Technische Universität München
Zentrum Mathematik

Generalized Wiener expansions for individual based control: theory and applications

Faidra Stavropoulou

Technische Universität München
Zentrum Mathematik

Generalized Wiener expansions for individual based control: theory and applications

Faidra Stavropoulou

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:		Univ.-Prof. Dr. Donna Pauler Ankerst
Prüfer der Dissertation:	1.	Univ.-Prof. Dr. Johannes Müller
	2.	Hon.-Prof. Dr. Dr. h.c. Albert Gilg
	3.	Prof. Dr. Youssef M. Marzouk Massachusetts Institute of Technology, Cambridge, USA

Die Dissertation wurde am 13.05.2013 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 09.09.2013 angenommen.

Abstract

This thesis deals with two problems arising in the application of polynomial chaos (PC) in dynamical systems with parametric uncertainty. In the first part, an algorithm for the parametrization of a random variable in terms of a polynomial basis from given observations is presented. The proofs for the corresponding convergence results are based on the theory of optimal transportation. PC expansions do not preserve positivity in general, although this property is central for problems in mathematical biology. A solution to this problem is proposed in the second part. It is based on the construction of positive summability kernels. The last chapter of the thesis is concerned with the optimization of the euglycemic clamp experiment. Two methods based on Monte Carlo and PC are analyzed and compared.

Zusammenfassung

Die vorliegende Doktorarbeit befasst sich mit zwei Problemen in der Anwendung von polynomialen Chaos (PC) in dynamischen Systemen mit Parameterunsicherheit. Zunächst wird ein Algorithmus zur Darstellung einer Zufallsvariablen bezüglich einer polynomialen Basis aus gegebener Beobachtungen entwickelt. Die Beweise für entsprechende Konvergenzergebnisse basieren auf der Theorie des optimalen Transportes. PC Entwicklungen erhalten die Positivität generisch nicht, obwohl diese Eigenschaft zentral bei Problemen der mathematischen Biologie ist. Es wird eine Lösung für dieses Problem vorgestellt, die auf der Konstruktion von positiven summierbaren Kernen basiert. Der letzte Teil der Arbeit beschäftigt sich mit der Optimierung des euglykämischen Clamp Experiments. Zwei Methoden basierend auf Monte Carlo und PC werden analysiert und verglichen.

Acknowledgments

First of all, I would like to thank my supervisor Prof. Dr. Johannes Müller for his help and endless support during my studies. For encouraging, motivating and believing in me, for his patience and suggestions, and for the knowledge and values he conveyed to me, I will always be grateful.

Further I would like to thank Dr. Burkhard Hense for establishing the collaboration with the German Mouse Clinic and thus providing me with an exciting research topic. I am also indebted to Dr. Susanne Neschen, Melanie Kahle and Nicole Boche from the German Mouse Clinic for their cooperation and for helping me understand complicated biological concepts by taking the time and effort to explain them in a simple way.

Special thanks and appreciation go to Dr. Josef Obermaier who communicated to me his love for mathematics and welcomed me every time with a smile. Working with him was essential for the development of my thesis.

I am also grateful to Prof. Dr. Youssef M. Marzouk for giving me the opportunity to visit his group, for his guidance and cooperation and to Dr. Tarek El Moselhy for being a great teacher and for his help and patience during my visit.

Moreover, I also would like to thank everyone in the Aerospace Computational Design Laboratory for the warmth with which they welcomed me, for considering me since my very first day as an equal member of the group and for making my stay in Boston an unforgettable experience.

To the members of my examination committee I would like to express my gratitude for reviewing the thesis and providing me with constructive feedback and valuable comments.

Finally, I would like to thank my friends, my sister, and my parents for always being there for me.

Contents

1 Introduction

Pour définir le probable il faut posséder le vrai.
Jean-Paul Sartre

1.1 Motivation and objectives

Mathematical models for biological systems often take the form of dynamical systems that involve unknown parameters. These have to be estimated from data, which are subjected to measurement errors. If individual organisms are considered, due to natural variation or different health conditions, they will have similar but different values for the same parameter. This makes it sometimes impossible to find one parameter value that fits various data sets corresponding to different subjects, and thus stresses the need to introduce stochasticity in models for biological processes.

