\pi^T \mathsf{b}^\pi + \sum_{p_i \in \mathcal{N}_\jmath \backslash \mathcal{N}_\jmath^D} \Phi \left(\frac{1}{\sqrt{\eta_\pi}} \left[\mathfrak{v}_\pi \right]_i \right) h_{p_i} \eta_\pi \\ \text{subject to } \frac{1}{\sqrt{\eta_\pi}} \left[\mathfrak{v}_\pi \right]_i \geq u_c \quad \forall p_i \in \mathcal{N}_\jmath \backslash \mathcal{N}_\jmath^D \end{aligned}$$

for all $\pi \in \mathcal{Q}^P$, independently of each other. We define $\mathsf{v}^\pi = \frac{1}{\sqrt{\eta_\pi}} \mathfrak{v}_\pi$ and obtain

$$\begin{aligned} \underset{\mathsf{v}^\pi \in \mathbb{R}^{\bar{N}_\jmath}}{\arg\min} \eta_\pi \left(\frac{1}{2} (\mathsf{v}^\pi)^T \mathsf{A}^\pi \mathsf{v}^\pi - (\mathsf{v}^\pi)^T \mathsf{b}^\pi + \sum_{p_i \in \mathcal{N}_\jmath \backslash \mathcal{N}_\jmath^D} \Phi \left(v_i^\pi \right) h_{p_i} \right) \\ \text{subject to } v_i^\pi \geq u_c \quad \forall p_i \in \mathcal{N}_\jmath \backslash \mathcal{N}_\jmath^D \end{aligned}$$

for all $\pi \in \mathcal{Q}^P$, which is equivalent to (4.2.38). □

Remark 4.2.9. There are several factors which lead to the result of Theorem 4.2.8. First, the conditions (3.3.23) and (4.2.26) ensure that the number of collocation points and the number of quadrature points and the number of polynomials coincide. Secondly, the approximation $\phi_\jmath^P$ of the convex functional defined in (3.3.33) is based on the quadrature scheme which is used for the collocation. Finally, the approximation $\mathcal{K}_\jmath^P$ from (3.3.30) for the convex set also suits the pointwise view of the collocation approach.

Note that by the considerations of Section 3.2, the equivalence of the stochastic Galerkin and the stochastic collocation approach can be stated for a whole class of stochastic variational inequalities

$$u \in \mathcal{K}: \quad a(u, v-u) - \ell(v-u) + \phi(v) - \phi(u) \geq 0 \quad \forall v \in \mathcal{K}$$

if the convex functional ϕ has the form (3.2.12). In this way, we generalized the result for the linear case from [13, Section 2.1]. The same result can be obtained if the biorthogonal PC basis $\{\Psi_k\} = \{\Psi_k^b\}$ is used directly, cf. Forster and Kornhuber [37].

4.2.3 Block minimization

The approach in the former subsection cannot be applied for the classical PC basis $\{\Psi_k\} = \{\Psi_k^c\}$. The only assertion from Proposition 4.2.4 which is still valid in this case is $\mathsf{B}^T\mathsf{B} = \mathrm{I}$. Due to $P + 1 < \Pi$, the matrix $\mathsf{B} \in \mathbb{R}^{\Pi\times(P+1)}$ is no longer quadratic, and even if one tries to replace the inverse in (4.2.17) by the Moore–Penrose pseudoinverse $\mathsf{B}^+ \in \mathbb{R}^{(P+1)\times\Pi}$ (see [49] for definition), the resulting matrix

$$(\mathsf{B}^+)^T\bar{\mathsf{A}}\mathsf{B}^+ \in \mathbb{R}^{\Pi\times\Pi}$$

is not diagonal, but a dense matrix with rank $P + 1 < \Pi$. Since the basis $\{\Psi_k^c\}$ is smaller than $\{\Psi_k^t\}$ for equal $P_r = P_0$, it is however of interest albeit a decoupling in stochastic direction is not possible. In this subsection, we give a short idea of what are the main challenges one has to deal with in this case.

The starting point is the minimization problem (4.2.13). In terms of convex optimization, it is of the form of finding a $v \in \mathbb{R}^{P+1}$ which fulfills

$$\min \check{\mathcal{J}}(v) \qquad \text{subject to} \quad \check{g}_\pi(v) \le 0, \tag{4.2.43}$$

where $\check{\mathcal{J}}$ is given by

$$\check{\mathcal{J}}(v) := \frac{1}{2}v^T\bar{\mathsf{A}}v + v^T([\mathsf{Aw}]_i - \mathsf{b}_i) + \sum_{\pi\in\mathcal{Q}^P} \Phi\left(\frac{1}{\sqrt{\eta_\pi}}\left[\mathsf{B}(\mathsf{w}_i + v)\right]_\pi\right) h_{p_i}\eta_\pi$$

and the constraints $\check{g}_\pi$ are given by

$$\check{g}_\pi(v) := u_c - \frac{1}{\sqrt{\eta_\pi}}\left[\mathsf{B}(\mathsf{w}_i + v)\right]_\pi = u_c - \sum_{k=0}^{P}(w_{ik} + v_k)\Psi_k(\pi)$$

for all $\pi \in \mathcal{Q}^P$. The constraint functions $\check{g}_\pi$ are linear and the energy function $\check{\mathcal{J}}$ is convex. There exists a variety of methods to solve this problem; however, many optimization algorithms like, e.g., widely used SQP (see [42, Section 5.5]) as generalization of Newton's method are not applicable, since $\check{\mathcal{J}} \notin C^2(\mathbb{R}^{P+1})$.

If the saturation H is continuous, then Φ is differentiable and hence $\check{\mathcal{J}} \in C^1(\mathbb{R}^{P+1})$. In this case, a promising way is to define the *Lagrange function* $\check{L} : \mathbb{R}^{P+1}\times\mathbb{R}^{\Pi} \to \mathbb{R}$ by

$$\check{L}(v, \check{\lambda}) := \check{\mathcal{J}}(v) + \sum_{\pi\in\mathcal{Q}^P} \check{\lambda}_\pi\check{g}_\pi(v) \tag{4.2.44}$$

and try to solve the *Karush–Kuhn–Tucker (KKT) conditions* of (4.2.43) given by

$$\nabla_v\check{L}(v, \check{\lambda}) = 0, \tag{4.2.45}$$

$$\check{\lambda} \ge 0,\ \check{g}(v) \le 0,\ \check{\lambda}^T\check{g}(v) = 0. \tag{4.2.46}$$

This can be done by penalty methods or barrier methods (see [42] and [111]). Note that all these methods have in common that the utilization of the gradient of $\check{L}$ implies that there are $\mathcal{O}(|\mathcal{Q}^P|)$ evaluations of H in each iteration step for solving

(4.2.45)–(4.2.46). Thus, we cannot expect that we can solve the minimization problem (4.2.13) faster by using the classical PC basis $\{\Psi_k\} = \{\Psi_k^c\}$ than if using the bases $\{\Psi_k\} = \{\Psi_k^t\}$ or $\{\Psi_k\} = \{\Psi_k^b\}$.

This is substantiated by another approach recently presented in [52]. We write the minimization problem (4.2.13) as

$$v^* \in \mathbb{R}^{P+1} : \tilde{\mathcal{J}}(v^*) \leq \tilde{\mathcal{J}}(v) \quad \forall v \in \mathbb{R}^{P+1} \tag{4.2.47}$$

with

$$\tilde{\mathcal{J}}(v) := \frac{1}{2} v^T \bar{\mathsf{A}} v + v^T([\mathsf{A}\mathsf{w}]_i - \mathsf{b}_i) + \sum_{\pi \in \mathcal{Q}^P} \Phi\left(\frac{1}{\sqrt{\eta_\pi}} \left[\mathsf{B}(\mathsf{w}_i + v)\right]_\pi\right) h_{p_i} \eta_\pi,$$

where the constraints

$$\frac{1}{\sqrt{\eta_\pi}} \left[\mathsf{B}(\mathsf{w}_i + v)\right]_\pi \geq u_c \tag{4.2.48}$$

are incorporated in Φ by extending the function by $\Phi(u) = +\infty$ for $u < u_c$, cf. Remark 3.2.19. The domain of $\tilde{\mathcal{J}}$ is a polyhedron, since (4.2.48) can be rewritten as

$$\langle \mathsf{c}_\pi, v \rangle \geq u_c \sqrt{\eta_\pi} - \langle \mathsf{c}_\pi, \mathsf{w}_i \rangle \quad \forall \pi \in \mathcal{Q}^P,$$

where c_π are the rows of B as in Proposition 4.2.4. Now, (4.2.47) can be solved in a Gauß–Seidel-like way if one can find a suitable finite set of search directions $\tilde{S} = \{v_1, \ldots, v_S\} \subset \mathbb{R}^{P+1}$ reflecting the shape of $\operatorname{dom} \tilde{\mathcal{J}}$. As shown in [52], it is very costly to find $\tilde{S}$, and this set can be quite large. If, however, we have $\{\Psi_k\} = \{\Psi_k^t\}$ or $\{\Psi_k\} = \{\Psi_k^b\}$, then we can just use the set $\tilde{S} = \{\mathfrak{w}_1, \ldots, \mathfrak{w}_\Pi\} \subset \mathbb{R}^{P+1} = \mathbb{R}^\Pi$ and obtain again the method described in Subsection 4.2.2. In this way, one can see that the tensor product PC bases with $P + 1 = \Pi$ are the most appropriate ones for our problem.

4.3 Monotone multigrid method

In this section, we present monotone multigrid methods with constrained Newton linearization which improve the solution method from Section 4.2. This idea originates from Kornhuber [69] basing upon earlier works [68] and we show that it can be applied in our context. It is presented in our hydrological setting with Brooks–Corey functions and their limit cases in analogy to [18].

4.3.1 Multilevel corrections

With the Block Gauß–Seidel method and the approaches to solve within the single blocks as described in Section 4.2, we developed a method to solve problem (4.0.1). It however turns out to be inefficient in general, since it is well known that even in the linear case with $\phi_j^P \equiv 0$ and no constraints, the convergence rates of the normal Gauß–Seidel method already deteriorate when passing to more and more refined spatial triangulations $\mathcal{T}_j$. This is due to the form of the subspaces V_i in (4.2.1), where only high frequency functions $s_{p_i}^{(j)}$ with small support are involved in spatial direction. Thus, the application of $\mathcal{M}_j^P$ rapidly reduces high frequency contributions of the error $(u_j^P)^\nu - u_j^P$ but hardly affects low frequencies.

A possible remedy is to extend the set of search directions by functions of larger support in an ordered subset

$$R^\nu := (r_1^\nu, \ldots, r_{N_{R^\nu}}^\nu)$$

of $\mathcal{S}_J$ for any $\nu \geq 0$. The first N_J functions are chosen as

$$r_i^\nu = s_{p_i}^{(J)}, \quad i = 1, \ldots, N_J,$$

and represent the fine grid functions, whereas the functions r_i^ν for $i > N_J$ are suitable functions with larger support.

This gives rise to an *extended relaxation method* by performing a successive minimization

$$\begin{aligned} \bar{v}_i^\nu \in V_i^\nu \ \text{ with } \ w_i^\nu := w_{i-1}^\nu + \bar{v}_i^\nu \in \mathcal{K}_J^P : \ & \mathcal{J}(w_{i-1}^\nu + \bar{v}_i^\nu) + \phi_J^P(w_{i-1}^\nu + \bar{v}_i^\nu) \\ \leq \mathcal{J}(w_{i-1}^\nu + v) + \phi_J^P(w_{i-1}^\nu + v) \quad & \forall v \in V_i^\nu \ \text{ with } \ w_{i-1}^\nu + v \in \mathcal{K}_J^P \end{aligned} \tag{4.3.1}$$

in the subspaces $V_i^\nu := \operatorname{span}\{r_i^\nu\} \otimes \mathcal{Z}^P$ analogously to (4.2.3). Denoting by $\bar{u}_J^{P,\nu}$ the smoothed iterate from (4.2.4), i.e.

$$\bar{u}_J^{P,\nu} := w_{N_J}^\nu = \mathcal{M}_J^P\left((u_J^P)^\nu\right),$$

the next iterate now reads

$$(u_J^P)^{\nu+1} = w_{N_{R^\nu}}^\nu = (u_J^P)^\nu + \sum_{i=1}^{N_J} \bar{v}_i^\nu + \sum_{i=N_J+1}^{N_{R^\nu}} \bar{v}_i^\nu =: \tilde{\mathcal{C}}_J^{P,\nu}\left(\mathcal{M}_J^P\left((u_J^P)^\nu\right)\right) = \tilde{\mathcal{C}}_J^{P,\nu}(\bar{u}_J^{P,\nu}).$$

In general, an exact evaluation of the coarse grid correction $\tilde{\mathcal{C}}_J^{P,\nu}(\bar{u}_J^{P,\nu})$ is too costly in practice due to the form of the nonlinearity ϕ_J^P such that one contents oneself with an approximation $\mathcal{C}_J^{P,\nu}$ of $\tilde{\mathcal{C}}_J^{P,\nu}$. The decisive condition to ensure the convergence of the extended relaxation method nevertheless is surprisingly simple and thus powerful and consists of the monotonicity

$$\mathcal{J}\left(\mathcal{C}_J^{P,\nu}(w)\right) + \phi_J^P\left(\mathcal{C}_J^{P,\nu}(w)\right) \leq \mathcal{J}(w) + \phi_J^P(w) \quad \forall w \in \mathcal{K}_J^P. \tag{4.3.2}$$

Theorem 4.3.1. *Let the assumptions of Theorem 4.2.1 and condition (4.3.2) hold. Then, for any initial iterate $(u_J^P)^0 \in \mathcal{K}_J^P$, the sequence of iterates $\left((u_J^P)^\nu\right)_{\nu\geq 0}$ provided by the extended relaxation method*

$$\begin{aligned} \bar{u}_J^{P,\nu} &= \mathcal{M}_J^P\left((u_J^P)^\nu\right) \\ (u_J^P)^{\nu+1} &= \mathcal{C}_J^{P,\nu}(\bar{u}_J^{P,\nu}) \end{aligned}$$

converges to the solution u_J^P of the discrete problem (4.0.1).

Proof. The proof is almost literally the same as the proof for Theorem 4.2.1 if condition (4.3.2) is added to the monotonicity (4.2.5). □

Note that the coarse grid corrections alone do not need to be convergent. This provides flexibility for the construction of $\mathcal{C}_J^{P,\nu}$. We resume this part after some remarks.

Remark 4.3.2. If one assumes the monotonicity in each correction step, i.e.

$$\mathcal{J}(w_i^\nu) + \phi_J^P(w_i^\nu) \leq \mathcal{J}(w_{i-1}^\nu) + \phi_J^P(w_{i-1}^\nu) \quad \forall i = 1, \ldots, N_{R^\nu}$$

for all $\nu \geq 0$, then it can be seen by

$$\mathcal{J}\left((u_J^P)^{\nu+1}\right) + \phi_J^P\left((u_J^P)^{\nu+1}\right) \leq \mathcal{J}(w_i^\nu) + \phi_J^P(w_i^\nu) \leq \mathcal{J}\left((u_J^P)^\nu\right) + \phi_J^P\left((u_J^P)^\nu\right)$$

and the continuity of $\mathcal{J} + \phi_J^P$ on $\mathcal{K}_J^P$ that the whole sequence $(w_i^\nu)_{v\geq 0, i=1,\dots,N_{R^\nu}}$ of intermediate iterates converges with

$$w_i^\nu \to u_J^P \quad \text{as } \nu \to \infty, \text{ for } i = 1, \dots, N_{R^\nu}.$$

In either case, it holds

$$\bar{u}_J^{P,\nu} \to u_J^P \quad \text{as } \nu \to \infty. \tag{4.3.3}$$

Remark 4.3.3. In the linear case with $\phi_J^P \equiv 0$ and no constraints and with an appropriate coarsening of the spatial grid, we can use a Gauß–Seidel iteration also for the coarse corrections and obtain a multigrid algorithm for PDEs (cf. e.g. [56]) for the outer (spatial) iteration. In our SPDE context, this multigrid idea was first investigated by Le Maître et al. [75] (who used an SOR solver for the inner (stochastic) iterations) and extended by Elman and Furnival [34].

4.3.2 Constrained Newton linearization with local damping

In the following, we use the convex function $\Phi : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ given by (1.1.17) with $p_b = -1$ using the Brooks–Corey parameter functions with the extension $\Phi(u) = +\infty$ for $u < u_c$ according to Remark 3.2.19. In order to incorporate Dirichlet boundary points, we define the point-dependent convex functions

$$\Phi_{p,\pi} : u \mapsto \begin{cases} \Phi(u) & \text{for } p \in \mathcal{N}_J \backslash \mathcal{N}_J^D \\ \chi_{\{u_D(p,\pi)\}} & \text{for } p \in \mathcal{N}_J^D \end{cases}$$

with the subdifferentials

$$\partial\Phi_{p,\pi} : u \mapsto \begin{cases} \partial\Phi(u) & \text{for } p \in \mathcal{N}_J \backslash \mathcal{N}_J^D \\ \partial\chi_{\{u_D(p,\pi)\}} & \text{for } p \in \mathcal{N}_J^D, \end{cases}$$

where

$$\partial\chi_{\{u_D(p,\pi)\}} : u \mapsto \begin{cases} \mathbb{R} & \text{if } u \equiv u_D(p,\pi) \\ \emptyset & \text{else.} \end{cases}$$

Note that for $p \in \mathcal{N}_J \backslash \mathcal{N}_J^D$ the function $\Phi_{p,\pi}$ is infinitely times differentiable on the intervals $I_1 := (u_c, -1)$ and $I_2 := (-1, \infty)$. Thus, we call $(p,\pi) \in \mathcal{N}_J \times \mathcal{Q}^P$ a *critical node* of $v \in \mathcal{K}_J^P$ if p is a Dirichlet node or if $v(p,\pi)$ takes a *critical value*, i.e.

$$v(p,\pi) \in \{u_c, -1\}.$$

We define by $\mathcal{N}_J^\bullet(v)$ the set of critical nodes of v. Accordingly, the complement $\mathcal{N}_J^\circ(v) := (\mathcal{N}_J \times \mathcal{Q}^P) \backslash \mathcal{N}_J^\bullet(v)$ is the set of *regular nodes* of v.