The estimation of parameters in a model from noisy observations consists an *inverse problem* [75, 77]. Although deterministic methods to solve this inverse problem exist, the focus in this thesis will be stochastic methods and more precisely *Bayesian methods* [16,108]. In this context, unknown parameters and measurements are modeled mathematically as random variables in an abstract probability space $(\Omega, \mathcal{A}, P)$. Any existing information on the model parameters may be incorporated in the inference procedure by the use of a *prior* distribution assigned initially to the parameters. In the presence of data, this prior distribution is updated by Bayes' theorem to the *posterior* distribution. The latter is the estimate of the parameters in the Bayesian framework. This is one of the main differences between Bayesian and deterministic approaches, which result in point estimates.

Markov Chain Monte Carlo (MCMC) methods are a class of techniques introduced to explore the posterior distribution and compute location and dispersion estimates, such as moments and credible intervals, and are an active research area [33, 62]. They are employed to discretize the posterior distribution, which is usually analytically intractable, by drawing a sample of realizations from it. This sample is then used to numerically evaluate the aforementioned estimates, which often take the form of high-dimensional integrals with respect to the posterior distribution. In the case that the data come sequentially and real-time estimation is required, *sequential Monte Carlo (SMC) methods* are employed to solve this problem faster and more efficiently [37, 44, 49].

Bayesian methods naturally require the propagation of uncertainty in the parameters through the model equations. *Monte Carlo (MC) simulation*, a method which dates back to the work of von Neumann and Ulam provides a simple solution to this problem: the model equations are solved for a set of realizations from the distribution of the model parameters and thus produce an ensemble of solution realizations [30]. This ensemble is then used to obtain density estimates and reveal statistical properties of the solution. An alternative to Monte Carlo methods, which can also speed up computations in a Bayesian framework as

was shown by Marzouk, Najm and Rahn [92, 93] are methods based on *polynomial chaos (PC)* expansions.

The term was first used by Wiener [126] and gained much attention by the engineering community due to the work of Ghanem and Spanos [61]. The idea is that given a *basic* random variable $\boldsymbol{\Xi}$ and a corresponding sequence of orthogonal polynomials $\{P_{\boldsymbol{n}} : \boldsymbol{n} \in \mathbb{N}_0^d\}$, every random variable X defined on the probability space $(\Omega, \sigma(\boldsymbol{\Xi}), P)$ with finite variance can be decomposed in a series of polynomials

$$X = g_X(\boldsymbol{\Xi}) = \sum_{\boldsymbol{n} \in \mathbb{N}_0^d} \widehat{g_X}_{\boldsymbol{n}} P_{\boldsymbol{n}}(\boldsymbol{\Xi})$$

provided that the polynomials are dense in the Hilbert space $L^2(\Omega, \sigma(\boldsymbol{\Xi}), P)$ [52, 72, 128]. Polynomial chaos expansions can be found in the literature also under the names *Wiener-Askey* [128] or *Wiener-Haar* expansions [86]. Instead of orthogonal polynomials, other basis functions, as Haar wavelets may be chosen [84]. In this thesis, only global orthogonal polynomials will be considered.

In a dynamical system framework, both the model parameters and initial conditions may be random variables and thus admit such expansions. Then, the solution of the model equations will be a stochastic process and its time and/or space dependent PC coefficients $\{\widehat{g_X}_{\boldsymbol{n}} : \boldsymbol{n} \in \mathbb{N}_0^d\}$ can be computed via *intrusive (Galerkin)* and *non-intrusive* spectral methods [86, 131]. These coefficients summarize all the statistical information about the solution.

In many applications one may not only be interested in the propagation of parametric uncertainty through a given model, but also in the design of an optimal controller. Deterministic solutions for such problems based on robust control theory and worst-case analysis usually result in very conservative controls [50, 132]. The field of *probabilistic robust control* theory seeks for solutions to the optimal control problem in a probabilistic sense: one does not seek for a controller which is optimal for all possible realizations of the parameters, but for a controller that has a given probability of being optimal over the range of parameters. Sampling and MC methods has been used to handle such problems [29, 119]. As an alternative to these methods, PC approximations can be used for the control design of systems with parametric uncertainty. A number of works exist already in this direction. The idea of the application of PC methods in control problems with parametric uncertainty first appeared in the literature in the paper of Monti, Ponci and Lovett [96], where the PC techniques were applied for the control of a power converter. Later, Hover and Triantafyllou [71] stressed out the potential of PC methods for the control of nonlinear systems and as an alternative to costly MC simulations. In the paper of Fisher and Bhattacharya [53] the stochastic stability and the optimality conditions were analyzed for a linear quadratic regulator (LQR) of systems with parametric uncertainty represented by PC expansions. In the PhD thesis by Blanchard [23] a numerical method for the solution of an LQR optimization problem related with the PC framework was proposed and in the one by Templeton [118] a theoretical framework for the extension of H_2 and LQR design to systems with parametric uncertainty approximated by PC expansions was given. More recently, Peng, Ghanem and Li [102] studied the problem of the control of a Duffing oscillator subjected to stochastic excitation and proposed a solution based on PC expansions.

Two problems that may arise on considerations regarding the development of control methods based on PC and may limit their applicability are next stated.

Problem 1: There is no generally accepted method for the parametrization of a random variable X based on a sample from its distribution in terms of a basic random variable Ξ. The problem translates in finding a measure preserving transformation g_X such that $X = g_X(\Xi)$. The situation becomes of course more difficult when X is a random vector with dependent components. Taking into account the dependences is essential in order to have a useful and accurate PC representation. Le Maître and Knio note in [86] that *"the impact that UQ [uncertainty quantification] schemes can bring to such situations [elaborate physical models] is in large part conditioned on a suitable representation of the uncertainty in the model inputs"*.

This problem may arise in a control framework when PC methods are combined with Bayesian parameter estimation. In this case, one needs to find an appropriate transformation g_X of a random variable X, whose distribution is given by a complicated posterior distribution.

Problem 2: The second problem arises when approximating positive (or more generally bounded) random variables by truncated polynomial expansions. In such cases, one cannot guarantee that the finite approximation will stay positive for all realizations of the basis random vector Ξ. The problem was addressed in [41, 106] and it was pointed out how this situation can lead to instabilities when propagating a finite expansion through a dynamical system. In addition, it may result in meaningless control policies when designing a controller based on PC methods for nonlinear models, which may exhibit blow-up behavior in finite time.

Particularly in mathematical biology the preservation of positivity is of central importance: comparing to engineering problems where the law of physics guarantee the validity of the models, positivity is one of the minimal properties to require for mathematical models which represent biological processes in order for them to be meaningful.

1.2 Outline

The thesis is organized as follows.

In chapter 2 basic concepts are introduced and notation is fixed. Sections 2.1 and 2.2 contain results from probability and polynomial chaos theory. The application of PC expansions in ordinary differential equations with parametric uncertainty is discussed in section 2.3, where also spectral Galerkin and non-intrusive methods are briefly reviewed.

Chapter 3 deals with the problem 1 stated above. An algorithm for the parametrization of a random variable X in terms of another random variable Ξ from given observations is presented. The method is based on the construction of a discrete transformation which converges to a continuous transformation g_X as the number of observations grows. This discrete map is used to estimate the PC coefficients via a regression approach. After a review of existing work in section 3.1, basic results from the theory of order statistics and optimal transportation are summarized in sections 3.2 and 3.3. Section 3.4 deals with the problem in the one-dimensional case, where the discrete map is constructed with the help of order statistics and with the general multi-dimensional case, where the discrete map is constructed by solving a discrete optimal transportation problem. The chapter closes with numerical simulations in section 3.5.

The preservation of positivity in finite polynomial approximations is examined in chapter 4. A solution is proposed based on positive summability methods. Known results from approximation theory are generalized to the multi-dimensional case and applied to the stochastic framework. In section 4.1, tools from functional analysis are summarized. The summability methods are analyzed in section 4.2, where convergence results and approximation errors for the proposed method are also given. Examples of positive kernels are stated as well as numerical simulations to validate the theoretical results in sections 4.3 and 4.4 respectively.

The results in this chapter were developed in cooperation with PD Dr. Josef Obermaier from Helmholtz Zentrum München, Institute of Computational Biology. Original papers on this topic include [10, 11, 81–83, 100].

Chapter 5 deals with the optimization of the euglycemic hyperinsulinemic clamp (EHC) test, a widely accepted experimental test used in diabetes research. The optimization is mathematically a problem of individual based real-time optimal control. Section 5.1 provides with the essential biological background on the glucose-insulin system and details on the conduction of the EHC test. A mathematical model for the glucose dynamics during the test is presented in section 5.2. The model is based on data from clamps run on mice provided by Dr. Susanne Neschen from Helmholtz Zentrum München, Institute of Experimental Genetics. Bayesian methods for parameter inference are reviewed in section 5.3 and in section 5.4 a first optimization algorithm is given based on the exact solution of the quadratic optimization control problem and SMC methods. In the last section 5.5, a second algorithm based on PC expansions, SMC methods and the results from chapter 3 is presented and analyzed.