Now, consider a regular node $(p,\pi) \in \mathcal{N}_J^\circ(\bar{u}_J^{P,\nu})$ for a given smoothed iterate $\bar{u}_J^{P,\nu}$. We can find real numbers

$$\underline{\varphi}_{\bar{u}_J^{P,\nu}}(p,\pi) < \bar{u}_J^{P,\nu}(p,\pi) < \overline{\varphi}_{\bar{u}_J^{P,\nu}}(p,\pi)$$

such that on the neighborhood $[\underline{\varphi}_{\bar{u}_J^{P,\nu}}(p,\pi), \overline{\varphi}_{\bar{u}_J^{P,\nu}}(p,\pi)]$ of $\bar{u}_J^{P,\nu}(p,\pi)$ the function $\Phi_{p,\pi}$ is twice differentiable with

$$|\Phi''_{p,\pi}(z_1) - \Phi''_{p,\pi}(z_2)| \leq L_{p,\pi}^\nu |z_1 - z_2| \quad \forall z_1, z_2 \in [\underline{\varphi}_{\bar{u}_J^{P,\nu}}(p,\pi), \overline{\varphi}_{\bar{u}_J^{P,\nu}}(p,\pi)]$$

and a pointwise Lipschitz constant $L_{p,\pi}^\nu > 0$. For instance, for $\bar{u}_J^{P,\nu}(p,\pi) \in I_1$ we set

$$[\underline{\varphi}_{\bar{u}_J^{P,\nu}}(p,\pi), \overline{\varphi}_{\bar{u}_J^{P,\nu}}(p,\pi)] := [(u_c + \bar{u}_J^{P,\nu}(p,\pi))/2, (\bar{u}_J^{P,\nu}(p,\pi) - 1)/2]$$

and for $\bar{u}_J^{P,\nu}(p,\pi) \in I_2$ we set

$$[\underline{\varphi}_{\bar{u}_J^{P,\nu}}(p,\pi), \overline{\varphi}_{\bar{u}_J^{P,\nu}}(p,\pi)] := [(-1+\bar{u}_J^{P,\nu}(p,\pi))/2, 2|\bar{u}_J^{P,\nu}(p,\pi)|+1].$$

Furthermore, we set

$$\underline{\varphi}_{\bar{u}_J^{P,\nu}}(p,\pi) = \overline{\varphi}_{\bar{u}_J^{P,\nu}}(p,\pi) = \bar{u}_J^{P,\nu}(p,\pi)$$

for $(p,\pi) \in \mathcal{N}_J^\bullet(\bar{u}_J^{P,\nu})$ and define the closed and convex set

$$\mathcal{K}_{\bar{u}_J^{P,\nu}} := \{w \in \mathcal{S}_J \otimes \mathcal{Z}^P : \underline{\varphi}_{\bar{u}_J^{P,\nu}}(p,\pi) \le w(p,\pi) \le \overline{\varphi}_{\bar{u}_J^{P,\nu}}(p,\pi) \ \ \forall p \in \mathcal{N}_J \forall \pi \in \mathcal{Q}^P\}. \tag{4.3.4}$$

The special form of ϕ_J^P given by (3.3.33) allows us to write

$$\phi_J^P(w) = \phi_{\bar{u}_J^{P,\nu}}(w) + \text{const.} \quad \forall w \in \mathcal{K}_{\bar{u}_J^{P,\nu}} \tag{4.3.5}$$

with the smooth functional

$$\phi_{\bar{u}_J^{P,\nu}} : w \mapsto \sum_{(p,\pi)\in\mathcal{N}_J^\circ(\bar{u}_J^{P,\nu})} \sum \Phi(w(p,\pi))h_p\eta_\pi \quad \forall w \in \mathcal{K}_{\bar{u}_J^{P,\nu}}. \tag{4.3.6}$$

Let us now consider the *constrained minimization* of the *smooth energy* $\mathcal{J} + \phi_{\bar{u}_J^{P,\nu}}$

$$u_{\bar{u}_J^{P,\nu}} \in \mathcal{K}_{\bar{u}_J^{P,\nu}} : \quad \mathcal{J}(u_{\bar{u}_J^{P,\nu}}) + \phi_{\bar{u}_J^{P,\nu}}(u_{\bar{u}_J^{P,\nu}}) \le \mathcal{J}(v) + \phi_{\bar{u}_J^{P,\nu}}(v) \quad \forall v \in \mathcal{K}_{\bar{u}_J^{P,\nu}}. \tag{4.3.7}$$

By (4.3.3), we have $\text{dist}(u_J^P, \mathcal{K}_{\bar{u}_J^{P,\nu}}) \to 0$ as $\nu \to \infty$, which means that the solutions of (4.3.7) tend to u_J^P. Thus, we are interested in approximate solutions of the constrained minimization problem (4.3.7). Since the critical nodes of $\bar{u}_J^{P,\nu}$ do not contribute to $\phi_{\bar{u}_J^{P,\nu}}$ according to (4.3.6), we can regard $\phi_{\bar{u}_J^{P,\nu}}$ as being smooth on a neighborhood of $\bar{u}_J^{P,\nu}$ and we can thus apply the Taylor expansion to obtain

$$\phi_{\bar{u}_J^{P,\nu}}(w) \approx \phi_{\bar{u}_J^{P,\nu}}(\bar{u}_J^{P,\nu}) + \phi'_{\bar{u}_J^{P,\nu}}(\bar{u}_J^{P,\nu})(w-\bar{u}_J^{P,\nu}) + \frac{1}{2}\phi''_{\bar{u}_J^{P,\nu}}(\bar{u}_J^{P,\nu})(w-\bar{u}_J^{P,\nu}, w-\bar{u}_J^{P,\nu}).$$

This enables us to approximate $\mathcal{J} + \phi_{\bar{u}_J^{P,\nu}}$ by the quadratic energy functional $\mathcal{J}_{\bar{u}_J^{P,\nu}}$ defined as

$$\begin{aligned}&\mathcal{J}_{\bar{u}_J^{P,\nu}}(w) := \frac{1}{2}a_{\bar{u}_J^{P,\nu}}(w,w) - \ell_{\bar{u}_J^{P,\nu}}(w) := \\ &\frac{1}{2}\left(a(w,w) + \phi''_{\bar{u}_J^{P,\nu}}(\bar{u}_J^{P,\nu})(w,w)\right) - \left(\ell(w) - \phi'_{\bar{u}_J^{P,\nu}}(\bar{u}_J^{P,\nu})w + \phi''_{\bar{u}_J^{P,\nu}}(\bar{u}_J^{P,\nu})(\bar{u}_J^{P,\nu}, w)\right),\end{aligned} \tag{4.3.8}$$

and we can regard the quadratic obstacle problem

$$w_{\bar{u}_J^{P,\nu}} \in \mathcal{K}_{\bar{u}_J^{P,\nu}} : \quad \mathcal{J}_{\bar{u}_J^{P,\nu}}(w_{\bar{u}_J^{P,\nu}}) \le \mathcal{J}_{\bar{u}_J^{P,\nu}}(v) \quad \forall v \in \mathcal{K}_{\bar{u}_J^{P,\nu}} \tag{4.3.9}$$

as a *constrained Newton linearization* of (4.3.7).

This obstacle problem is now solved by an *extended underrelaxation*, where we minimize successively in the subspaces $V_i^\nu = \text{span}\{r_i^\nu\} \otimes \mathcal{Z}^P$ for $i = N_J+1, \dots, N_{R^\nu}$. More exactly, we solve the $(P+1)$-dimensional problems

$$v_i^\nu \in \mathcal{D}_i^\nu : \quad \mathcal{J}_{\bar{u}_J^{P,\nu}}(w_{i-1}^\nu + v_i^\nu) \le \mathcal{J}_{\bar{u}_J^{P,\nu}}(w_{i-1}^\nu + v) \quad \forall v \in \mathcal{D}_i^\nu \tag{4.3.10}$$

with constraints $\mathcal{D}_i^\nu \subset V_i^\nu$ satisfying

$$0 \in \mathcal{D}_i^\nu \subset \{v \in V_i^\nu : w_{i-1}^\nu + v \in \mathcal{K}_{\bar{u}_\jmath^{P,\nu}}\}. \tag{4.3.11}$$

We show how to solve this in case of the tensor product PC basis $\{\Psi_k\} = \{\Psi_k^t\}$. To this end, we take the assumptions and notations from Subsection 4.2.2 and assume that we have a sequence of (nested) triangulations $\mathcal{T}_0, \mathcal{T}_1, \ldots, \mathcal{T}_\jmath$ of D resulting from uniform refinement, i.e. each triangle $\mathrm{t} \in \mathcal{T}_\imath$ is subdivided into four congruent subtriangles constituting $\mathcal{T}_{\imath+1}$, $\imath = 0, \ldots, \jmath-1$. This procedure also provides nested sets of nodes $\mathcal{N}_0 \subset \cdots \subset \mathcal{N}_\jmath$ and a nested sequence $\mathcal{S}_0 \subset \cdots \subset \mathcal{S}_\jmath$ of subspaces of $\mathcal{S}_\jmath$ which correspond to the levels $\imath = 0, \cdots, \jmath$. We define the multilevel nodal basis as

$$\Lambda_\mathcal{S} := (s_{p_1}^{(\jmath)}, \ldots, s_{p_{N_\jmath}}^{(\jmath)}, s_{p_1}^{(\jmath-1)}, \ldots, s_{p_{N_{\jmath-1}}}^{(\jmath-1)}, \ldots, s_{p_1}^{(0)}, \ldots, s_{p_{N_0}}^{(0)}),$$

which consists of all $N_\mathcal{S} = N_\jmath + \cdots + N_0$ nodal basis functions from all refinement levels such that

$$r_i^\nu = s_{p_i}^{(\imath_i)}, \quad i = N_\jmath + 1, \ldots, N_\jmath + N_\mathcal{S}. \tag{4.3.12}$$

Observe that the functions are ordered from fine to coarse and that the nodal basis functions on the finest grid are part of the coarse grid correction.

The stiffness matrix A and the vector b are defined as in (4.1.5) and (4.1.8), respectively, in dependence of the underlying spatial grid, the same holds for the diagonal matrix D_i from (4.2.17). Due to the Kronecker product structure of A according to Remark 4.1.3, the restriction and prolongation between different grids can be performed in the usual way, see [34] for details, whereas the vector b needs a recalculation with the new basis functions.

Now, let $\mathsf{v} \in \mathbb{R}^{P+1}$ be the coefficient vector of $v_i^\nu \in \mathcal{D}_i^\nu$ with

$$v_i^\nu(p,\pi) = \sum_{k=0}^{P} v_{ik} \Psi_k(\pi) r_i^\nu(p)$$

and $\mathfrak{v} = \mathsf{B}\mathsf{v} \in \mathbb{R}^\Pi$ its local evaluation by means of the matrix B from (4.2.12). Then, inserting (4.3.8) into (4.3.10), passing to matrix notation, omitting all constant terms, and using (4.2.16) and (4.2.17), we can rewrite the minimization problem (4.3.10)–(4.3.11) as

$$\begin{aligned}
\operatorname*{arg\,min}_{\mathfrak{v}\in\mathbb{R}^\Pi} \frac{1}{2} \Bigg(& \mathfrak{v}^T \mathsf{D}_i \mathfrak{v} + \sum_{(p,\pi)\in\mathcal{N}_\jmath^\circ(\bar{u}_\jmath^{P,\nu})} \sum H'(\bar{u}_\jmath^{P,\nu}) \left(\frac{1}{\sqrt{\eta_\pi}} \mathfrak{v}_\pi\right)^2 (r_i^\nu(p))^2 h_p \eta_\pi \Bigg) \\
& - \Bigg(\mathfrak{v}^T \mathsf{B}\mathsf{b}_i - \mathfrak{v}^T \mathsf{B}[\mathsf{A}\mathsf{w}]_i - \sum_{(p,\pi)\in\mathcal{N}_\jmath^\circ(\bar{u}_\jmath^{P,\nu})} \sum H'(\bar{u}_\jmath^{P,\nu}) \mathfrak{v}_\pi r_i^\nu(p) w_{i-1}^\nu(p,\pi) h_p \sqrt{\eta_\pi} \\
& - \sum_{(p,\pi)\in\mathcal{N}_\jmath^\circ(\bar{u}_\jmath^{P,\nu})} \sum H(\bar{u}_\jmath^{P,\nu}) \mathfrak{v}_\pi r_i^\nu(p) h_p \sqrt{\eta_\pi} + \sum_{(p,\pi)\in\mathcal{N}_\jmath^\circ(\bar{u}_\jmath^{P,\nu})} \sum H'(\bar{u}_\jmath^{P,\nu}) \mathfrak{v}_\pi r_i^\nu(p) \bar{u}_\jmath^{P,\nu}(p,\pi) h_p \sqrt{\eta_\pi} \Bigg)
\end{aligned} \tag{4.3.13}$$

with constraints

$$\underline{\varphi}_{\bar{u}_\jmath^{P,\nu}}(p,\pi) - w_{i-1}^\nu(p,\pi) \le \frac{1}{\sqrt{\eta_\pi}} \mathfrak{v}_\pi r_i^\nu(p) \le \overline{\varphi}_{\bar{u}_\jmath^{P,\nu}}(p,\pi) - w_{i-1}^\nu(p,\pi). \tag{4.3.14}$$

At this point, we have $\Pi = P + 1$ constrained one-dimensional quadratic problems which are uncoupled. As in Proposition 4.2.7, we can deduce the analogy in each step with the constrained minimization of the smooth energy $\mathcal{J}^\pi + \phi^\pi_{\bar{u}_J^{P,\nu}}$ and its Newton linearization

$$\mathcal{J}^\pi_{\bar{u}_J^{P,\nu}} = \frac{1}{2} a^\pi_{\bar{u}_J^{P,\nu}}(\cdot,\cdot) - \ell^\pi_{\bar{u}_J^{P,\nu}}(\cdot)$$

for all $\pi \in \mathcal{Q}^P$ as resulting from the stochastic collocation approach, cf. Remark 4.3.6. This enables us to transfer the basic theory from Kornhuber [69] to our case. In particular, there exist damping parameters $\vartheta^\nu_{i,\pi} \in [0,1]$ (see [69, Section 4]) such that

$$\mathcal{J}^\pi(w_i^\nu(\pi)) + \phi^\pi_{\bar{u}_J^{P,\nu}}(w_i^\nu(\pi)) \leq \mathcal{J}^\pi(w_{i-1}^\nu(\pi)) + \phi^\pi_{\bar{u}_J^{P,\nu}}(w_{i-1}^\nu(\pi)) \tag{4.3.15}$$

if

$$w_i^\nu(\pi) = w_{i-1}^\nu(\pi) + \vartheta^\nu_{i,\pi} v_i^\nu(\pi) \quad \forall \pi \in \mathcal{Q}^P \tag{4.3.16}$$

in the collocation approach, which means

$$w_i^\nu(p,\pi) = w_{i-1}^\nu(p,\pi) + \vartheta^\nu_{i,\pi} \frac{1}{\sqrt{\eta_\pi}} \mathfrak{v}_\pi r_i^\nu(p)$$

in our setting. Defining

$$\vartheta_i^\nu := \min_\pi \vartheta^\nu_{i,\pi} \tag{4.3.17}$$

and using the identity

$$\check{\mathsf{B}}[\mathsf{w}]_i + \vartheta_i^\nu \mathsf{H}^{-\frac{1}{2}} \mathfrak{v} = \check{\mathsf{B}}([\mathsf{w}]_i + \vartheta_i^\nu \mathsf{v}),$$

it is easy to see that the next iterate defined as

$$w_i^\nu = w_{i-1}^\nu + \vartheta_i^\nu v_i^\nu \tag{4.3.18}$$

provides the energy reduction

$$\mathcal{J}(w_i^\nu) + \phi_{\bar{u}_J^{P,\nu}}(w_i^\nu) \leq \mathcal{J}(w_{i-1}^\nu) + \phi_{\bar{u}_J^{P,\nu}}(w_{i-1}^\nu). \tag{4.3.19}$$

With this choice of ϑ_i^ν, we obtain the *monotone coarse grid correction*

$$\mathcal{C}_J^{P,\nu}(\bar{u}_J^{P,\nu}) = \bar{u}_J^{P,\nu} + \sum_{i=N_J+1}^{N_J+N_\mathcal{S}} \vartheta_i^\nu v_i^\nu \tag{4.3.20}$$

with local damping, which satisfies condition (4.3.2) and preserves therefore global convergence in light of Theorem 4.3.1.

In order to obtain optimal numerical complexity with $\mathcal{O}(N_J \cdot \Pi)$ point operations for each iteration step

$$(u_J^P)^{\nu+1} = \mathcal{C}_J^{P,\nu}\left(\mathcal{M}_J^P\left((u_J^P)^\nu\right)\right),$$

the implementation as a multigrid V-cycle in which calculations of corrections on a level $\imath \in \{0,\dots,J\}$ only require to access information on nodes $p \in \mathcal{N}_\imath$ is necessary. To this end, we introduce new local coarse grid obstacles $\underline{\psi}_i^\nu, \overline{\psi}_i^\nu \in V_i^\nu$ which satisfy

$$\underline{\varphi}_{\bar{u}_J^{P,\nu}}(p,\pi) - w_{i-1}^\nu(p,\pi) \leq \underline{\psi}_i^\nu(p,\pi) \leq 0 \leq \overline{\psi}_i^\nu(p,\pi) \leq \overline{\varphi}_{\bar{u}_J^{P,\nu}}(p,\pi) - w_{i-1}^\nu(p,\pi)$$
$$\forall p \in \mathcal{N}_J \, \forall \pi \in \mathcal{Q}^P.$$

The inductive construction of such obstacles presented in [68, Section 3.1.3] can be used immediately in our setting by just performing it successively for each fixed $\pi \in \mathcal{Q}^P$. The local constraints $\mathcal{D}_i^\nu$ in the local problems (4.3.10) are then given by

$$\mathcal{D}_i^\nu := \{v \in V_i^\nu : \underline{\psi}_i^\nu \leq v \leq \overline{\psi}_i^\nu\}, \quad i = N_\jmath + 1, \ldots, N_\jmath + N_\mathcal{S},$$

and the constraints (4.3.14) in the uncoupled problem read

$$\underline{\psi}_i^\nu(p,\pi) \leq \frac{1}{\sqrt{\eta_\pi}} \mathfrak{v}_\pi r_i^\nu(p) \leq \overline{\psi}_i^\nu(p,\pi). \tag{4.3.21}$$

By using these constraints, we have found a ν-independent coarse grid correction $\mathcal{C}_\jmath^{P,\mathrm{std}} = \mathcal{C}_\jmath^{P,\nu}$ and call this iteration the *stochastic version of the standard monotone multigrid method* (sSMMG)

$$(u_\jmath^P)^{\nu+1} = \mathcal{C}_\jmath^{P,\mathrm{std}}\left(\mathcal{M}_\jmath^P\left((u_\jmath^P)^\nu\right)\right), \quad \nu \geq 0. \tag{4.3.22}$$

At this point, several remarks are in order.

Remark 4.3.4. This method has the drawback that it provides, due to the definition of $\mathcal{K}_{\bar{u}_\jmath^{P,\nu}}$, the trivial correction $v_i^\nu = 0$ whenever

$$\left((\operatorname{int}\operatorname{supp}(r_i^\nu)) \times \mathcal{Q}^P\right) \cap \mathcal{N}_\jmath^\bullet(\bar{u}_\jmath^{P,\nu}) \neq \emptyset. \tag{4.3.23}$$

The same problem occurs in the deterministic case [18, p. 114], where a possible remedy is a modification of the basis functions $s_p^{(\imath)} \in \Lambda_\mathcal{S}$. This leads to the *truncated monotone multigrid method* (TMMG), see [68] for details.

Note that (4.3.23) holds already if there is only one $\pi \in \mathcal{Q}^P$ such that $\bar{u}_\jmath^{P,\nu}(p,\pi)$ takes a critical value for a $p \in \operatorname{int}\operatorname{supp}(r_i^\nu)$. Thus, it is advisable to concentrate on the uncoupled constraint (4.3.14) whenever transferring ideas of the TMMG method to our setting.

Remark 4.3.5. The choice (4.3.17) of the damping parameter ϑ_i^ν seems to be stricter than necessary. Another possibility consists in taking the diagonal matrix $\Theta_i^\nu = \operatorname{diag}(\theta_{i,\pi}^\nu) \in \mathbb{R}^{\Pi\times\Pi}$ and to retransform it. Note that the matrix

$$\tilde{\Theta}_i^\nu := \check{\mathsf{B}}^{-1}\Theta_i^\nu\check{\mathsf{B}} = \mathsf{B}^T\mathsf{H}^{\frac{1}{2}}\Theta_i^\nu\mathsf{H}^{-\frac{1}{2}}\mathsf{B},$$

which then occurs as damping in the coarse grid correction step

$$w_i^\nu = w_{i-1}^\nu + \tilde{\Theta}_i^\nu v_i^\nu \tag{4.3.24}$$

instead of (4.3.18), is no longer diagonal.