The results in this chapter were partly obtained in cooperation with Prof. Dr. Youssef M. Marzouk and Dr. Tarek El Moselhy from Massachusetts Institute of Technology, Department of Aeronautics and Astronautics.

Finally, conclusions are drawn in chapter 6.

2 Preliminaries

2.1 Basics from probability theory

In this section, some basic concepts and results from probability theory are recalled. For a rigorous introduction to the topic, the reader is referred to [78, 112].

Let Ω be a set containing all the possible outcomes of a random experiment.

Definition 2.1. *Let $\Omega \neq \emptyset$. A collection $\mathcal{A}$ of subsets of Ω with the properties*

(i) $\Omega \in \mathcal{A}$,

(ii) *if* $A \in \mathcal{A}$, *then* $\Omega \setminus A \in \mathcal{A}$,

(iii) *if* $\{A_n\}_{n\in\mathbb{N}} \subseteq \mathcal{A}$, *then* $\cup_{n\in\mathbb{N}} A_n \in \mathcal{A}$,

is called a σ-algebra on Ω. An element $A \in \mathcal{A}$ of a σ-algebra $\mathcal{A}$ is called an event.

The tuple $(\Omega, \mathcal{A})$ is called a *measurable space*. Given a non-empty collection of subsets $\mathcal{C}$ of Ω, there exists the smallest σ-algebra that contains $\mathcal{C}$. This will be denoted by $\sigma(\mathcal{C})$ and is called the σ-algebra *generated* by $\mathcal{C}$. The smallest σ-algebra generated by the open sets in the Euclidean space $\mathbb{R}^d$ is called the *Borel* σ-algebra and is denoted by $\mathcal{B}(\mathbb{R}^d)$.

Definition 2.2. *A function $\mu\colon \mathcal{A} \to \mathbb{R}$ is called a measure if it satisfies the following*

(i) $\mu(\emptyset) = 0$,

(ii) $\mu(A) \geq 0, \forall\ A \in \mathcal{A}$,

(iii) *if* $\{A_n\}_{n\in\mathbb{N}} \subseteq \mathcal{A}$ *are disjoint sets, then* $\mu(\cup_{n\in\mathbb{N}} A_n) = \sum_{n\in\mathbb{N}} \mu(A_n)$.

The triple $(\Omega, \mathcal{A}, \mu)$ is called a *measure space*. If in addition $\mu(\Omega) = 1$, then μ is called a *probability measure*, it is usually denoted by P and the triple $(\Omega, \mathcal{A}, P)$ is called a *probability space*. The *support* of a measure μ on a measurable space $(\Omega, \mathcal{A})$ is defined as

$$\text{supp}\mu = \{\omega \in \Omega\colon \mu(U) > 0 \text{ for every neighborhood } U \text{ of } \omega\}. \tag{2.1}$$

Next, functions on Ω are defined.

Definition 2.3. *Let $(\Omega, \mathcal{A})$ be a measurable space. A function $X\colon \Omega \to \mathbb{R}^d$ that satisfies*

$$\forall\ B \in \mathcal{B}(\mathbb{R}^d),\ X^{-1}(B) = \{\omega \in \Omega\colon X(\omega) \in B\} \in \mathcal{A} \tag{2.2}$$

is called a $\mathcal{A} - \mathcal{B}(\mathbb{R}^d)$ measurable function. In the context of probability spaces, measurable functions are called random variables. The corresponding σ-algebras when referring to measurable functions will be omitted, if these are clear from the context.

The *values or realizations* $X(\omega) = x \in \mathbb{R}^d$ of a random variable X will be denoted with small letters. Random variables mapping on $\mathbb{R}^d$ for $d \geq 2$ are called *random vectors* and will be also denoted with bold capital letters. Each component $X_i \colon \Omega \to \mathbb{R}, i = 1, \ldots, d$ of a random vector $\boldsymbol{X} = (X_1, \ldots, X_d) \colon \Omega \to \mathbb{R}^d$ is a random variable.

The notion of a stochastic process is now defined. These are random quantities that also depend on time or space. For more details see in [48].

Definition 2.4. *A stochastic process is a collection $\{X_t\}_{t \in T}$ of random variables, indexed by a set T and defined on a common probability space $(\Omega, \mathcal{A}, P)$. The set T can be countable or not. For fixed t, X_t is a random variable and for fixed $\omega \in \Omega$, $\{X_t(\omega)\}_{t \in T}$ is a function on T and is called a sample path or realization of the process.*

The *smallest σ-algebra generated by a random variable X*, denoted $\sigma(X)$, is the σ-algebra $\sigma(\mathcal{C})$ generated by the collection of sets $\mathcal{C} = \{X^{-1}(B), B \in \mathcal{B}(\mathbb{R}^d)\}$. An important concept in probability theory is the independence of random variables. This is based on the independence of σ-algebras.