Remark 4.3.6. In this remark, we enlighten the coarse grid correction from the stochastic collocation approach mentioned in (4.3.15). Let $\bar{u}_\jmath^{P,\nu}(\pi)$ be the smoothed iterate at $\pi \in \mathcal{Q}^P$ from (4.2.4) or (4.2.28), which is the same according to Theorem 4.2.8. Defining

$$\mathcal{K}_{\bar{u}_\jmath^{P,\nu}}(\pi) := \{w(\pi) \in \mathcal{S}_\jmath : \underline{\varphi}_{\bar{u}_\jmath^{P,\nu}}(p,\pi) \leq w(p,\pi) \leq \overline{\varphi}_{\bar{u}_\jmath^{P,\nu}}(p,\pi) \ \ \forall p \in \mathcal{N}_\jmath\} \tag{4.3.25}$$

and

$$\phi_{\bar{u}_\jmath^{P,\nu}}^\pi : w \mapsto \sum_{\substack{p \text{ with} \\ (p,\pi)\in\mathcal{N}_\jmath^\circ(\bar{u}_\jmath^{P,\nu})}} \Phi(w(p,\pi))h_p \qquad \forall w(\cdot,\pi) \in \mathcal{K}_{\bar{u}_\jmath^{P,\nu}}(\pi), \tag{4.3.26}$$

the Newton linearization $\mathcal{J}^{\pi}_{\bar{u}_J^{P,\nu}}$ of the smooth energy $\mathcal{J}^{\pi} + \phi^{\pi}_{\bar{u}_J^{P,\nu}}$ reads

$$\mathcal{J}^{\pi}_{\bar{u}_J^{P,\nu}}(w) := \frac{1}{2} a^{\pi}_{\bar{u}_J^{P,\nu}}(w,w) - \ell^{\pi}_{\bar{u}_J^{P,\nu}}(w) := \frac{1}{2}\left(a^{\pi}(w,w) + (\phi^{\pi}_{\bar{u}_J^{P,\nu}})''(\bar{u}_J^{P,\nu}(\pi))(w,w)\right) - \left(\ell^{\pi}(w) - (\phi^{\pi}_{\bar{u}_J^{P,\nu}})'(\bar{u}_J^{P,\nu}(\pi))w + (\phi^{\pi}_{\bar{u}_J^{P,\nu}})''(\bar{u}_J^{P,\nu}(\pi))(\bar{u}_J^{P,\nu}(\pi), w)\right). \quad (4.3.27)$$

Using the same nodal basis (4.3.12), one can see analogously to Proposition 4.2.7 that the minimization

$$w_{\bar{u}_J^{P,\nu}}(\pi) \in \mathcal{K}_{\bar{u}_J^{P,\nu}}(\pi): \quad \mathcal{J}^{\pi}_{\bar{u}_J^{P,\nu}}(w_{\bar{u}_J^{P,\nu}}(\pi)) \le \mathcal{J}^{\pi}_{\bar{u}_J^{P,\nu}}(v) \quad \forall v \in \mathcal{K}_{\bar{u}_J^{P,\nu}}(\pi) \quad (4.3.28)$$

can be written equivalently as (4.3.13)–(4.3.14). This justifies the derivation of (4.3.19) from (4.3.15).

Comparing the new coarse grid corrections in the iterates (4.3.16) and (4.3.18), one can see that they are no longer identical (after transformation). This is due to (4.3.17), which also causes the inconvenience described in the second paragraph of Remark 4.3.4. This difference is resolved if one takes the damping matrix $\tilde{\Theta}_i^{\nu}$ as in (4.3.24).

From a computational point of view, it is however expedient to solve directly the minimizations (4.3.28) for all $\pi \in \mathcal{Q}^P$ and to retransform the solution to the PC basis afterwards. This procedure allows in particular the use of existing numerical multigrid solvers for deterministic problems without extensive changes in the code.

Remark 4.3.7. The results in this section can be applied to many other functions Φ without any changes. We refer to [69] for the treatment of more general functions and to [68] for the case of more than two critical values.

Remark 4.3.8. We close this section with a look at the limit cases for the Brooks–Corey parameter functions as introduced in (1.1.18) and (1.1.20). As mentioned in Remark 3.2.32, the function Φ is then linear on the interval $[u_c, \infty)$, and we obtain with the subdifferential

$$\partial\Phi(u) = \tilde{H}(u) = \begin{cases} \emptyset & \text{for } u < u_c \\ (-\infty, \theta_M] & \text{for } u = u_c \\ \theta_M & \text{for } u > u_c \end{cases} \quad (4.3.29)$$

a linear constrained problem, where $u_c \in \{-2, -1\}$ according to the considered setting. Using (4.3.29) in (4.2.25), the one-dimensional problems illustrated in Figure 4.2 become very easy. Furthermore, since u_c remains the only critical value, we can choose as the constraint set the whole convex set $\mathcal{K}_J^P$ from (3.3.30), i.e.

$$\mathcal{K}_{\bar{u}_J^{P,\nu}} = \mathcal{K}_J^P.$$

Moreover, by the linearity of Φ with $\Phi'' \equiv 0$ and $\Phi(0) = 0$, we have

$$\phi'_{\bar{u}_J^{P,\nu}}(\bar{u}_J^{P,\nu})w = \phi_{\bar{u}_J^{P,\nu}}(w),$$

from which follows that the constrained Newton linearization (4.3.10) is just the original problem (4.3.1) with $R^{\nu} = \Lambda_{\mathcal{S}}$ due to

$$\mathcal{J}_{\bar{u}_J^{P,\nu}} = \mathcal{J} + \phi_{\bar{u}_J^{P,\nu}} = \mathcal{J} + \phi_J^P \quad (4.3.30)$$

on $\mathcal{K}_J^P$; consequently, (4.3.13) reduces to (4.2.19). In particular, damping is no longer necessary and we end up with a block version of a usual multigrid method for a linear constrained problem.

4.3.3 Numerical results

The similarity of the nonlinear fine grid solver and the coarse grid correction in our sSMMG method with their deterministic counterparts within a stochastic collocation approach suggests that we can transfer known robustness and efficiency results (see Kornhuber [68] and following works) of the deterministic standard monotone multigrid method to our case. In order to underline this, we will investigate the convergence rates observed when applying our method to the two-dimensional test problem from the previous chapter.

Recall the setting of Subsection 3.4.2. We employ a tensor product PC basis created from Hermite polynomials in $\mathcal{Z}^P$ and a nested sequence of triangulations $\mathcal{T}_\jmath$ with corresponding nodal bases for finite element spaces $\mathcal{S}_\jmath$. We use the sSMMG solver as a $V(3,3)$ cycle with 3 pre-smoothing and 3 post-smoothing steps, where the fine grid smoother $\mathcal{M}_\jmath^P$ is the nonlinear Block Gauß–Seidel method with transformation from Subsections 4.2.1–4.2.2 and the coarse grid correction $\mathcal{C}_\jmath^{P,\mathrm{std}}$ is defined in Subsection 4.3.2 with damping matrix $\tilde{\Theta}_i^\nu$ from (4.3.24). We take the solution from former time step as initial iterate $(u_\jmath^P)^0$ and perform multigrid iterations $\nu = 0, 1, \ldots$ until the relative accuracy condition

$$\frac{\|(u_\jmath^P)^{\nu^*} - (u_\jmath^P)^{\nu^*-1}\|_a}{\|(u_\jmath^P)^{\nu^*}\|_a} \leq \mathrm{TOL} \tag{4.3.31}$$

with $\mathrm{TOL} = 10^{-12}$ is satisfied, where we utilize the energy norm

$$\|v\|_a^2 := a(v,v) \quad \forall v \in \mathcal{S}_\jmath \otimes \mathcal{Z}^P. \tag{4.3.32}$$

In order to determine the average convergence rate ϱ, we take the geometric mean of the relative corrections in relation to the previous corrections, i.e.

$$\varrho = \left(\prod_{\nu=2}^{\nu^*} \frac{\|(u_\jmath^P)^\nu - (u_\jmath^P)^{\nu-1}\|_a}{\|(u_\jmath^P)^{\nu-1} - (u_\jmath^P)^{\nu-2}\|_a} \right)^{\frac{1}{\nu^*-1}} = \left(\frac{\|(u_\jmath^P)^{\nu^*} - (u_\jmath^P)^{\nu^*-1}\|_a}{\|(u_\jmath^P)^1 - (u_\jmath^P)^0\|_a} \right)^{\frac{1}{\nu^*-1}}. \tag{4.3.33}$$

We want to carry out a comparison with usual standard monotone multigrid method (SMMG) from [68] if applied within the stochastic collocation approach from 4.1.2. In this case, we solve for each $\pi \in \mathcal{C} = \mathcal{Q}^P$, where $\mathcal{Q}^P$ is the quadrature point set corresponding to $\mathcal{Z}^P$, with the SMMG method in a $V(3,3)$ cycle until the accuracy condition

$$\frac{\|(u_\jmath)^{\nu^*}(\pi) - (u_\jmath)^{\nu^*-1}(\pi)\|_{a,\pi}}{\|(u_\jmath)^{\nu^*}(\pi)\|_{a,\pi}} \leq \mathrm{TOL} \tag{4.3.34}$$

is satisfied for iterates $(u_\jmath)^\nu(\pi) \in \mathcal{S}_\jmath$. Here, we use the norm

$$\|v\|_{a,\pi}^2 := a^\pi(v,v) \quad \forall v \in \mathcal{S}_\jmath$$

and the average convergence rates

$$\varrho^\pi = \left(\frac{\|(u_\jmath)^{\nu^*}(\pi) - (u_\jmath)^{\nu^*-1}(\pi)\|_{a,\pi}}{\|(u_\jmath)^1(\pi) - (u_\jmath)^0(\pi)\|_{a,\pi}} \right)^{\frac{1}{\nu^*-1}}$$

instead of (4.3.32) and (4.3.33), respectively. Note that this is done independently for each $\pi \in \mathcal{C}$ such that we obtain different number of iterates $\nu^* = \nu^*(\pi)$ and different convergence rates ϱ^π for different collocation points.

First, we take fixed PC order $P_1 = P_2 = 3$ and vary the maximal refinement level $\jmath$. Comparing the convergence rates of the sSMMG method with the minimal and

$\jmath$	$\min \varrho^\pi$	ϱ	$\max \varrho^\pi$
4	0.036	0.143	0.142
5	0.181	0.272	0.254
6	0.253	0.362	0.357
7	0.333	0.482	0.490
8	0.476	0.667	0.671

Table 4.1: Convergence rates for varying maximal refinement level $\jmath$.

maximal convergence rates of the SMMG in the collocation approach, we detect in Table 4.1 that ϱ and $\max_\pi \varrho^\pi$ develop more or less identically. This suggests that convergence results for SMMG can be transferred to the sSMMG method.

		P_2						
		0	1	2	3	4	5	6
P_1	0	0.373	0.396	0.359	0.369	0.371	0.363	0.366
		0.475	0.491	0.475	0.478	0.477	0.489	0.477
		0.373	0.400	0.373	0.391	0.385	0.393	0.391
	1	0.367	0.370	0.367	0.343	0.362	0.360	0.346
		0.490	0.474	0.490	0.489	0.482	0.478	0.488
		0.380	0.393	0.405	0.400	0.387	0.401	0.418
	2	0.342	0.372	0.336	0.342	0.342	0.347	0.340
		0.487	0.489	0.487	0.482	0.487	0.492	0.487
		0.421	0.404	0.436	0.439	0.421	0.445	0.465
	3	0.364	0.367	0.336	0.333	0.341	0.335	0.347
		0.507	0.487	0.508	0.482	0.501	0.490	0.491
		0.464	0.417	0.480	0.490	0.464	0.488	0.507
	4	0.362	0.358	0.344	0.340	0.343	0.325	0.335
		0.496	0.484	0.498	0.487	0.493	0.497	0.493
		0.504	0.419	0.531	0.534	0.504	0.541	0.563
	5	0.354	0.359	0.354	0.334	0.338	0.345	0.336
		0.509	0.503	0.526	0.486	0.503	0.488	0.501
		0.559	0.434	0.588	0.592	0.559	0.607	0.630
	6	0.357	0.353	0.341	0.325	0.338	0.334	0.326
		0.501	0.495	0.504	0.484	0.498	0.500	0.494
		0.650	0.432	0.664	0.662	0.650	0.662	0.675

Table 4.2: Convergence rates for varying maximal polynomial degree P_1 and P_2.

Another interesting result is given in Table 4.2, where we fix the maximal refinement level $\jmath = 7$ and solve the problem for different PC orders P_1 and P_2. Recall that the size of the PC basis is $(P_1 + 1) \cdot (P_2 + 1)$ and that we can connect a quadrature point set $\mathcal{Q}^P$ to each of these basis sets. In each field of Table 4.2, one can read the following information:

$$\begin{array}{|c|}\hline \min \varrho^\pi \\ \varrho \\ \max \varrho^\pi \\ \hline \end{array}$$

Hence, ϱ with values 0.47–0.52 remains almost constant for different polynomial degrees and is therefore mainly dependent on the refinement level $\jmath$ and the problem — like its deterministic counterpart. For the collocation, the values of $\min_\pi \varrho^\pi$ are

also almost constant for different collocation point sets while $\max_\pi \varrho^\pi$ is increasing as $|\mathcal{C}|$ becomes larger. This is a consequence of the fact that deterministic problems are solved for more and more extreme collocation points π with corresponding $K(\cdot, \pi)$ the larger $|\mathcal{C}|$ is.

Note that the collocation approach solves successively for each $\pi \in \mathcal{C}$ and treats all single problems in an equal way by solving them up to a precision determined by (4.3.34). If, however, one wants to compute moments like $\mathbb{E}[u]$, the values of the solution $u(\pi)$ for such extreme π only contribute to a more and more negligible extent to this evaluation, since the quadrature weights corresponding to these points tend to zero. On the other hand by the definition of the norm (4.3.32), this distinction into more important and less important collocation and quadrature points is detected by the sSMMG method and part of the stopping condition (4.3.31), which prevents the increase of the convergence rates for large P_r, as can be seen in Table 4.2. This underlines the efficiency of the sSMMG solver and points the way of how to make the collocation approach more efficient: it is expedient to perform each SMMG step parallel for all $\pi \in \mathcal{C}$, to create after each multigrid step the approximated function as in (4.1.17) and to use a weighted norm which is equivalent to (4.3.32) to determine the stopping condition for all multigrid iterations.

4.4 Post-processing

The results presented in this chapter allow the approximation of the function $u(x, \xi(\omega)) \in H^1(D) \otimes L^2(\Omega)$ and its representation as

$$u(x, \xi(\omega)) = \sum_{i=1}^{N_J} \sum_{k=0}^{P} u_{ik} s_{p_i}^{(J)}(x) \Psi_k(\xi(\omega)) \in \mathcal{S}_J \otimes \mathcal{Z}^P. \tag{4.4.1}$$

Whilst the pointwise knowledge of u in $x \in D$ is of main interest, it is hardly relevant to know the specific value of u in a certain $\omega \in \Omega$; instead of that, we want to be able to compute statistics and probability density functions and to make statements about certain probabilities. In this section, we will show how appropriate and powerful the PC representation (4.4.1) is to achieve this goal.

We start with some elementary definitions which can be found in textbooks in statistics like [36].

Definition 4.4.1. Let X be a random variable on Ω. Then

$$m_n^X := \mathbb{E}\left[X^n\right]$$

is called the *moment of order* n of X and

$$\mu_n^X := \mathbb{E}\left[(X - \mathbb{E}[X])^n\right]$$

is called the *central moment of order* n of X if the respective expectation values exist.

The most common statistics which describe the nature of X besides the expection value $m_1^X = \mathbb{E}[X]$ are the following.

Definition 4.4.2. Let X be a random variable on Ω. Then we call

$$\begin{aligned}\mathrm{Var}[X] &:= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right],\\ \gamma_1[X] &:= \frac{\mathbb{E}\left[(X - \mathbb{E}[X])^3\right]}{\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]^{3/2}},\\ \gamma_2[X] &:= \frac{\mathbb{E}\left[(X - \mathbb{E}[X])^4\right]}{\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]^2}\end{aligned}$$

the *variance* $\mathrm{Var}[X]$, the *skewness* $\gamma_1[X]$ and the *kurtosis* $\gamma_2[X]$ of X if the respective expectation values exist and the denominators are not zero.

Obviously, these values can be rewritten according to

$$\mathrm{Var}[X] = \mu_2^X = m_2^X - (m_1^X)^2, \tag{4.4.2}$$

$$\gamma_1[X] = \frac{\mu_3^X}{(\mu_2^X)^{3/2}} = \frac{m_3^X - 3m_2^X m_1^X + 2(m_1^X)^3}{(m_2^X - (m_1^X)^2)^{3/2}}, \tag{4.4.3}$$

$$\gamma_2[X] = \frac{\mu_4^X}{(\mu_2^X)^2} = \frac{m_4^X - 4m_3^X m_1^X + 6m_2^X (m_1^X)^2 - 3(m_1^X)^4}{(m_2^X - (m_1^X)^2)^2}. \tag{4.4.4}$$

The statistics from Definition 4.4.2 are hence known once either the moments or central moments have been computed.

Writing the solution u from (4.4.1) as

$$u(x, \xi(\omega)) = \sum_{k=0}^{P} u_k(x)\Psi_k(\xi(\omega)), \tag{4.4.5}$$

it is easy to see that the moments of order n of $u(x,\cdot)$ are given by

$$m_n^u(x) = \mathbb{E}\left[u^n(x,\cdot)\right] = \sum_{k_1=0}^{P} \cdots \sum_{k_n=0}^{P} \left(\prod_{i=1}^{n} u_{k_i}(x)\right) \mathbb{E}\left[\prod_{i=1}^{n} \Psi_{k_i}\right]. \tag{4.4.6}$$

For $\{\Psi_k\} = \{\Psi_k^c\}$ or $\{\Psi_k\} = \{\Psi_k^t\}$, the central moments are computed in one go, since

$$\mu_n^u(x) = \mathbb{E}\left[(u(x,\cdot) - \mathbb{E}[u(x,\cdot)])^n\right] = \sum_{k_1=1}^{P} \cdots \sum_{k_n=1}^{P} \left(\prod_{i=1}^{n} u_{k_i}(x)\right) \mathbb{E}\left[\prod_{i=1}^{n} \Psi_{k_i}\right].$$

due to (3.3.11) and $\mathbb{E}[u(x,\cdot)] = u_0(x)$. Here, the expectation values on the right-hand side of (4.4.6) can be calculated once and stored for the rest of the time. Observe that $\mathbb{E}[\Psi_{k_1}]$ and $\mathbb{E}[\Psi_{k_1}\Psi_{k_2}]$ in the case $n = 1$ and $n = 2$ are already known from the assembly in Subsection 4.1.1, whereas the expectation values for higher order n, for example $\mathbb{E}[\Psi_{k_1}\Psi_{k_2}\Psi_{k_3}]$ for $k_1, k_2, k_3 = 0, \ldots, P$ in the case $n = 3$, have to be computed by now using quadrature formulas of appropriate order.

In our hydrological setting, we are rather interested in the statistics of the physical pressure $p(x,\cdot) = \kappa^{-1}(u(x,\cdot))$ than in the corresponding values of u. Since the Kirchhoff transformation κ is not linear in general, a direct conversion from, say, $\mathrm{Var}[u](x)$ to $\mathrm{Var}[p](x)$ is not possible (cf. (2.3.7)). Rather, we compute the moments

$$m_n^p(x) = \mathbb{E}\left[p^n(x,\cdot)\right] = \mathbb{E}\left[\left(\kappa^{-1}(u(x,\cdot))\right)^n\right]$$

directly be means of high-order stochastic quadrature and obtain the desired values via (4.4.2)–(4.4.4).

In order to answer questions like (Q3) from the introduction, it is moreover essential to know at arbitrary $x \in D$ the *probability density functions* $\mathrm{pdf}_x^u : \mathbb{R} \to \mathbb{R}^+$ and $\mathrm{pdf}_x^p : \mathbb{R} \to \mathbb{R}^+$, which are defined such that

$$\mathbb{P}(a \leq u(x,\cdot) \leq b) = \int_a^b \mathrm{pdf}_x^u(s)\,\mathrm{d}s \quad \text{and} \quad \mathbb{P}(a \leq p(x,\cdot) \leq b) = \int_a^b \mathrm{pdf}_x^p(s)\,\mathrm{d}s,$$

respectively, or, equivalently, the corresponding *cumulative distribution functions* $\mathrm{cdf}_x^u : \mathbb{R} \to [0,1]$ and $\mathrm{cdf}_x^p : \mathbb{R} \to [0,1]$ with

$$\mathrm{cdf}_x^u(y) = \mathbb{P}(u(x,\cdot) \leq y) = \int_{-\infty}^y \mathrm{pdf}_x^u(s)\,\mathrm{d}s, \tag{4.4.7}$$

$$\mathrm{cdf}_x^p(y) = \mathbb{P}(p(x,\cdot) \leq y) = \int_{-\infty}^y \mathrm{pdf}_x^p(s)\,\mathrm{d}s. \tag{4.4.8}$$

A direct estimation of pdf_x^u from the representation (4.4.5) is possible in special cases, for instance if the underlying random variables ξ_r from (3.1.11) are Gaussian, cf. [102] and the references therein; our approach, however, should work for arbitrary ξ_r and for the density functions of u as well as of $p = \kappa^{-1}(u)$. To this end, we introduce for fixed $x \in D$ a partition $\mathcal{A} = \{A_1, \ldots, A_{N_\mathcal{A}}\}$ of the space $\Omega^{(M)}$ and approximate the functions $u(x,\cdot)$ and $p(x,\cdot)$ by step functions with values $u(x,\bar{a}_i)$ and $p(x,\bar{a}_i)$ for certain points $\bar{a}_i \in A_i$. An estimation of the distribution functions is then obtained by

$$\mathrm{cdf}_x^u(y) = \sum_{\substack{A_i \in \mathcal{A} \\ u(x,\bar{a}_i) \leq y}} \mathbb{P}^{(M)}(A_i) \quad \text{and} \quad \mathrm{cdf}_x^p(y) = \sum_{\substack{A_i \in \mathcal{A} \\ p(x,\bar{a}_i) \leq y}} \mathbb{P}^{(M)}(A_i).$$

The functions pdf_x^u and pdf_x^p can then be derived by virtue of (4.4.7) and (4.4.8).

At the end of the following chapter with a hydrological example, the described post-processing methods will be applied and will provide further insights for the analysis of the problem.

Chapter 5

A hydrological example

This final chapter is devoted to the application of the presented results to a typical hydrological problem. Our computations are carried out on a realistic geometry, and we describe which further numerical challenges are connected with it. Then, we use measured data to create the Karhunen–Loève expansion for the permeability and solve the Richards equation over a certain time interval. Our aim is to point out the benefits and the limits of the polynomial chaos approach and our solution method.

5.1 Line smoothers for an anisotropic grid

We consider the bend of a river passing through a landscape. In horizontal direction the region is of the size $1500 \times 1500\,[m^2]$, while in z-direction the geometry is depending on the surface of the landscape and the aquifers and is varying between 4 and $14\,[m]$. This is depicted in Figure 5.1, where the river is colored in blue and the z-direction is scaled (throughout this chapter) by the factor 20.

We take a spatial grid as it is often utilized in the hydrological context. Recalling the coordinates $x = (x_1, x_2, z) \in D$ from Chapter 1, the coarse grid consists of hexahedra and is constructed in the following way: first, create a regular two-dimensional grid in the horizontal x_1-x_2-plane consisting of 30^2 squares with size 50×50; then divide the line along the z-axis at each grid point (x_1, x_2) into intervals in order to create grid points (x_1, x_2, z_i) with $z_{\min} = z_1 < z_2 < \ldots < z_7 = z_{\max}$. We obtain a total of $31^2 \cdot 7 = 6727$ nodes and $30^2 \cdot 6 = 5400$ hexahedral elements in six layers on top of each other, see Figure 5.2. The fine grids are obtained by uniform refinement and feature the following size:

level j	# elements	# nodes
0	5 400	6 727
1	43 200	48 373
2	345 600	366 025
3	2 764 800	2 845 969

The form of the Richards equation and the structure of the grid need further attention. First, we detect that the convective part in (3.2.2) can cause stability problems. This is often solved by upwind schemes and artificial viscosity (see [61,

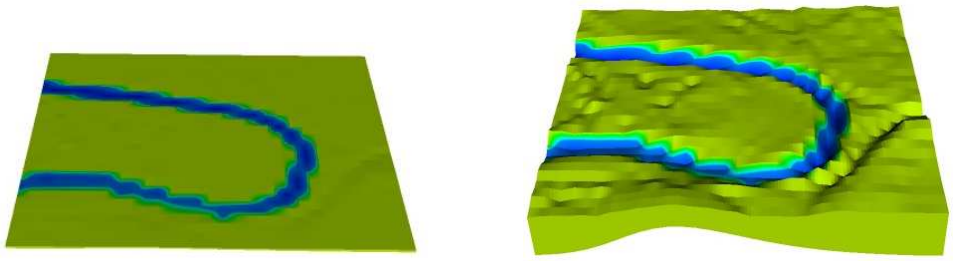


Figure 5.1: The computational domain in original scale (left) and scaled in z-direction (right).

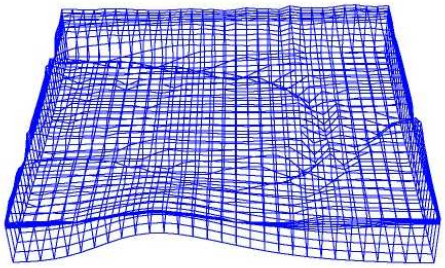


Figure 5.2: The coarse grid.

Chapter 9]). For the deterministic Richards equation and under consideration of the special form of the grid, where nodes lie on top of each other on lines parallel to the z-axis, we refer to the investigations in [18]. Since these modifications only affect the spatial discretization of the differential operators, we can adopt this approach immediately to our case, see the assembling procedure in Subsection 4.1.1.

Secondly, observe that the elements are anisotropic, because $h_z \ll h_{x_1} = h_{x_2}$, if h_z denotes the maximal diameter of an element parallel to the z-axis. As can be seen by transformation, this is equivalent to an anisotropic problem on an isotropic grid, since

$$-u_{x_1x_1} - u_{x_2x_2} - \frac{1}{\varepsilon} u_{zz} = 0 \quad \text{in } [0,1]^3$$

yields the same discretization as

$$-u_{x_1x_1} - u_{x_2x_2} - u_{zz} = 0 \quad \text{in } [0,1]^2 \times [0, \sqrt{\varepsilon}]$$

for the simple Laplace's equation. It is known (see [56, Chapter 10]) that convergence rates for multigrid solvers with Gauß–Seidel pre-smoothers tend to 1 as $\varepsilon \to 0$. A possible remedy proposed in [21] are line smoothers in direction of the anisotropy. We will extend this idea to our case with obstacle conditions and stochastic coordinates for the coarse grid corrections as described in Subsection 4.3.2.

At the beginning, we define a line $\mathcal{N}_\imath^{[i]}$ as the set of nodes $p \in \mathcal{N}_\imath$ having the same x_1 and x_2 coordinates, see the marked nodes in Figure 5.3. Let $N_\imath^{[\,]}$ denote the number of lines and $n_\imath^{[i]}$ the number of nodes in $\mathcal{N}_\imath^{[i]}$. For example in our coarse grid $\imath = 0$, we identify $N_0^{[\,]} = 31^2$ lines each consisting of $n_0^{[i]} = 7$ nodes. Define

$$R_\imath^{\nu,[i]} := \left\{ s_p^{(\imath)} : s_p^{(\imath)}(q) = 0 \text{ for all } q \notin \mathcal{N}_\imath^{[i]} \right\}$$

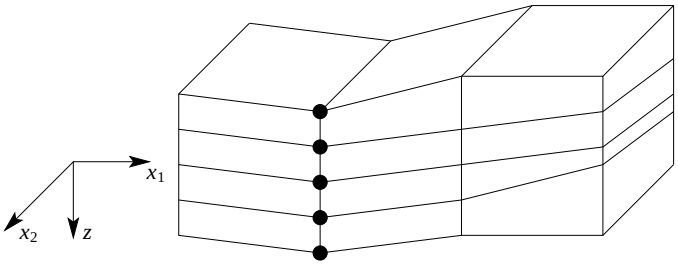


Figure 5.3: The line smoother solves along the z-axis.

as the set of nodal basis functions corresponding to the line $\mathcal{N}_i^{[i]}$. We want to apply the line smoother in the coarse grid correction $\mathcal{C}_\jmath^{P,\nu}$ and solve the quadratic obstacle problem (4.3.9) by a successive minimization as in (4.3.10), but now in the subspaces $V_{[i]}^\nu = \operatorname{span} R_i^{\nu,[i]} \times \mathcal{Z}^P$ for all lines $[i] = 1, \dots, N_i^{[]}$ obtaining iterates $w_{[i]}^\nu = w_{[i-1]}^\nu + v_{[i]}^\nu$ and for all levels $\imath = \jmath, \jmath-1, \dots, 0$. The only modification compared with the method described in (4.3.10)–(4.3.22) is that we now encounter blocks of size $(P+1) \cdot n_i^{[i]}$ and that we have to solve within this block in a way that the energy reduction property (4.3.19) is still valid.

To this end, denote by A^ϕ and b^ϕ the matrix and vector to $a_{\bar{u}_\jmath^{P,\nu}}(\cdot,\cdot)$ and $\ell_{\bar{u}_\jmath^{P,\nu}}(\cdot)$ from (4.3.8), respectively, and by $\mathsf{A}^{\pi,\phi}$ and $\mathsf{b}^{\pi,\phi}$ the matrix and vector to their counterparts $a^\pi_{\bar{u}_\jmath^{P,\nu}}(\cdot,\cdot)$ and $\ell^\pi_{\bar{u}_\jmath^{P,\nu}}(\cdot)$ from (4.3.27), respectively. We assume again that we have a tensor product PC basis $\{\Psi_k\} = \{\Psi_k^t\}$ and adopt the notation from Subsections 4.2.1 and 4.2.2 with the modification that the block diagonal matrix $\underline{\mathsf{B}}$ with entries $\underline{\mathsf{B}}_{jj} = \mathsf{B} \in \mathbb{R}^{\Pi\times(P+1)}$ is now of the size $\underline{\mathsf{B}} \in \mathbb{R}^{n_i^{[i]}\Pi \times n_i^{[i]}(P+1)}$. The minimization problem (4.3.10)–(4.3.11) within the introduced subspaces $V_{[i]}^\nu$ is

$$\begin{aligned}
&\operatorname*{arg\,min}_{\bar{\mathsf{v}}\in\mathbb{R}^{n_i^{[i]}(P+1)}} \frac{1}{2}(\mathsf{w}+\mathsf{v})^T\mathsf{A}^\phi(\mathsf{w}+\mathsf{v}) - (\mathsf{w}+\mathsf{v})^T\mathsf{b}^\phi \\
&\text{subject to } \underline{\varphi}_{\bar{u}_\jmath^{P,\nu}}(p,\pi) \le w_{[i-1]}^\nu(p,\pi) + v_{[i]}^\nu(p,\pi) \le \overline{\varphi}_{\bar{u}_\jmath^{P,\nu}}(p,\pi) \quad \forall \pi \in \mathcal{Q}^P \, \forall p \in \mathcal{N}_\jmath,
\end{aligned} \tag{5.1.1}$$

where the long block vector $\mathsf{v} \in \mathbb{R}^{N_i(P+1)}$ as coefficient vector of $v_{[i]}^\nu$ has the block entry $\mathsf{v}_{[i]} = \bar{\mathsf{v}}$ and $\mathsf{v}_{[j]} = 0$ for $[j] \neq [i]$. In the same way as in the derivation of (4.3.13), (5.1.1) can be rewritten with $\mathfrak{v} = \underline{\mathsf{B}}\mathsf{v}_{[i]}$ as

$$\begin{aligned}
&\operatorname*{arg\,min}_{\mathfrak{v}_\pi\in\mathbb{R}^{n_i^{[i]}}} \frac{1}{2}\mathfrak{v}_\pi^T\left[\underline{\mathsf{B}}\mathsf{A}^\phi_{[i][i]}\underline{\mathsf{B}}^T\right]_{\pi\pi}\mathfrak{v}_\pi - \mathfrak{v}_\pi^T\left[\underline{\mathsf{B}}\left(\mathsf{b}^\phi_{[i]} - [\mathsf{A}^\phi\mathsf{w}]_{[i]}\right)\right]_\pi \\
&\text{subject to } \underline{\varphi}_{\bar{u}_\jmath^{P,\nu}}(p_j,\pi) - w_{[i-1]}^\nu(p_j,\pi) \le \frac{1}{\sqrt{\eta_\pi}}[\mathfrak{v}_\pi]_j \le \overline{\varphi}_{\bar{u}_\jmath^{P,\nu}}(p_j,\pi) - w_{[i-1]}^\nu(p_j,\pi) \\
&\qquad\qquad \forall p_j \in \mathcal{N}_i^{[i]}
\end{aligned} \tag{5.1.2}$$

for all $\pi \in \mathcal{Q}^P$, since we have again the uncoupling in the quadrature points $\pi \in \mathcal{Q}^P$, and which is (up to a scalar factor) the same as the minimization problem from the

collocation approach reading

$$\begin{aligned}&\arg\min_{\mathsf{v}^\pi\in\mathbb{R}^{n_i^{[i]}}} \frac{1}{2}(\mathsf{v}^\pi)^T \left[\mathsf{A}^{\pi,\phi}\right]_{[i][i]} \mathsf{v}^\pi - (\mathsf{v}^\pi)^T \left(\mathsf{b}_{[i]}^{\pi,\phi} - \left[\mathsf{A}^{\pi,\phi}\mathsf{w}^\pi\right]_{[i]}\right)\\ &\text{subject to } \underline{\varphi}_{\bar{u}_j^{P,\nu}}(p_j,\pi) - w_{[i-1]}^\nu(p_j,\pi) \le v_{[i]}^\nu(p_j,\pi) \le \overline{\varphi}_{\bar{u}_j^{P,\nu}}(p_j,\pi) - w_{[i-1]}^\nu(p_j,\pi)\\ &\qquad\qquad \forall p_j \in \mathcal{N}_i^{[i]} \quad (5.1.3)\end{aligned}$$

with v^π as coefficient vector of $v_{[i]}^\nu(\pi)$, see the arguments in the proof of Theorem 4.2.8. The matrix $\left[\underline{\mathsf{B}}\mathsf{A}_{[i][i]}^{\phi}\underline{\mathsf{B}}^T\right]_{\pi\pi} = \left[\mathsf{A}^{\pi,\phi}\right]_{[i][i]} \in \mathbb{R}^{n_i^{[i]}\times n_i^{[i]}}$ describes $a^\pi_{\bar{u}_j^{P,\nu}}(s_{p_{j_1}}^{(\imath)}, s_{p_{j_2}}^{(\imath)})$ for nodal basis functions $s_{p_{j_1}}^{(\imath)}, s_{p_{j_2}}^{(\imath)} \in R_i^{\nu,[i]}$ in the one-dimensional vertical line $\mathcal{N}_i^{[i]}$ and is thus (after ordering) a tridiagonal matrix $\mathsf{T}^{\pi,[i]}$. This means that whether using the stochastic collocation or the stochastic Galerkin approach with transformation, one ends up for each $\pi \in \mathcal{Q}^P$ and each line $\mathcal{N}_i^{[i]}$ with a quadratic minimization problem of the form

$$\arg\min_{\mathsf{v}\in\mathbb{R}^{n_i^{[i]}}} \frac{1}{2}\mathsf{v}^T\mathsf{T}^{\pi,[i]}\mathsf{v} - \mathsf{v}^T\mathsf{r}^{\pi,[i]} \tag{5.1.4}$$

with constraints as in (5.1.2) or (5.1.3), respectively. This can be solved by any adequate descent method provided that it yields the energy reduction

$$\mathcal{J}(w_{[i]}^\nu) + \phi_{\bar{u}_j^{P,\nu}}(w_{[i]}^\nu) \le \mathcal{J}(w_{[i-1]}^\nu) + \phi_{\bar{u}_j^{P,\nu}}(w_{[i-1]}^\nu). \tag{5.1.5}$$

For our computations, we solve (5.1.4) by an iteration $(\mathsf{v}^k)_k$ by taking a variant of the truncated nonsmooth Newton multigrid method from [51, Section 6] with a projected Gauß–Seidel method as pre-smoother, a linear correction c^k which is obtained by the solution of

$$\mathsf{T}^{\pi,[i]}\mathsf{c}^k = \mathsf{r}^{\pi,[i]} - \mathsf{T}^{\pi,[i]}\mathsf{v}^{k-1},$$

and a damping in order to provide (5.1.5). Finally incorporate the constraints $\underline{\psi}_j^\nu$, $\overline{\psi}_j^\nu$ from (4.3.21) for all $p_j \in \mathcal{N}_i^{[i]}$ into the current block iteration step to reduce numerical complexity. With these modifications to our sSMMG method, we observe in the following computations average convergence rates ranging from 0.9 to 0.95.

5.2 Solution of the Richards equation

In this final section, we present an example of the stochastic Richards equation for the three-dimensional domain introduced at the beginning of this chapter.

We solve for the pressure p, which is given in meters of water column and set atmospheric pressure equal to $0\,[m]$. We use the Brooks–Corey model from Section 1.1 corresponding to sandy soil with $\lambda = 1$, $e(\lambda) = 5$, $u_c = -1.25$, $\theta_m = 0.2$, $\theta_M = 0.95$. The bubbling pressure is given by $-0.1\,[m]$. Since we normalized $p_b = -1\,[m]$ throughout this thesis, we introduce the dimensionless factor p_r such that $p_r p_b = -0.1\,[m]$. Consequently, the Kirchhoff transformation to obtain the generalized pressure u reads $u = \kappa(p/p_r)$ and $p = p_r\kappa^{-1}(u)$. Other parameters in our model are $\mathfrak{n} = 0.35$, $\rho = 10^3\,[kg/m^3]$, $g = 10\,[m/s]$, and $\eta = 10^{-3}\,[kg/ms]$.