Definition 2.5. *Let $(\Omega, \mathcal{A}, P)$ be a probability space and $\{\sigma_i\}_{i=1,\ldots,I} \subseteq \mathcal{A}$ a collection of σ-algebras on Ω. One says that the σ_i are independent, if for any $n \in \mathbb{N}$ and any sets $A_{i_1} \in \sigma_{i_1}, \ldots, A_{i_n} \in \sigma_{i_n}$, the following holds:*

$$P(A_{i_1} \cap \ldots \cap A_{i_n}) = P(A_{i_1}) \cdots P(A_{i_n}). \tag{2.3}$$

A collection of random variables $\{X_i\}_{i=1,\ldots,I}$ are called independent if their generated σ-algebras $\{\sigma(X_i)\}_{i=1,\ldots,I}$ are independent.

Each random variable X defines a probability measure P_X on the measurable space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ by the equation

$$P_X(B) = P(X^{-1}(B)), \ \forall \ B \in \mathcal{B}(\mathbb{R}^d). \tag{2.4}$$

This measure is called the *image measure* or *distribution* of the random variable X and is also denoted by $(X)_{\#}P$. A random variable Y which has the same distribution as X will be called a *copy* of X.

One can easily show that if $X \colon \Omega \to \mathbb{R}^d$ is a random variable and $g \colon (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) \to (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ a measurable function, then the function $Y \colon \Omega \to \mathbb{R}^d, \ Y = g(X)$ is again a random variable. The following lemmata play a central role in what follows.

Lemma 2.6. *(Doob-Dynkin) Let Ξ, X be two random variables defined on a common measurable space $(\Omega, \mathcal{A})$ and denote by $\sigma(\Xi)$ the σ-algebra generated by Ξ. Then, X is $\sigma(\Xi) - \mathcal{B}(\mathbb{R}^d)$ measurable if and only if there exists a measurable function*

$$g_X \colon (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) \to (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$$

such that

$$X = g_X(\Xi). \tag{2.5}$$

Lemma 2.7. *(Change of variables) Let $X \colon (\Omega, \mathcal{A}, P) \to (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ be a random variable and $g \colon (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) \to (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ a measurable function. The function g is integrable with respect to P_X if and only if the random variable $g(X)$ is integrable with respect to P and in such cases*

$$\int_{\Omega} g(X(\omega)) dP(\omega) = \int_{\mathbb{R}^d} g(x) dP_X(x). \tag{2.6}$$

The *distribution function* of a random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ is the function $F\colon \mathbb{R}^d \to \mathbb{R}$ defined by

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = P(X_1 \leq x_1, \ldots, X_d \leq x_d),\ \boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d. \tag{2.7}$$

If there exists a function $f_X\colon \mathbb{R}^d \to \mathbb{R}$ such that $f_X \geq 0$, $\int_{\mathbb{R}^d} f_X(\boldsymbol{x})d\boldsymbol{x} = 1$ and

$$F_X(\boldsymbol{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} f_X(\boldsymbol{y})d\boldsymbol{y},\ \boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d, \tag{2.8}$$

then f_X is called the *density* of F_X. The random vector $\boldsymbol{X}$ is called in this case *continuous* with respect to the Lebesgue measure. Related to distribution functions are marginal and conditional distributions. The *marginal distribution function* of the random variable X_i, $i = 1, \ldots, d$ is defined as

$$F_{X_i}(x_i) = \lim_{x_1 \to \infty} \cdots \lim_{x_{i-1} \to \infty} \lim_{x_{i+1} \to \infty} \cdots \lim_{x_d \to \infty} F_{\boldsymbol{X}}(\boldsymbol{x}), \tag{2.9}$$

and its *conditional distribution function* given the random variables $\{X_j\}_{j \in J}$, where $J \subseteq \{1, \ldots, d\} \setminus \{i\}$, is defined as

$$F_{X_i|X_{j,j\in J}}(x_i \mid x_{j,j\in J}) \equiv F_{i|j\in J}(x_i \mid x_{j,j\in J}) = P(X_i \leq x_i \mid X_j = x_j, j \in J). \tag{2.10}$$

For the independence of random variables the following characterization based on their distribution functions holds.