Our goal is to model the infiltration of river water into a dry soil. Figure 5.4 shows the deterministic initial condition $p^0(x) = p(t = 0, x)$. In the riverbed, the pressure is equal to the water column above, i.e. equal to the difference between the

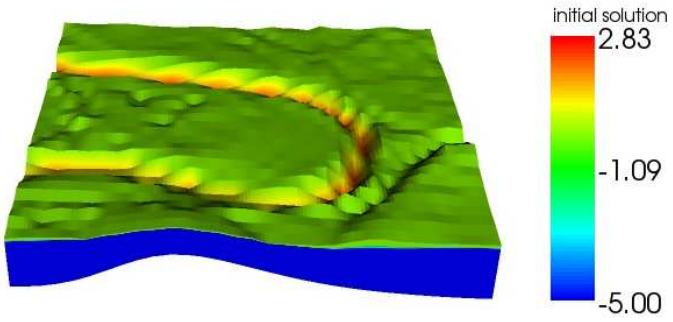


Figure 5.4: The initial condition $p^0(x)$.

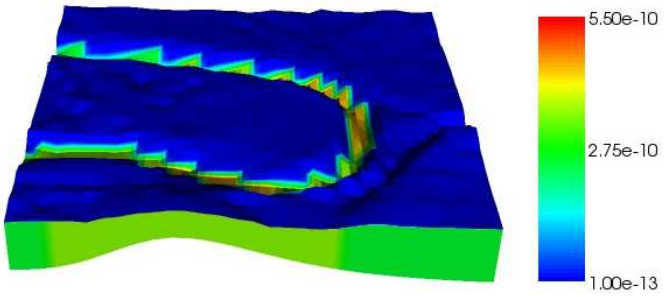


Figure 5.5: The function $\exp(\bar{K})$.

water table of the river and the ground. The water table is modeled with a linear incline along the course of the river and is constant in time. Thus, we use Dirichlet boundary conditions in the riverbed which are constant in t. The remaining top surface is modeled with Dirichlet boundary conditions equal to the atmospheric pressure (we will not consider seepage faces which can be modeled by Signorini boundary conditions, cf. [18]), and the same pressure can be found in the soil layers directly below the surface. The rest of the soil has the initial pressure $-5\,[m]$, which means that it is completely dry. Finally, all remaining boundaries (the vertical ones and the one at the bottom of our domain) have Neumann conditions $f_N = 0$.

Let us now create the stochastic permeability K. We assume a lognormal distribution as it is usually done in hydrology (see the references in Remarks 2.3.8 and 3.1.27) and consider

$$K(x,\omega) = \exp\left(\bar{K}(x) + \sum_{r=1}^{M} \sqrt{\lambda_r} g_r(x)\xi_r(\omega)\right) \tag{5.2.1}$$

with normally distributed $\xi_r \propto \mathcal{N}(0,1)$. In order to calculate the eigenfunctions g_r, knowledge of the covariance structure is needed. Frequently used in hydrology (see e.g. [44, 76, 100, 106]) is the exponential covariance kernel from Appendix A. For our three-dimensional domain D with points $x = (x_1, x_2, z)$ and $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \tilde{z})$, take

$$V_K(x,\tilde{x}) = \exp\left(-\left(\frac{|x_1-\tilde{x}_1|^2}{\gamma_1^2} + \frac{|x_2-\tilde{x}_2|^2}{\gamma_2^2} + \frac{|z-\tilde{z}|^2}{\gamma_3^2}\right)^{1/2}\right). \tag{5.2.2}$$

As correlation lengths, we estimate from the data $\gamma_1 = 200$, $\gamma_2 = 150$, $\gamma_3 = 5$, which is in accordance with estimates from other measurements (cf. [44, Section 6.1]). The

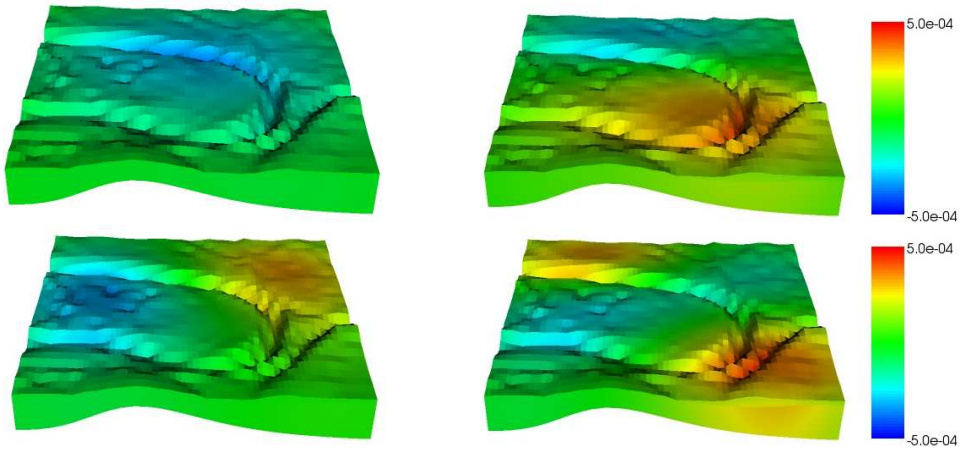


Figure 5.6: The eigenfunctions g_r, $r = 1, \dots, 4$.

next step is to assemble the matrices in the generalized eigenvalue problem

$$\mathsf{V}\mathsf{g} = \lambda \mathsf{M}\mathsf{g}$$

from (3.1.36). This is done on level $\jmath = 0$ such that V is a dense 6727×6727-matrix. Using common eigenproblem solvers, one obtains the following eigenvalues:

$$\begin{aligned}
\lambda_1 &= 1\,328\,177 \\
\lambda_2 &= 955\,317 \\
\lambda_3 &= 884\,811 \\
\lambda_4 &= 691\,483 \\
\lambda_5 &= 617\,473 \\
\lambda_6 &= 514\,826 \\
&\dots
\end{aligned}$$

We decide to truncate the expansion for all $\lambda_i < \lambda_1/2$, which means $M = 4$. Note that the numerical complexity increases rapidly with M and that less strict truncations (e.g. $\lambda_i < \lambda_1/10$ with $M = 26$ or $\lambda_i < \lambda_1/100$ with $M = 143$) would lead to stochastic dimensions where the polynomial chaos approach is no longer superior to the Monte Carlo method. The corresponding eigenfunction g_r are normalized and interpolated to the refined grids, see Figure 5.6. Finally, the function $\bar{K}(x)$ is given by the measured data, where we have $\exp(\bar{K}(x)) = 10^{-13}\,[m^2]$ on the surface and a variation $\exp(\bar{K}(x)) \in [2.5 \cdot 10^{-10}, 5.5 \cdot 10^{-10}]\,[m^2]$ in the remaining soil, see Figure 5.5. Altogether, the stochastic permeability $K(x, \omega)$ varies over a broad range and satisfies

$$\mathbb{P}\left(K(x,\omega) \in [7.9 \cdot 10^{-12}, 1.1 \cdot 10^{-8}]\right) \geq 0.989$$

outside the surface.

For the time evolution, we fix a constant time step size $\tau = 100\,[s]$. As spatial grid, we choose the uniformly refined grid with $\jmath = 2$ such that we perform our multigrid solver on three different grid levels. In order to choose the dimension of the PC basis, we proceed by applying an adaptive algorithm following an idea in [12, Section 9.3]. Start with $P_1 = P_2 = P_3 = P_4 = 0$, shortly $(0, 0, 0, 0)$, solve the Richards equation for the first time step, and compute $E_0 = \mathbb{E}[u]$. Then, carry

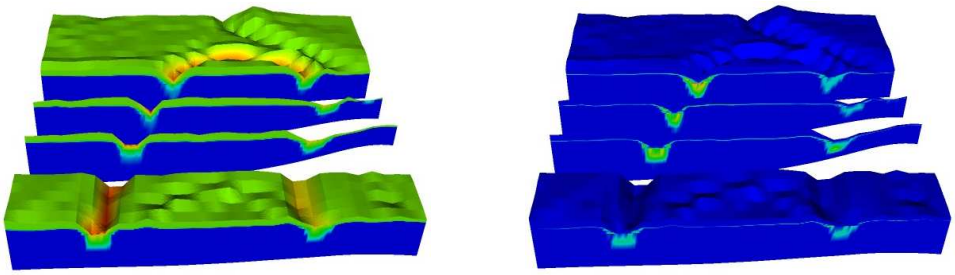


Figure 5.7: $\mathbb{E}[p]$ and $\mathrm{Var}[p]$ at $t = 100$.

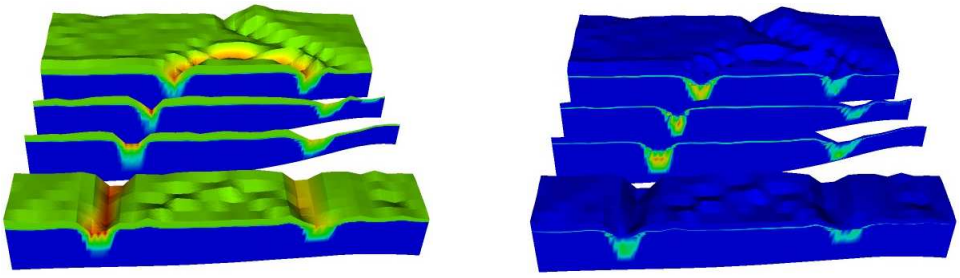


Figure 5.8: $\mathbb{E}[p]$ and $\mathrm{Var}[p]$ at $t = 200$.

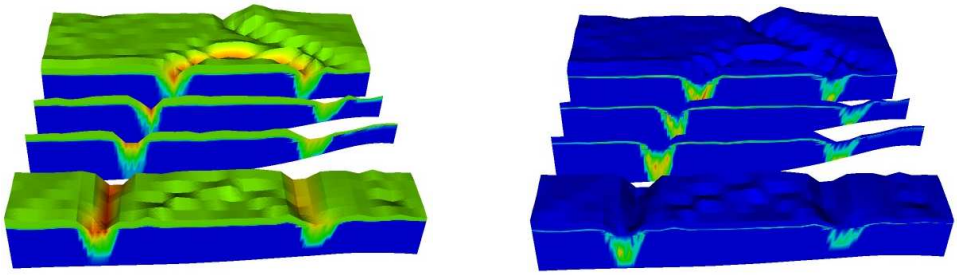


Figure 5.9: $\mathbb{E}[p]$ and $\mathrm{Var}[p]$ at $t = 500$.

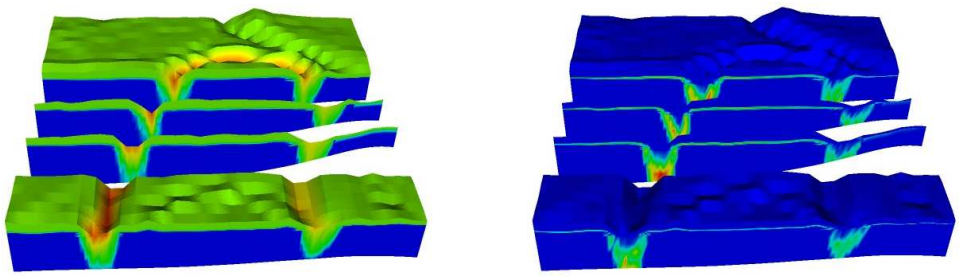


Figure 5.10: $\mathbb{E}[p]$ and $\mathrm{Var}[p]$ at $t = 1\,000$.

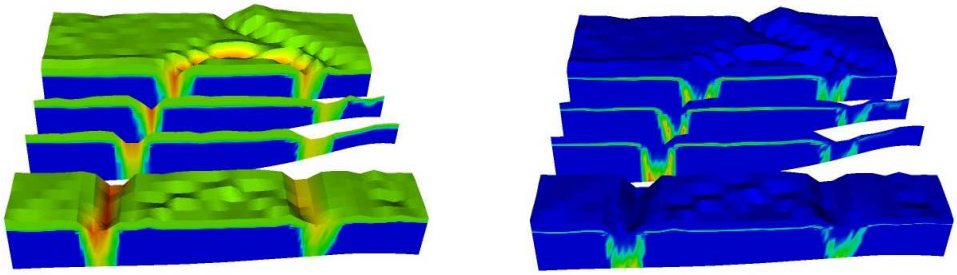


Figure 5.11: $\mathbb{E}[p]$ and $\mathrm{Var}[p]$ at $t = 2\,000$.

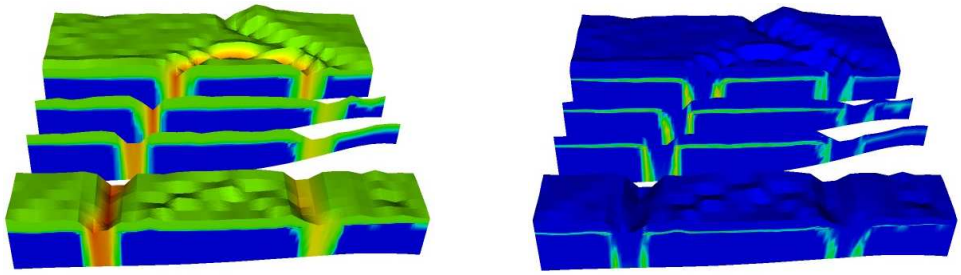


Figure 5.12: $\mathbb{E}[p]$ and $\mathrm{Var}[p]$ at $t = 5\,000$.

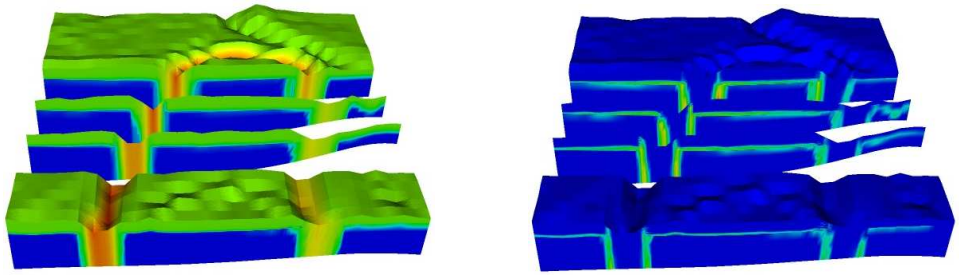


Figure 5.13: $\mathbb{E}[p]$ and $\mathrm{Var}[p]$ at $t = 10\,000$.

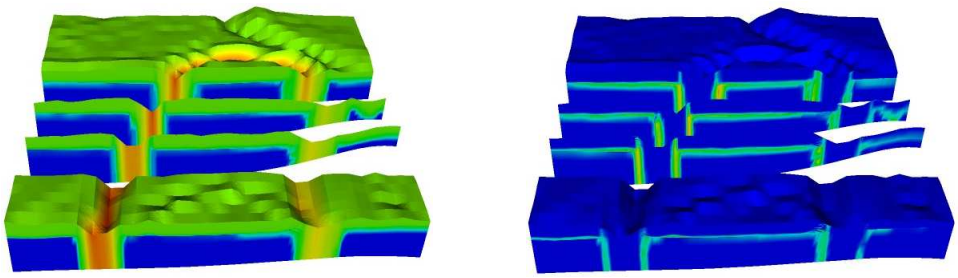


Figure 5.14: $\mathbb{E}[p]$ and $\mathrm{Var}[p]$ at $t = 15\,000$.

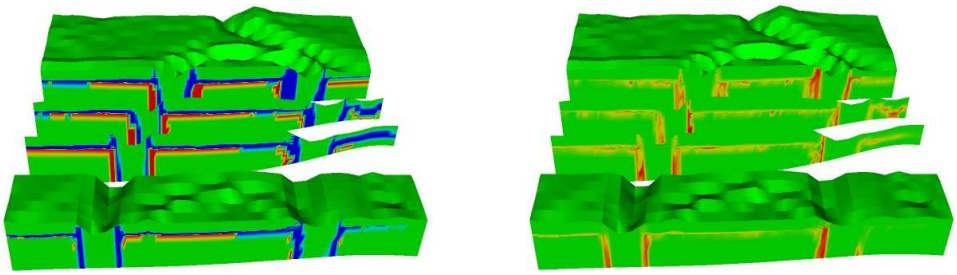


Figure 5.15: $\gamma_1[p]$ and $\gamma_2[p]$ at $t = 15\,000$.

out the same for $P_1 = 1$ and $P_2 = P_3 = P_4 = 0$, shortly $(1, 0, 0, 0)$, and obtain $E_1 = \mathbb{E}[u]$. If

$$\frac{\|E_1 - E_0\|_{L^2(D)}}{\|E_0\|_{L^2(D)}} > \text{TOL},$$

accept this step and compute $E_2 = \mathbb{E}[u]$ for the PC basis set $(1, 1, 0, 0)$, otherwise reject this step and compute $E_1 = \mathbb{E}[u]$ for the set $(0, 1, 0, 0)$. In the following, increase successively the polynomial degrees in each dimension $r = 1, \ldots, 4$ until

$$\frac{\|E_i - E_{i-1}\|_{L^2(D)}}{\|E_{i-1}\|_{L^2(D)}} < \text{TOL}$$

for all possible steps i. Doing this for $\text{TOL} = 5 \cdot 10^{-4}$, we end up with the set $(3, 1, 1, 1)$, i.e. we have $P_1 = 3$, $P_2 = P_3 = P_4 = 1$, which means $P + 1 = 32$. Consequently, the total number of degrees of freedom is $366\,025 \times 32 = 11\,712\,800$ for each time step.

Figures 5.7–5.14 show the results of the time evolution until $T = 15\,000\,[s]$ at selected time steps. The left pictures always display the expectation value $\mathbb{E}[p]$ and the right pictures the variance $\text{Var}[p]$. We changed the angle of view and cut the spatial domain into slices to focus on the relevant processes in the soil.

One can observe in the left pictures that the infiltration mainly takes place below the riverbed (where the maximal pressure difference can be found) and that this infiltration proceeds in vertical direction due to the gravitational parts in the Richards equation. For $t > 5\,000$, the soil below the riverbed is fully saturated and water flows in adjacent regions (in horizontal direction). One can also detect an infiltration in the remaining parts of the domain outside the riverbed, which takes place on a larger time scale and which is depending on the size of the permeability (it is higher in the southern part of the domain, here at the right-hand side of the pictures). Looking at the evolution of the variance in the right pictures, it is clear that it is zero at the Dirichlet boundaries and in the region where the pressure remains equal to -5. Apart from that, the variance is also small in regions which are already saturated, e.g. in the soil below the riverbed for $t > 5\,000$, while we observe the greatest variance near the saturation front, where the determination whether the soil is saturated or not strongly depends on the randomness of the stochastic permeability K. A comparison of the variance for $t = 1\,000$ in Figure 5.10 and $t = 10\,000$ in Figure 5.13 also suggests that the variance is higher the faster the saturation front is moving. This answers questions (Q1) and (Q2) from the introduction.

This variability near the saturation front and the resulting asymmetry of the distribution can also be measured by the higher moments $\gamma_1[p]$ and $\gamma_2[p]$, see Figure 5.15 for $t = 15\,000$. The reason for the special form of these moments is elucidated

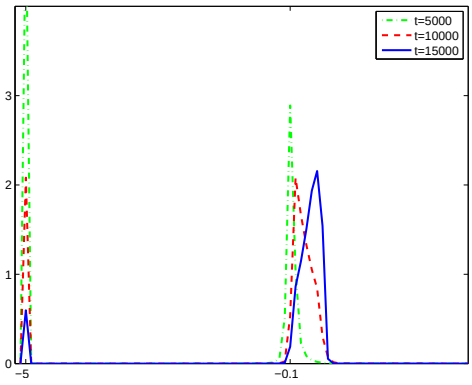


Figure 5.16: $y \mapsto \mathrm{pdf}_{\bar{x}}^{p}(y)$

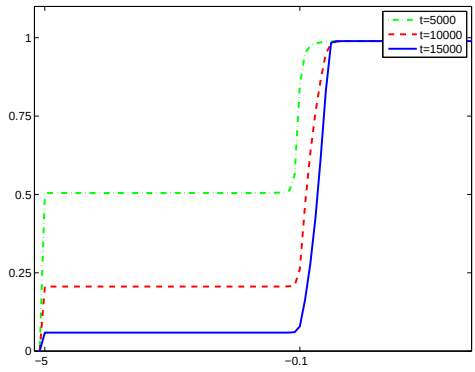


Figure 5.17: $y \mapsto \mathrm{cdf}_{\bar{x}}^{p}(y)$

if we plot the probability density function at a point $\bar{x}$, which is located $10\,m$ south of the river bank and approximately $3.50\,m$ below the surface. As shown in Figure 5.16, where the solid blue line indicates $t = 15\,000$, the density function has two peaks, which corresponds to the fact that the soil can be still unsaturated (with a pressure $p(\bar{x}, \omega) \approx -5$) or already saturated (with a pressure $p(\bar{x}, \omega) > -0.1$). The value $\mathbb{P}(p(\bar{x}, \cdot) \in [-4.9, -0.1]) < 0.03$ supports the observation that the infiltration proceeds very fast at the saturation front with corresponding steep gradients in spatial direction.

We are now ready to answer question (Q3) from the introduction. Observe first that the expectation value alone—which can also be computed by Monte Carlo methods—does not provide information about this task, since we obtain

$$\begin{aligned}
\mathbb{E}[p](t = 5\,000, \bar{x}) &= -3.604, \\
\mathbb{E}[p](t = 10\,000, \bar{x}) &= -0.650, \\
\mathbb{E}[p](t = 15\,000, \bar{x}) &= -0.157,
\end{aligned}$$

which could suggest that the soil is still unsaturated. Consulting the cumulative distribution functions $\mathrm{cdf}_{\bar{x}}^{p}(\cdot)$ in Figure 5.17 for $t \in \{5\,000, 10\,000, 15\,000\}$, which are generated by the representation of $p(t, x, \omega)$ in the stochastic domain by means of polynomial chaos, however, one can read

$$\begin{aligned}
&\mathbb{P}(\text{soil saturated in } \bar{x} \text{ at } t = 5\,000) \\
&\qquad = \mathbb{P}(p(5\,000, \bar{x}, \cdot) \geq -0.1) \\
&\qquad = 1 - \mathrm{cdf}_{\bar{x}}^{p(5\,000, \bar{x}, \cdot)}(-0.1) \approx 43.6\%, \\
&\mathbb{P}(\text{soil saturated in } \bar{x} \text{ at } t = 10\,000) \approx 79.1\%, \\
&\mathbb{P}(\text{soil saturated in } \bar{x} \text{ at } t = 15\,000) \approx 94.0\%
\end{aligned}$$

immediately. Finally, question (Q4) can be answered in the same way.

These results show that the methods developed in this thesis can be applied to realistic hydrological problems.

Appendix A

Exponential covariance

We consider the exponential covariance kernel

$$V_K(x_1, x_2) = \exp(-|x_1 - x_2|/\gamma) \tag{A.1}$$

with a parameter $\gamma > 0$. This kernel represents Markov processes and is used extensively to model processes in various fields [115, Section 31]. The parameter γ denotes the *correlation length* since the correlation between two points x_1, $x_2 \in D$ attenuates rapidly for small γ. In the limit $\gamma \to 0$, the corresponding process tends to white noise [115, Section 14].

We derive the eigenvalues and eigenfunctions of the operator (3.1.13) for the kernel (A.1) in the one-dimensional case $D = (-a, a)$, i.e. solve

$$\int_{-a}^{a} e^{-|x_1 - x_2|/\gamma} g_r(x_2)\,\mathrm{d}x_2 = \lambda_r g_r(x_1).$$

We split up the integral according to

$$\int_{-a}^{x} e^{-(x-x_2)/\gamma} g_r(x_2)\,\mathrm{d}x_2 + \int_{x}^{a} e^{(x-x_2)/\gamma} g_r(x_2)\,\mathrm{d}x_2 = \lambda_r g_r(x) \tag{A.2}$$

and differentiate twice with respect to x to obtain

$$-\frac{1}{\gamma}\int_{-a}^{x} e^{-(x-x_2)/\gamma} g_r(x_2)\,\mathrm{d}x_2 + \frac{1}{\gamma}\int_{x}^{a} e^{(x-x_2)/\gamma} g_r(x_2)\,\mathrm{d}x_2 = \lambda_r g_r'(x) \tag{A.3}$$

and

$$\frac{1}{\gamma^2}\lambda_r g_r(x) - \frac{2}{\gamma} g_r(x) = \lambda_r g_r''(x). \tag{A.4}$$

Introducing the new variables

$$\vartheta_r^2 := \frac{2}{\gamma\lambda_r} - \frac{1}{\gamma^2}, \tag{A.5}$$

we can write (A.4) as

$$g_r''(x) + \vartheta_r^2 g_r(x) = 0.$$

All solutions of this ordinary differential equation have the form

$$g_r(x) = \alpha_r \cos(\vartheta_r x) + \beta_r \sin(\vartheta_r x), \tag{A.6}$$

and we take the boundary conditions

$$g_r'(a) + \frac{1}{\gamma} g_r(a) = 0, \quad g_r'(-a) - \frac{1}{\gamma} g_r(-a) = 0 \tag{A.7}$$

derived from (A.2) and (A.3) into account. Combining (A.6) and (A.7) and dividing the equations by the factor $\cos(\vartheta_r a)$, we deduce that the (normalized) eigenfunctions g_r are given by

$$g_r(x) = \begin{cases} \frac{\sin(\vartheta_r x)}{\sqrt{a - \frac{\sin(2\vartheta_r a)}{2\vartheta_r}}} & \text{if } r \text{ is even} \\ \frac{\cos(\vartheta_r x)}{\sqrt{a + \frac{\sin(2\vartheta_r a)}{2\vartheta_r}}} & \text{if } r \text{ is odd,} \end{cases} \tag{A.8}$$

where ϑ_r are the solutions of the transcendental equations

$$\vartheta_r + \frac{1}{\gamma} \tan(\vartheta_r a) = 0 \quad \text{for } r \text{ even} \tag{A.9}$$

$$\frac{1}{\gamma} - \vartheta_r \tan(\vartheta_r a) = 0 \quad \text{for } r \text{ odd.} \tag{A.10}$$

The corresponding eigenvalues are given by

$$\lambda_r = \frac{2}{\gamma(\vartheta_r^2 + \frac{1}{\gamma^2})}$$

due to (A.5). We can easily notice from (A.9) and (A.10) that

$$\vartheta_r \in \left[\frac{(r-1)\pi}{2a}, \frac{r\pi}{2a} \right] \quad \forall r = 1, 2, \ldots$$

such that we can estimate

$$\lambda_r \leq \frac{2}{\gamma \left(\left(\frac{(r-1)\pi}{2a} \right)^2 + \frac{1}{\gamma^2} \right)} = \frac{8a^2\gamma}{((r-1)\gamma\pi)^2 + (2a)^2} \tag{A.11}$$

and

$$\lambda_r \geq \frac{2}{\gamma \left(\left(\frac{r\pi}{2a} \right)^2 + \frac{1}{\gamma^2} \right)} = \frac{8a^2\gamma}{(r\gamma\pi)^2 + (2a)^2}. \tag{A.12}$$

These calculations can be adapted to arbitrary intervals $D = (a, b)$ by translating and to d dimensions by taking products if the covariance kernel remains separable, e.g. in the case

$$V_K(x, \tilde{x}) = \exp\left(- \sum_{i=1}^{d} |x_i - \tilde{x}_i| / \gamma_i \right)$$

for $x = (x_1, \ldots, x_d)$, $\tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_d) \in D = \prod_{i=1}^{d} (a_i, b_i) \subset \mathbb{R}^d$.

Appendix B

Orthogonal polynomials

We summarize the definitions and properties of some important orthogonal polynomials. For theoretical background, we refer to the monograph [41].

Denote by $\{Q_k(x)\}$ the set of orthogonal polynomials with regard to a nonnegative integrable function on $\mathbb{R}$ called the *weight function* $w : \mathbb{R} \to \mathbb{R}$, i.e.

$$\int_I Q_k(x)Q_l(x)w(x)\,\mathrm{d}x = e_k\delta_{kl},$$

where $I \subset \mathbb{R}$ is the support of w.

The polynomials Q_k are given by the three-term recurrence relation

$$Q_{k+1}(x) = (a_k x + b_k)Q_k(x) - c_k Q_{k-1}(x) \tag{B.1}$$

with initial conditions $Q_{-1}(x) = 0$ and $Q_0(x) = 1$. Another possibility to determine the "classical" polynomials is Rodrigues' formula

$$Q_k(x) = \frac{1}{d_k w(x)} \frac{\mathrm{d}^k}{\mathrm{d}x^k} \left(w(x) \left(P(x)\right)^k \right)$$

with a polynomial $P(x)$ which is at most quadratic and standardization factors d_k.

B.1 Hermite polynomials

Consider the *normal distribution* $\mathcal{N}(0, 1)$ with density function

$$w(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \tag{B.2}$$

The corresponding orthogonal polynomials $\{Q_k\} = \{\mathrm{H}_k\}$ are called the *Hermite polynomials* given by the recurrence relation

$$\mathrm{H}_{k+1}(x) = x\mathrm{H}_k(x) - k\mathrm{H}_{k-1}(x)$$

or the formula

$$\mathrm{H}_k(x) = (-1)^k e^{x^2/2} \frac{\mathrm{d}^k}{\mathrm{d}x^k} e^{-x^2/2}.$$

They have the support $I = \mathbb{R}$ and satisfy

$$\int_{-\infty}^{\infty} \mathrm{H}_k(x)\mathrm{H}_l(x)w(x)\,\mathrm{d}x = k!\delta_{kl}.$$

The polynomials up to order six read

$$\begin{aligned}
\mathrm{H}_0(x) &= 1,\\
\mathrm{H}_1(x) &= x,\\
\mathrm{H}_2(x) &= x^2 - 1,\\
\mathrm{H}_3(x) &= x^3 - 3x,\\
\mathrm{H}_4(x) &= x^4 - 6x^2 + 3,\\
\mathrm{H}_5(x) &= x^5 - 10x^3 + 15x,\\
\mathrm{H}_6(x) &= x^6 - 15x^4 + 45x^2 - 15.
\end{aligned}$$

For the normal distribution $\mathscr{N}(\mu, \sigma^2)$ with expectation μ, variance σ^2 and density function

$$w(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{B.3}$$

scale the Hermite polynomials

$$\mathrm{H}_k^{(\mu,\sigma^2)}(x) = \mathrm{H}_k\left(\frac{x-\mu}{\sigma}\right).$$

B.2 Legendre polynomials

Consider the *uniform distribution* $\mathscr{U}(-1, 1)$ with density function

$$w(x) = \frac{1}{2}$$

on $I = [-1, 1]$. The corresponding orthogonal polynomials $\{Q_k\} = \{\mathrm{Le}_k\}$ are called the *Legendre polynomials* given by the recurrence relation

$$\mathrm{Le}_{k+1}(x) = \frac{2k+1}{k+1} x\mathrm{Le}_k(x) - \frac{k}{k+1}\mathrm{Le}_{k-1}(x)$$

or the formula

$$\mathrm{Le}_k(x) = \frac{1}{2^k k!} \frac{\mathrm{d}^k}{\mathrm{d}x^k}\left((x^2-1)^k\right).$$

The orthogonality condition reads

$$\int_{-1}^{1} \mathrm{Le}_k(x)\mathrm{Le}_l(x)w(x)\,\mathrm{d}x = \frac{1}{2k+1}\delta_{kl},$$

and the first polynomials up to order six are given by

$$\begin{aligned}
\mathrm{Le}_0(x) &= 1,\\
\mathrm{Le}_1(x) &= x,\\
\mathrm{Le}_2(x) &= \frac{1}{2}(3x^2 - 1),\\
\mathrm{Le}_3(x) &= \frac{1}{2}(5x^3 - 3x),\\
\mathrm{Le}_4(x) &= \frac{1}{8}(35x^4 - 30x^2 + 3),\\
\mathrm{Le}_5(x) &= \frac{1}{8}(63x^5 - 70x^3 + 15x),\\
\mathrm{Le}_6(x) &= \frac{1}{16}(231x^6 - 315x^4 + 105x^2 - 5).
\end{aligned}$$

For the uniform distribution $\mathscr{U}(a,b)$ on the more general support $I=[a,b]$, take the density

$$w(y)=\frac{1}{b-a}$$

and scale the polynomials by $y=\frac{b-a}{2}x+\frac{a+b}{2}$.

B.3 Jacobi polynomials

Consider the *beta distribution* $\mathscr{B}(\alpha,\beta;-1,1)$ with density function

$$w(x)=\frac{\Gamma(\alpha+\beta+2)}{2^{\alpha+\beta+1}\Gamma(\alpha+1)\Gamma(\beta+1)}(1-x)^{\alpha}(1+x)^{\beta} \tag{B.4}$$

on the support $I=[-1,1]$. The parameters α, $\beta>-1$ determine the shape of the distribution and $\Gamma(\cdot)$ denotes the gamma function interpolating the factorial. The name of the distribution derives from the beta function

$$B(x):=\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$$

used as the normalization factor in (B.4). The corresponding orthogonal polynomials $\{Q_k\}=\{\mathrm{Ja}_k^{(\alpha,\beta)}\}$ are called the *Jacobi polynomials* given by the recurrence relation (B.1) with

$$\begin{aligned}
a_k&=\frac{(2k+\alpha+\beta+1)(2k+\alpha+\beta+2)}{2(k+1)(k+\alpha+\beta+1)},\\
b_k&=-\frac{\beta^2-\alpha^2}{(2k+\alpha+\beta)(2k+\alpha+\beta+2)}a_k,\\
c_k&=\frac{2(k+\alpha)(k+\beta)}{(2k+\alpha+\beta)(2k+\alpha+\beta+1)}a_k
\end{aligned}$$

or by Rodrigues' formula

$$(1-x)^{\alpha}(1+x)^{\beta}\mathrm{Ja}_k^{(\alpha,\beta)}(x)=\frac{(-1)^k}{2^k k!}\frac{\mathrm{d}^k}{\mathrm{d}x^k}\left((1-x)^{k+\alpha}(1+x)^{k+\beta}\right).$$

Then, we obtain the orthogonality condition

$$\begin{aligned}
&\int_{-1}^{1}\mathrm{Ja}_k^{(\alpha,\beta)}(x)\mathrm{Ja}_l^{(\alpha,\beta)}(x)w(x)\,\mathrm{d}x\\
&\qquad=\frac{\Gamma(\alpha+k+1)\Gamma(\beta+k+1)\Gamma(\alpha+\beta+2)}{k!(2k+\alpha+\beta+1)\Gamma(\alpha+\beta+k+1)\Gamma(\alpha+1)\Gamma(\beta+1)}\delta_{kl}.
\end{aligned}$$

For $\alpha=\beta=0$, we recover the Legendre polynomials, more precisely we have $\mathrm{Ja}_k^{(0,0)}=\mathrm{Le}_k$.

For the beta distribution $\mathscr{B}(\alpha,\beta;a,b)$ on the more general support $I=[a,b]$, take the density

$$w(y)=\frac{\Gamma(\alpha+\beta+2)}{(b-a)^{\alpha+\beta+1}\Gamma(\alpha+1)\Gamma(\beta+1)}(b-y)^{\alpha}(y-a)^{\beta}$$

and scale the polynomials by $y=\frac{b-a}{2}x+\frac{a+b}{2}$.

B.4 Laguerre polynomials

Consider the *gamma distribution* $\mathscr{G}(\alpha)$ with density function

$$w(x) = \frac{e^{-x}}{\Gamma(\alpha+1)} x^{\alpha}$$

on the support $I = [0, \infty)$. The parameter $\alpha > -1$ determines the shape of the distribution and $\Gamma(\cdot)$ denotes the gamma function interpolating the factorial. The corresponding orthogonal polynomials $\{Q_k\} = \{\mathrm{La}_k^{(\alpha)}\}$ are called the *Laguerre polynomials* given by the recurrence relation

$$\mathrm{La}_{k+1}^{(\alpha)}(x) = \frac{2k+\alpha+1-x}{k+1}\mathrm{La}_k^{(\alpha)}(x) - \frac{k+\alpha}{k+1}\mathrm{La}_{k-1}^{(\alpha)}(x)$$

or by Rodrigues' formula

$$\mathrm{La}_k^{(\alpha)}(x) = \frac{e^x x^{-\alpha}}{k!}\frac{\mathrm{d}^k}{\mathrm{d}x^k}\left(e^{-x}x^{k+\alpha}\right).$$

The Laguerre polynomials are orthogonal according to

$$\int_0^{\infty} \mathrm{La}_k^{(\alpha)}(x)\mathrm{La}_l^{(\alpha)}(x)w(x)\,\mathrm{d}x = \frac{1}{k!}\frac{\Gamma(\alpha+k+1)}{\Gamma(\alpha+1)}\delta_{kl},$$

and the first polynomials up to order 3 are given by

$$\begin{aligned}
&\mathrm{La}_0^{(\alpha)}(x) = 1,\\
&\mathrm{La}_1^{(\alpha)}(x) = -x+\alpha+1,\\
&\mathrm{La}_2^{(\alpha)}(x) = \frac{1}{2}\left(x^2 - 2(\alpha+2)x + (\alpha+2)(\alpha+1)\right),\\
&\mathrm{La}_3^{(\alpha)}(x) = \frac{1}{6}(-x^3 + 3(\alpha+3)x^2 - 3(\alpha+2)(\alpha+3)x + (\alpha+1)(\alpha+2)(\alpha+3)).
\end{aligned}$$

If $X \propto \mathscr{G}(\alpha)$, then it has the expectation value $\mathbb{E}[X] = \alpha+1$. In order to satisfy condition (3.1.19) in the Karhunen–Loève expansion, one can scale $y = x - (\alpha+1)$. Note that the gamma distribution is bounded from below and hence a practical approximation of the normal distribution since the distribution of $(X - \mathbb{E}[X])$ converges to $\mathscr{N}(0, \alpha+1)$.

Appendix C

Gaussian quadrature

We recall the notation from Appendix B. We assume that the moments

$$m_n := \int_{\mathbb{R}} x^n \, w(x) \, \mathrm{d}x \tag{C.1}$$

exist and are finite. Moreover, for sake of simplicity, let the (possibly unbounded) support I of the weight function $w(x)$ be a connected interval.

In order to approximate the integral

$$\int_I f(x) \, w(x) \, \mathrm{d}x, \tag{C.2}$$

we take as usual the sum

$$\sum_{i=1}^{k} f(\pi_i^{(k)}) \eta_i^{(k)} \tag{C.3}$$

with quadrature weights η_i given by

$$\eta_i^{(k)} = \int_I \mathcal{L}_i^{(k)}(x) \, w(x) \, \mathrm{d}x,$$

where $\pi_i^{(k)}$ are certain points in the interval I and $\mathcal{L}_i^{(k)}$ are the Lagrange polynomials with $\mathcal{L}_i^{(k)}(\pi_j^{(k)}) = \delta_{ij}$. The Gaussian quadrature now consists in choosing as evaluation points $\pi_i^{(k)}$ the zeros of Q_k, i.e. of the orthogonal polynomial with respect to $w(\cdot)$. This is possible due to the following theorem.

Theorem C.1 ([41])**.** *All zeros $\pi_i^{(k)}$ of Q_k are real, simple, and located in the interior of the support interval I. Furthermore, the zeros of Q_{k+1} alternate with those of Q_k, i.e.*

$$\pi_1^{(k+1)} < \pi_1^{(k)} < \pi_2^{(k+1)} < \pi_2^{(k)} < \ldots < \pi_k^{(k)} < \pi_{k+1}^{(k+1)}.$$

It is well known that the weights $\eta_i^{(k)}$ are always positive and that this quadrature has order $2(k+1)$, i.e. it computes polynomials of degree $2k+1$ exactly, see [97, Theorem 3.6.12]. To perform the Gaussian quadrature, we only need to know the zeros and the weights. They can be computed directly from the orthogonal polynomials, see [31, Subsection 9.3.2], or by the recursion formula (B.1) with the coefficients of

the corresponding orthogonal polynomials. For the classical polynomial sets, these values are tabulated [98] or can be computed by codes like GAUSSQ [1].

For the next results, we concentrate on weight functions which satisfy the moment estimate

$$m_n < C\, R^{2n}\, (2n+1)! \tag{C.4}$$

with constants C and R. The first useful theorem states that the zeros of the orthogonal polynomials are "dense" in I.

Theorem C.2 ([105]). *Let condition (C.4) hold for certain constants C and R. Then for any given interval $(a,b) \subset I$, there is a number N such that there are quadrature points $\pi_i^{(k)}$ with $\pi_i^{(k)} \in (a,b)$ for all $k > N$.*

The next theorem states under which conditions on the integrand f we achieve convergence of the quadrature.

Theorem C.3 ([105]). *If the moments satisfy condition (C.4) for certain constants C and R, then the quadrature formula (C.3) converges to (C.2) for any integrable function f satisfying the inequality (for sufficiently large x)*

$$|f(x)| < \frac{e^{r|x|}}{|x|^{1+\alpha}},$$

where $r = 1/R$ and $0 < \alpha < 1$.

In special cases, we can say even more. According to Uspensky [105], we obtain convergence if

$$|f(x)| < \frac{e^{x^2}}{|x|^{1+\alpha}}, \quad 0 < \alpha < 1,$$

for the weight function $w(x) = e^{-x^2}$ or, more generally,

$$|f(x)| < \frac{e^{\frac{x^2}{2\sigma^2}}}{|x|^{1+\alpha}}, \quad 0 < \alpha < 1,$$

for the density (B.3) of $\mathcal{N}(0, \sigma^2)$.

List of Symbols

Miscellaneous

$(\cdot,\cdot)_{0,0}$, $(\cdot,\cdot)_{1,0}$	scalar product on tensor space	20
$(\cdot,\cdot)_X$	scalar product on X	17
$\langle\cdot,\cdot\rangle$	Euclidean scalar product	107
${}_{V'}\langle\cdot,\cdot\rangle_V$	duality bracket	14
$\lvert\cdot\rvert_p$	p-norm in $\mathbb{R}^d$	82
$\lVert\cdot\rVert_{0,0}$, $\lVert\cdot\rVert_{1,0}$	norm on tensor space	20
$\lVert\cdot\rVert_X$	norm on X	
$A \otimes B$	operator tensor product	21
$\mathsf{E} \otimes \mathsf{F}$	Kronecker product	87
$x \otimes y$	tensor product	18
$X \otimes Y$	tensor space	18
$\mathbf{1}(\cdot)$	constant one function	47
$\mathbf{1}_X(\cdot)$	indicator function on X	41
2^X	power set of X	16

Greek letters

γ	covariance parameter	34, 131
$\gamma_1[X]$	skewness of random variable X	118
$\gamma_2[X]$	kurtosis of random variable X	118
Γ_D	Dirichlet boundary	12
Γ_N	Neumann boundary	12
δ_{ij}	Kronecker delta	
∂D	boundary of D	12
$\partial_v F$	directional derivative of F	47
$\partial F(\cdot)$	subdifferential of F	54
η	viscosity	8
η_π, η_{π_i}	quadrature weights	62
$\theta(\cdot)$	saturation	7
θ_m, θ_M	minimal/maximal saturation	8
$\vartheta^\nu_{i,\pi}$, ϑ^ν_i	damping factor	112
$\kappa(\cdot)$	Kirchhoff transformation	8
λ	pore size distribution factor	8
λ_r	eigenvalues in the KL expansion	32
$\Lambda_\jmath$, $\Lambda^D_\jmath$	nodal basis on level $\jmath$	64
$\Lambda_{\mathcal{S}}$	multilevel nodal basis	111
μ^X_n	central moment of order n of random variable X	117
μ_π	eigenvalue of $\bar{\mathsf{A}}$	99

ξ_r, ξ	random variables from the Karhunen–Loève expansion	32, 38
π, π_i	quadrature points	62
Π	cardinality of $\mathcal{Q}^P$	62
ρ	density of the water	7
ϱ, ϱ^π	convergence rate	115
$\sigma(\cdot), \sigma_r(\cdot)$	positive weight function	67
$\sigma(X)$	sigma-algebra generated by random variable X	16
τ, τ_n	time step size	14, 44
$\hat{\phi}(\cdot)$, $\phi(\cdot)$	convex functional	15, 47
$\phi^y(\cdot)$, $\phi^\pi(\cdot)$	convex functional	90
$\phi_j^P(\cdot)$, $\phi_{\bar{u}_j^{P,\nu}}(\cdot)$	discrete convex functional	65, 110
$\phi_j^y(\cdot)$, $\phi_j^\pi(\cdot)$	discrete convex functional	92
$\underline{\varphi}_{\bar{u}_j^{P,\nu}}(\cdot)$, $\overline{\varphi}_{\bar{u}_j^{P,\nu}}(\cdot)$	constraints in the Newton linearization	109
$\Phi(\cdot)$, $\hat{\Phi}(\cdot)$	antiderivative of H, convex function	10, 15, 46
χ_E	characteristic function of the set E	109
$\underline{\psi}_i^\nu(\cdot)$, $\overline{\psi}_i^\nu(\cdot)$	constraints in the Newton linearization	112
$\psi_k^r(\cdot)$, $\Psi_k(\cdot)$	orthogonal polynomials	58, 59
$\Psi(\cdot)$	obstacle function	72
$\{\Psi_k^c\}$, $\{\Psi_k^t\}$, $\{\Psi_k^b\}$	polynomial chaos basis	60, 61
ω	sample in Ω	17
Ω	sample space	16
$\Omega^{(M)}$	stochastic space	38
Ω_r	stochastic space $\Omega_r = \xi_r(\Omega)$	38

Roman letters

$\hat{a}(\cdot,\cdot)$, $a(\cdot,\cdot)$, $a_{\bar{u}_j^{P,\nu}}(\cdot,\cdot)$	bilinear form	15, 45, 110
$a^y(\cdot,\cdot), a^\pi(\cdot,\cdot), a^\pi_{\bar{u}_j^{P,\nu}}(\cdot,\cdot)$	bilinear form	90, 114
A, $\bar{\mathsf{A}}$, A^π	stiffness matrix	87, 98, 102
b, b^π	coefficient vector	88, 102
B, $\tilde{\mathsf{B}}$, $\underline{\mathsf{B}}$	local evaluation matrix	98, 103, 105, 122
$\mathscr{B}(\alpha,\beta;a,b)$	beta distribution	135
$\mathrm{Bor}(X)$	Borel sets on X	17
c_π	row of B	99
$\mathrm{cdf}_x^u(\cdot)$, $\mathrm{cdf}_x^p(\cdot)$	cumulative distribution function of random variable u or p	119
$\mathcal{C}$	collocation point set	90
$C^0_\sigma(\Omega^{(M)})$, $C^0_{\sigma,0}(\Omega^{(M)})$	space of continuous, bounded (and decreasing) functions on $\Omega^{(M)}$ w.r.t. σ	67
$C_0^\infty(X)$	space of smooth functions with compact support	13, 20
$C^k(X)$	space of k times continuously differentiable functions	
$C_K(\cdot,\cdot)$	correlation of K	30
$\mathcal{C}_K$	correlation operator w.r.t. K	31
$\mathcal{C}_j^{P,\nu}$	coarse grid correction	108
d	spatial dimension	12
$\mathrm{dom}\, A$	domain of an operator (Chapter 2)	21
$\mathrm{dom}\, F$	effective domain of a functional (Ch. 3)	50
$\mathrm{dom}\, \partial F$	domain of a subdifferential	54

D	spatial domain	12
D	special diagonal matrix	100
$\mathscr{D}$	tensor space of smooth functions with compact support	20
$\mathcal{D}_i^\nu$	constraint set	111
$e(\lambda)$	Brooks–Corey factor	8
e_z	vector in direction of gravity	12
$\mathcal{E}$	set of events	16
$\mathbb{E}[X]$	expectation of random variable X	17
$f(\cdot)$	source function, right-hand side function	8, 26
$f_N(\cdot)$	function on Γ_N	12, 26
g	gravitational constant	7
$g_r(\cdot)$	eigenfunctions of the KL expansion	32
$\mathscr{G}(\alpha)$	gamma distribution	136
h	piezometric head	7
$h_\jmath$	mesh size on level $\jmath$	63
h_p	weighting factor to a node p	65
$H(\cdot)$, $\tilde{H}(\cdot)$	generalized saturation	9, 55
H	scaling matrix	102
H_0, H_∞	limit cases of H	11
$H^{1/2}(\Gamma)$	trace space	13
$H^1_{\Gamma_D}(D)$	Sobolev space with homogeneous Dirichlet boundary conditions on Γ_D	45
$H^1(D;L^2(X))$	Sobolev space of functions $D \to L^2(X)$	20
$H^k(X)$, $H^k_0(X)$	Sobolev space	13
$\mathrm{H}_k(\cdot)$	Hermite polynomials	133
I	identity matrix	61, 99
$\mathscr{I}^P$	Lagrange interpolant operator	63
$\mathcal{I}_{\mathcal{S}_\jmath}$	linear interpolation operator	66
$\mathcal{J}(\cdot)$, $\mathcal{J}_{\bar{u}_\jmath^{P,\nu}}(\cdot)$	quadratic functional	49, 110
$\mathcal{J}^y(\cdot)$, $\mathcal{J}^\pi(\cdot)$, $\mathcal{J}^\pi_{\bar{u}_\jmath^{P,\nu}}(\cdot)$	quadratic functional	92, 114
$\mathrm{Ja}_k^{(\alpha,\beta)}(\cdot)$	Jacobi polynomials	135
$kr(\cdot)$	relative permeability	8
$K(\cdot)$	permeability	8, 22, 42
$\bar{K}(\cdot)$	expectation of permeability $K(\cdot)$	30
$\bar{\mathcal{K}}$	convex set (obstacle problem)	72
$\hat{\mathcal{K}}$, $\mathcal{K}$, $\mathcal{K}_{\Gamma_D}$	convex set	13, 25, 39, 45, 53
K_c	hydraulic conductivity	7
$K_M(\cdot)$	truncated KL expansion of $K(\cdot)$	33
$K_{\min}$, $K_{\max}$	bounds of permeability $K(\cdot)$	12
$\mathcal{K}_\jmath^P$, $\mathcal{K}_{\bar{u}_\jmath^{P,\nu}}$	discrete convex set	64, 110
$\mathcal{K}_\jmath^y$, $\mathcal{K}_\jmath^\pi$	discrete convex set	92
$\hat{\ell}(\cdot)$, $\ell(\cdot)$, $\ell_{\bar{u}_\jmath^{P,\nu}}(\cdot)$	linear functional	15, 46, 110
$\ell^y(\cdot)$, $\ell^\pi(\cdot)$, $\ell^\pi_{\bar{u}_\jmath^{P,\nu}}(\cdot)$	linear functional	90, 114
$L^2(0,T;X)$	L^2 space of functions $(0,T) \to X$	13
$L^2(X,\mu;H)$, $L^2(X,\mu)$	L^2 space w.r.t. measure μ	19
$L^p(X)$	space of Lebesgue integrable functions	13, 17
$\mathcal{L}_{\pi_i}(\cdot)$, $\mathcal{L}_{r,\pi_i^r}(\cdot)$, $\mathcal{L}_i^{(k)}(\cdot)$	Lagrange polynomials	62, 137
$\mathrm{La}_k^{(\alpha)}(\cdot)$	Laguerre polynomials	136
$\mathrm{Le}_k(\cdot)$	Legendre polynomials	134
m_n^X	moment of order n of random variable X	117
M	length of Karhunen–Loève expansion	33

$\mathcal{M}_\jmath^P$, $\mathcal{M}_\jmath^\pi$	Block Gauß–Seidel iteration operator	96, 102
$\mathbf{n}$	outer normal	12
$\mathfrak{n}(\cdot)$	porosity of the soil	8
$\mathscr{N}(\mu, \sigma^2)$	normal distribution	133
$N_\jmath$	cardinality of $\mathcal{N}_\jmath$	63
$\mathcal{N}_\jmath$, $\mathcal{N}_\jmath^D$	set of nodes on level $\jmath$	63, 64
$\mathcal{N}_\jmath^\bullet(v)$, $\mathcal{N}_\jmath^\circ(v)$	set of critical/regular nodes	109
N_{MC}	number of Monte Carlo iterations	27, 93
p	node in the triangulation $\mathcal{T}_\jmath$	64
$p(\cdot)$	pressure	7, 22
$p^0(\cdot)$	initial condition	123
p_b	bubbling pressure	8
$\mathrm{pdf}(\cdot)$, $\mathrm{pdf}_r(\cdot)$	probability density function of ξ and ξ_r	38
$\mathrm{pdf}_x^u(\cdot)$, $\mathrm{pdf}_x^p(\cdot)$	probability density function of random variable u or p	119
$\mathbb{P}$	probability measure	16
P, P_0, P_r	size of PC basis	60
$\mathbb{P}^{(M)}$	probability measure on $\Omega^{(M)}$	38
$\mathcal{P}^P$	projection operator to $\mathcal{Z}^P$	59
$\mathbb{P}_X$	distribution of random variable X	17
$\mathrm{Pol}^k(X)$	space of polynomials of degree $\leq k$ defined on X	60
Q	time cylinder	13
$\mathcal{Q}^P$	quadrature point set	62
r_i^ν	multilevel basis function	107
$s_p^{(\jmath)}$, $s_i^{(\jmath)}$	nodal basis function on level $\jmath$	64, 86
$\mathrm{supp}(f)$	support of a function f	20
$\mathcal{S}_\jmath$, $\mathcal{S}_\jmath^D$	finite element space on level $\jmath$	64
t	time variable	7
t	triangle in $\mathcal{T}_\jmath$	63
t_n	time step	14, 44
tr_Γ	trace operator on Γ	13, 26
T	end time	8
$\mathcal{T}_\jmath$	triangulation of D on level $\jmath$	63
$u(\cdot)$	generalized pressure	9, 22
u	coefficient vector	88
$u^0(\cdot)$	initial condition	62
u_c	critical generalized pressure	9
$u_D(\cdot)$	function on Γ_D	12, 26
$u_\jmath^P$, $u_\jmath$	discrete solution	70, 90
$(u_\jmath^P)^\nu$, $\bar{u}_\jmath^{P,\nu}$	Block Gauß–Seidel iterate	95, 108
$\mathscr{U}(a, b)$	uniform distribution	134
v, $\bar{\mathsf{v}}$, v^π, $\mathfrak{v}$	coefficient vector	98, 98, 103, 105
$\mathbf{v}(\cdot)$	water flux	7, 12, 23
$V_K(\cdot, \cdot)$	covariance of K	30
$\mathcal{V}_K$	covariance operator w.r.t. K	33
$\mathrm{Var}[X]$	variance of random variable X	17, 118
w, $\mathfrak{w}$, w^π	coefficient vector	98, 100, 103
w_i^ν	intermediate iterate	95
x	point in D	7
z	vertical coordinate	7
$\mathcal{Z}^P$	polynomial chaos space	59

Bibliography

[1] GAUSSQ. http://www.netlib.org/go/.

[2] Milton Abramowitz and Irene A. Stegun, editors. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, 1972.

[3] Robert A. Adams. *Sobolev Spaces*. Academic Press, 1975.

[4] Robert J. Adler. *The Geometry of Random Fields*. Wiley, 1981.

[5] Hans Wilhelm Alt and Stephan Luckhaus. Quasilinear elliptic–parabolic differential equations. *Math. Z.*, 183:311–341, 1983.

[6] Hans Wilhelm Alt, Stephan Luckhaus, and Augusto Visintin. On nonstationary flow through porous media. *Ann. Math. Pura Appl.*, 136:303–316, 1984.

[7] Theodore W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, 2nd edition, 1984.

[8] Jean-Pierre Aubin and Hélène Frankowska. *Set-valued Analysis*. Birkhäuser, 1990.

[9] Ivo Babuška and Panagiotis Chatzipantelidis. On solving elliptic stochastic partial differential equations. *Comput. Methods Appl. Mech. Eng.*, 191:4093–4122, 2002.

[10] Ivo Babuška, Kang-Man Liu, and Raúl Tempone. Solving stochastic partial differential equations based on experimental data. *Math. Mod. Meth. Appl. S.*, 13(3):415–444, 2003.

[11] Ivo Babuška, Raúl Tempone, and Georgios E. Zouraris. Galerkin finite element approximations of stochastic elliptic partial differential equations. *SIAM J. Num. Anal.*, 42(2):800–825, 2004.

[12] Ivo Babuška, Raúl Tempone, and Georgios E. Zouraris. Solving elliptic boundary value problems with uncertain coefficients by the finite element method: the stochastic formulation. *Comput. Methods Appl. Mech. Engrg.*, 194(12–16): 1251–1294, 2005.

[13] Ivo Babuška, Fabio Nobile, and Raúl Tempone. A stochastic collocation method for elliptic partial differential equations with random input data. *SIAM J. Num. Anal.*, 45(3):1005–1034, 2007.

[14] Viorel Barbu. *Nonlinear Semigroups and Differential Equations in Banach Spaces*. Noordhoff International, 1976.

[15] Jacob Bear. *Dynamics of Fluids in Porous Media*. Dover, 1988.

[16] Ehrhard Behrends. *Maß- und Integrationstheorie*. Springer, 1987.

[17] Fred E. Benth and Jon Gjerde. Convergence rates for finite element approximations of stochastic partial differential equations. *Stoch. & Stoch. Reports*, 63:313–326, 1998.

[18] Heiko Berninger. *Domain Decomposition Methods for Elliptic Problems with Jumping Nonlinearities and Application to the Richards Equation*. PhD thesis, Freie Universität Berlin, 2007.

[19] Paul Besold. *Solutions to Stochastic Partial Differential Equations as Elements of Tensor Product Spaces*. PhD thesis, Georg-August-Universität zu Göttingen, 2000.

[20] Marcel Bieri and Christoph Schwab. Sparse high order FEM for elliptic sPDEs. *Comp. Meth. Appl. Mech. Engrg.*, 198(13-14):1149–1170, 2009.

[21] James H. Bramble and Xuejun Zhang. Uniform convergence of the multigrid V-cycle for an anisotropic problem. *Math. Comp.*, 70(234):453–470, 2001.

[22] Franco Brezzi and Gianni Gilardi. Functional spaces. In Hayrettin Kardestuncer and Douglas H. Norrie, editors, *Finite Element Handbook*, chapter 1.2. Springer, 1987.

[23] Robert H. Cameron and William T. Martin. The orthogonal development of nonlinear functionals in series of Fourier–Hermite functionals. *Ann. Math.*, 48(2):385–392, 1947.

[24] Haiyan Cheng and Adrian Sandu. Uncertainty quantification and apportionment in air quality models using the polynomial chaos method. *Envir. Model. & Softw.*, 24:917–925, 2009.

[25] Philippe G. Ciarlet. Basic error estimates for elliptic problems. In *Finite Element Methods (Part 1)*, volume II of *Handbook of Numerical Analysis*, pages 17–351. North-Holland, 1991.

[26] Philippe G. Ciarlet. *The Finite Element Method for Elliptic Problems*. North-Holland, 1978.

[27] Gedeon Dagan and David G. Zeitoun. Seawater–freshwater interface in a stratified aquifer of random permeability distribution. *J. Contam. Hydrol.*, 29(3):185–203, 1998.

[28] Henry Darcy. *Les fontaines publiques de la ville de Dijon*. Dalmont, 1856.

[29] Manas K. Deb, Ivo M. Babuška, and J. Tinsley Oden. Solution of stochastic partial differential equations using Galerkin finite element techniques. *Comput. Methods Appl. Mech. Engrg.*, 190:6359–6372, 2001.

[30] Bert J. Debusschere, Habib N. Najm, Philippe P. Pébay, Omar M. Knio, Roger G. Ghanem, and Olivier P. Le Maître. Numerical challenges in the use of polynomial chaos representations for stochastic processes. *SIAM J. Sci. Comput.*, 26(2):698–719, 2005.

[31] Peter Deuflhard and Andreas Hohmann. *Numerische Mathematik I: Eine algorithmisch orientierte Einführung*. Walter de Gruyter, 2nd edition, 1993.

[32] Michael Eiermann, Oliver G. Ernst, and Elisabeth Ullmann. Computational aspects of the stochastic finite element method. *Comput. Visual. Sci.*, DOI 10.1007/s00791-006-0046-5 (electronic), 2007.

[33] Ivar Ekeland and Roger Temam. *Convex Analysis and Variational Problems*. North-Holland, 1976.

[34] Howard Elman and Darran Furnival. Solving the stochastic steady-state diffusion problem using multigrid. Technical Report UMCP-CSD:CS-TR-4786, University of Maryland, 2006.

[35] William Feller. *An Introduction to Probability Theory and Its Applications*, volume II. Wiley, 2nd edition, 1971.

[36] Marek Fisz. *Probability Theory and Mathematical Statistics*. Wiley, 3rd edition, 1963.

[37] Ralf Forster and Ralf Kornhuber. A polynomial chaos approach to stochastic variational inequalities. to appear in J. Numer. Math.

[38] Philipp Frauenfelder, Christoph Schwab, and Radu A. Todor. Finite elements for elliptic problems with stochastic coefficients. *Comput. Methods Appl. Mech. Engrg.*, 194(2–5):205–228, 2005.

[39] R. Allan Freeze. A stochastic-conceptual analysis of one-dimensional groundwater flow in nonuniform homogeneous media. *Water Resour. Res.*, 11(5): 725–741, 1975.

[40] Benjamin Ganis, Hector Klie, Mary F. Wheeler, Tim Wildey, Ivan Yotov, and Dongxiao Zhang. Stochastic collocation method and mixed finite elements for flow in porous media. *Comput. Methods Appl. Mech. Engrg.*, 197:3547–3559, 2008.

[41] Walter Gautschi. *Orthogonal Polynomials: Computation and Approximation*. Oxford University Press, 2004.

[42] Carl Geiger and Christian Kanzow. *Theorie und Numerik restringierter Optimierungsaufgaben*. Springer, 2002.

[43] Israïl M. Gel'fand and Naum Y. Vilenkin. *Generalized Functions. Volume 4: Applications of Harmonic Analysis*. Academic Press, 1964.

[44] Lynn W. Gelhar. *Stochastic Subsurface Hydrology*. Prentice Hall, 1993.

[45] Martinus Th. van Genuchten. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Sci. Soc. Am. J.*, 44:892–898, 1980.

[46] Roger G. Ghanem. Probabilistic characterization of transport in heterogeneous media. *Comput. Methods Appl. Mech. Engrg.*, 158:199–220, 1998.

[47] Roger G. Ghanem and Pol D. Spanos. *Stochastic Finite Elements: A Spectral Approach*. Springer, 1991.

[48] Roland Glowinski. *Numerical Methods for Nonlinear Variational Problems*. Springer, 1984.

[49] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 3rd edition, 1996.

[50] J. Jaime Gómez-Hernández and Xian-Huan Wen. To be or not to be multi-Gaussian? A reflection on stochastic hydrogeology. *Adv. Water Resour.*, 21 (1):47–61, 1998.

[51] Carsten Gräser and Ralf Kornhuber. Multigrid methods for obstacle problems. *J. Comput. Math.*, 27(1):1–44, 2009.

[52] Carsten Gräser and Oliver Sander. Polyhedral Gauß–Seidel converges. Matheon Preprint 696, 2010.

[53] Mircea Grigoriu and Rich V. Field Jr. On the accuracy of the polynomial chaos approximation. *Probabilist. Eng. Mech.*, 19:65–80, 2004.

[54] Joachim Gwinner. A class of random variational inequalities and simple random unilateral boundary value problems: Existence, discretization, finite element approximation. *Stoch. Anal. Appl.*, 18:967–993, 2000.

[55] Joachim Gwinner and Fabio Raciti. On a class of random variational inequalities on random sets. *Numer. Func. Anal. Opt.*, 27:619–636, 2006.

[56] Wolfgang Hackbusch. *Multi-Grid Methods and Applications*. Number 4 in Springer Series in Computational Mathematics. Springer, 1985.

[57] Robert J. Hoeksema and Peter K. Kitanidis. Analysis of the spatial structure of properties of selected aquifers. *Water Resour. Res.*, 21(4):563–572, 1985.

[58] Helge Holden, Bernt Øksendal, Jan Ubøe, and Tusheng Zhang. *Stochastic Partial Differential Equations: A Modeling, White Noise Functional Analysis Approach*. Probability and Applications. Birkhäuser, 1996.

[59] Zhi-Yuan Huang and Jia-An Yan. *Introduction to Infinite Dimensional Stochastic Analysis*. Kluwer Academic Publishers and Science Press, 2000.

[60] Joseph W. Jerome. *Approximation of Nonlinear Evolution Systems*. Academic Press, 1983.

[61] Claes Johnson. *Numerical Solutions of Partial Differential Equations by the Finite Element Method*. Cambridge University Press, 1987.

[62] Kari Karhunen. Über lineare Methoden in der Wahrscheinlichkeitsrechnung. *Ann. Acad. Sci. Fennicae, Ser. A. I.*, 37:1–79, 1947.

[63] George Em Karniadakis, Chau-Hsing Su, Dongbin Xiu, Didier Lucor, Christoph Schwab, and Radu A. Todor. Generalized polynomial chaos solution for differential equations with random inputs. Research report no. 2005–01, ETH Zürich, 2005.

[64] Andreas Keese. *Numerical Solution of Systems with Stochastic Uncertainties*. PhD thesis, Technische Universität Braunschweig, 2003.

[65] Achim Klenke. *Wahrscheinlichkeitstheorie*. Springer, 2006.

[66] Hermann König. *Eigenvalue Distribution of Compact Operators*, volume 16 of *Operator Theory: Advances and Applications*. Birkhäuser, 1986.

[67] Konrad Königsberger. *Analysis 1*. Springer, 3rd edition, 1995.

[68] Ralf Kornhuber. *Adaptive Monotone Multigrid Methods for Nonlinear Variational Problems*. Teubner, 1997.

[69] Ralf Kornhuber. On constrained Newton linearization and multigrid for variational inequalities. *Numer. Math.*, 91:699–721, 2002.

[70] Wojtek J. Krzanowski. *Principles of Multivariate Analysis: A User's Perspective*. Manchester University Press, 1992.

[71] Hui-Hsiung Kuo. *White Noise Distribution Theory.* CRC Press, 1996.

[72] Olivier P. Le Maître and Omar Knio. *Spectral Methods for Uncertainty Quantification.* Springer, 2010.

[73] Olivier P. Le Maître, Omar M. Knio, Habib N. Najm, and Roger G. Ghanem. A stochastic projection method for fluid flow: I. Basic formulation. *J. Comp. Phys.*, 173(2):481–511, 2001.

[74] Olivier P. Le Maître, Matthew T. Reagan, Habib N. Najm, Roger G. Ghanem, and Omar M. Knio. A stochastic projection method for fluid flow: II. Random process. *J. Comp. Phys.*, 181(1):9–44, 2002.

[75] Olivier P. Le Maître, Omar M. Knio, Bert J. Debusschere, Habib N. Najm, and Roger G. Ghanem. A multigrid solver for two-dimensional stochastic diffusion equations. *Comput. Methods Appl. Mech. Engrg.*, 192:4723–4744, 2003.

[76] Thomas van Lent and Peter K. Kitanidis. A numerical spectral approach for the derivation of piezometric head covariance functions. *Water Resour. Res.*, 25(11):2287–2298, 1989.

[77] Yusong Li, Eugene LeBoeuf, Prodyot K. Basu, and Sankaran Mahadevan. Stochastic modeling of the permeability of randomly generated porous media. *Adv. Water Resour.*, 28(8):835–844, 2005.

[78] Michel Loève. *Probability Theory.* D. Van Nostrand, 3rd edition, 1963.

[79] Doron S. Lubinsky. A survey of weighted polynomial approximation with exponential weights. *Approx. Theory*, 3:1–105, 2007.

[80] Didier Lucor, Chau-Hsing Su, and George Em Karniadakis. Generalized polynomial chaos and random oscillators. *Int. J. Numer. Meth. Eng.*, 60:571–596, 2004.

[81] Lionel Mathelin and M. Yousuff Hussaini. A stochastic collocation algorithm for uncertainty analysis. Technical report, NASA/CR-2003-212153, 2003.

[82] Hermann G. Matthies and Andreas Keese. Galerkin methods for linear and nonlinear elliptic stochastic partial differential equations. *Comp. Meth. Appl. Mech. Engrg.*, 194:1295–1331, 2005.

[83] Edward L. Melnick and Aaron Tenenbein. Misspecifications of the normal distribution. *The American Statistician*, 36(4):372–373, 1982.

[84] Maria da Glória Bastos de Freitas Mesquita, Sérgio Oliveira Moraes, and José Eduardo Corrente. More adequate probability distributions to represent the saturated soil hydraulic conductivity. *Sci. agric.*, 59(4):789–793, 2002.

[85] Nicholas Metropolis. The beginning of the Monte Carlo method. *Los Alamos Science*, 15:125–130, 1987.

[86] Habib N. Najm, Bert J. Debusschere, Youssef M. Marzouk, Steve Widmer, and Olivier P. Le Maître. Uncertainty quantification in chemical systems. *Int. J. Numer. Meth. Engng.*, 80:789–814, 2009.

[87] Fabio Nobile, Raúl Tempone, and Clayton G. Webster. A sparse grid stochastic collocation method for partial differential equations with random input. *SIAM J. Numer. Anal.*, 46(5):2309–2345, 2008.

[88] Bernt Øksendal. *Stochastic Differential Equations*. Springer, 6th edition, 2007.

[89] Alfio Quarteroni and Alberto Valli. *Domain Decomposition Methods for Partial Differential Equations*. Oxford Science Publications, 1999.

[90] Malempati Madhusudana Rao. *Probability Theory with Applications*. Academic Press, 1984.

[91] Lorenzo A. Richards. Capillary conduction of liquids through porous mediums. *Physics*, 1:318–333, 1931.

[92] Carl P. Rupert and Cass T. Miller. An analysis of polynomial chaos approximations for modeling single-fluid-phase flow in porous medium systems. *J. Comp. Phys.*, 226:2175–2205, 2007.

[93] Youcef Saad. *Numerical Methods for Large Eigenvalue Problems*. Manchester University Press, 1992.

[94] Thomas Sauer and Yuan Xu. On multivariate Lagrange interpolation. *Math. Comp.*, 64(211):1147–1170, 1995.

[95] Christoph Schwab and Radu A. Todor. Karhunen–Loève approximation of random fields by generalized fast multipole methods. *J. Comp. Phys.*, 217(1): 100–122, 2006.

[96] Liang-Sheng Shi, Jin-Zhong Yang, Shu-Ying Cai, and Lin Lin. Stochastic analysis of groundwater flow subject to random boundary conditions. *J. Hydrodyn.*, 20(5):553–560, 2008.

[97] Josef Stoer and Roland Bulirsch. *Introduction to Numerical Analysis*. Springer, 1980.

[98] Arthur H. Stroud and Don Secrest. *Gaussian Quadrature Formulas*. Prentice Hall, 1966.

[99] Witold G. Strupczewski, Vijay P. Singh, and Stanislaw Weglarczyk. Asymptotic bias of estimation methods caused by the assumption of false probability distribution. *J. Hydrol.*, 258:122–148, 2002.

[100] Edward A. Sudicky. A natural gradient experiment on solute transport in a sand aquifer: Spatial variability of hydraulic conductivity and its role in the dispersion process. *Water Resour. Res.*, 22(13):2069–2082, 1986.

[101] Bruno Sudret and Armen Der Kiureghian. Stochastic finite elements and reliability: A state-of-the-art report. Technical Report UCB/SEMM–2000/08, University of California, Berkeley, 2000.

[102] Bruno Sudret, Marc Berveiller, and Maurice Lemaire. A stochastic finite element procedure for moment and reliability analysis. *Rev. Eur. Méca. Num.*, 15(7,8):825–866, 2006.

[103] Radu A. Todor and Christoph Schwab. Convergence rates for sparse chaos approximations of elliptic problems with stochastic coefficients. *IMA J. Numer. Anal.*, 27(2):232–261, 2007.

[104] Kahraman Ünlü, M. Levent Kavvas, and Donald R. Nielsen. Stochastic analysis of field measured unsaturated hydraulic conductivity. *Water Resour. Res.*, 25(12):2511–2519, 1989.

[105] James V. Uspensky. On the convergence of quadrature formulas related to an infinite interval. *Trans. Amer. Math. Soc.*, 30(3):542–559, 1928.

[106] Brian J. Wagner and Steven M. Gorelick. Reliable aquifer remediation in the presence of spatially variable hydraulic conductivity: From data to design. *Water Resour. Res.*, 25(10):2211–2225, 1989.

[107] Xiaoliang Wan and George Em Karniadakis. An adaptive multi-element generalized polynomial chaos method for stochastic differential equations. *J. Comp. Phys.*, 209(2):617–642, 2005.

[108] Xiaoliang Wan and George Em Karniadakis. Beyond Wiener–Askey expansions: handling arbitrary PDFs. *J. Sci. Comput.*, 27(1–3):455–464, 2006.

[109] Dirk Werner. *Funktionalanalysis*. Springer, 3rd edition, 2000.

[110] Norbert Wiener. The homogeneous chaos. *Am. J. Math.*, 60:897–936, 1938.

[111] Margaret H. Wright. Interior methods for constrained optimization. *Acta Numerica*, 1:341–407, 1992.

[112] Dongbin Xiu and Jan S. Hesthaven. High-order collocation methods for differential equations with random inputs. *SIAM J. Sci. Comput.*, 27(3):1118–1139, 2005.

[113] Dongbin Xiu and George Em Karniadakis. Modeling uncertainty in steady state diffusion problems via generalized polynomial chaos. *Comput. Methods Appl. Mech. Engrg.*, 191:4927–4948, 2002.

[114] Dongbin Xiu and George Em Karniadakis. The Wiener–Askey polynomial chaos for stochastic differential equations. *SIAM J. Sci. Comput.*, 24(2):619–644, 2002.

[115] Akiva M. Yaglom. *An Introduction to the Theory of Stationary Random Functions*. Prentice Hall, 1962.

Zusammenfassung

Bei numerischen Berechnungen stößt man immer wieder auf die Schwierigkeit, daß gewisse Parameter in den beschreibenden Modellen aufgrund von Meßungenauigkeiten oder ihrer starken Variabilität nur mit einer gewissen Unsicherheit bestimmt werden können. In den letzten Jahren hat sich das Interesse an der Quantifizierung dieser Unsicherheiten und deren Auswirkungen auf die Lösung der numerischen Simulationen erhöht, wobei sich die sogenannte Polynomial-Chaos-Methode in einer Vielzahl von Anwendungen als effizientes Verfahren zur Beantwortung dieser Fragestellung erwiesen hat.

Das Ziel der vorliegenden Dissertation besteht in der Anwendung dieser Methode auf die Richards-Gleichung zur Modellierung von Grundwasserströmungen in gesättigten und ungesättigten Böden. Die Schwierigkeiten bei der numerischen Behandlung dieser Gleichung liegen darin begründet, daß die Sättigung und die hydraulische Leitfähigkeit, die in den Orts- und Zeitableitungen auftauchen, nichtlinear von der Lösung abhängen. Die Berücksichtigung unsicherer Parameter, worunter stochastische Anfangs- und Randbedingungen, vor allem aber eine stochastische Permeabilität fallen können, führt letztendlich auf die Untersuchung einer stochastischen Variationsungleichung zweiter Art mit Hindernisbedingungen und einem nichtlinearen konvexen Funktional in Form eines Superpositionsoperators.

Die Betrachtung von Variationsungleichungen im Zusammenhang mit unsicheren Parametern und der Polynomial-Chaos-Methode ist neu, so daß zunächst eine schwache Formulierung des Problems hergeleitet wird, bevor die Approximation der Parameter durch eine Karhunen-Loève-Entwicklung erfolgt. Für das zeitdiskrete Problem läßt sich nun durch Umformulierung in ein konvexes Minimierungsproblem die Existenz einer eindeutigen Lösung u in einem Tensorraum beweisen. Hiernach erfolgt die Diskretisierung mit finiten Elementen und polynomiellen Ansatzfunktionen, wobei das konvexe Funktional mit geeigneten Gauß-Quadratur-Formeln approximiert wird. Für den Spezialfall eines stochastischen Hindernisproblems wird die Konvergenz der Lösung des diskretisierten Problems gegen die Lösung u bewiesen. Hinzu kommen numerische Untersuchungen zur Abschätzung des Diskretisierungsfehlers, die mit bekannten Resultaten für den linearen Fall verglichen werden.

Im zweiten Teil der Arbeit wird ein effizientes numerisches Verfahren zur Lösung des diskretisierten Minimierungsproblems entwickelt. Als Grundlage dient ein Block-Gauß-Seidel-Verfahren, das global konvergiert und in dem eine Transformation zur Entkopplung der stochastischen Koeffizienten vorgestellt wird, die die Brücke zwischen stochastischen Galerkin- und stochastischen Kollokationsverfahren schlägt. Das ermöglicht letzthin auch die Erweiterung zu Mehrgitterverfahren, um die Konvergenzgeschwindigkeit deutlich zu verbessern.

Zum Abschluß wird die Leistungsfähigkeit des entwickelten Verfahrens an einem realistischen Beispiel mit lognormalverteilter Permeabilität und exponentieller Kovarianz gezeigt.

Lebenslauf

Mein Lebenslauf wird aus Gründen des Datenschutzes in der elektronischen Fassung meiner Arbeit nicht veröffentlicht.