Proposition 2.8. *Let $(\Omega, \mathcal{A}, P)$ be a probability space and let $X_1, \ldots, X_d$ be random variables defined on it with distribution functions $F_1, \ldots, F_d$ respectively. Then, the random variables are independent if and only if the distribution function of the random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ takes the product form, i.e.*

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = F_1(x_1) \cdots F_d(x_d),\ \boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d. \tag{2.11}$$

One is usually interested in statistical quantities of a random vector $\boldsymbol{X}$ such as its moments and covariance. The 1-st moment of a random vector, also called the *expectation*, is defined as the vector of expectations of the random variables $X_i, i = 1, \ldots, d$, which are given by

$$E[X_i] = \int_\Omega X_i(\omega)dP(\omega) = \int_{\mathbb{R}} x\, dP_{X_i}(x). \tag{2.12}$$

The 2-nd order central moment or *variance* is defined as the vector of variances of the components defined by

$$\mathrm{Var}[X_i] = \int_\Omega (X_i(\omega) - E[X_i])^2 dP(\omega) = \int_{\mathbb{R}} (x - E[X_i])^2\, dP_{X_i}(x). \tag{2.13}$$

Similarly, the m-th order *moments* are defined by the equations

$$E[X_i^m] = \int_\Omega X_i^m(\omega)dP(\omega) = \int_{\mathbb{R}} x^m\, dP_{X_i}(x). \tag{2.14}$$

Finally, the *covariance* of two random variables X_i, X_j is defined as

$$\mathrm{Cov}[X_i, X_j] = E[(X_i - E[X_i])(X_j - E[X_j])],\ i, j = 1, \ldots, d. \tag{2.15}$$

The L^p spaces of random variables are now introduced.

Definition 2.9. *A random variable X on a probability space $(\Omega, \mathcal{A}, P)$ is called p-integrable if*

$$\int_\Omega |X(\omega)|^p \, dP(\omega) < \infty, \text{ for } 1 \leq p < \infty. \tag{2.16}$$

The space of all p-integrable random variables on $(\Omega, \mathcal{A}, P)$ is denoted by $L^p(\Omega, \mathcal{A}, P)$. For $p = \infty$, one defines that a random variable X belongs to the space $L^\infty(\Omega, \mathcal{A}, P)$, if

$$\operatorname{ess\,sup}_{\omega \in \Omega} |X(\omega)| < \infty. \tag{2.17}$$

If one equips these spaces with the norms

$$\|X\|_p = \left(\int_\Omega |X(\omega)|^p dP(\omega) \right)^{\frac{1}{p}}, \ 1 \leq p < \infty \tag{2.18}$$

and

$$\|X\|_\infty = \operatorname{ess\,sup}_{\omega \in \Omega} |X(\omega)| \tag{2.19}$$

respectively, then they become Banach spaces. For the special case $p = 2$ the norm $\| \cdot \|_2$ is induced by the inner product

$$\langle X_1, X_2 \rangle = \int_\Omega X_1(\omega) X_2(\omega) dP(\omega), \tag{2.20}$$

thus making the space $L^2(\Omega, \mathcal{A}, P)$ a Hilbert space. In the sequel, the explicit reference to the underlying probability space will be omitted if this is clear from the context.

There exists several notions of convergence of sequences of random variables. These and some related results to be used in the sequence are summarized here.

Definition 2.10. *Let $\{X_n\}_{n \in \mathbb{N}}$ be a sequence of random variables defined on a common probability space $(\Omega, \mathcal{A}, P)$. One says that the sequence $\{X_n\}$*

(i) converges weakly to a random variable X, and write $X_n \rightharpoonup X$, if

$$E[f(X_n)] \to E[f(X)], \ n \to \infty \tag{2.21}$$

for all bounded continuous functions $f : \mathbb{R} \to \mathbb{R}$.

(ii) converges in probability to X, and write $X_n \xrightarrow{P} X$, if

$$\forall \, \varepsilon > 0, P(\omega \in \Omega : |X_n(\omega) - X(\omega)| > \varepsilon) \to 0, \ n \to \infty. \tag{2.22}$$

(iii) converges with probability 1 or almost surely (a.s.), and write $X_n \xrightarrow{wp1} X$ or equivalently $X_n \xrightarrow{a.s.} X$, if

$$P(\omega \in \Omega : X_n(\omega) \to X(\omega)) = 1, \ n \to \infty. \tag{2.23}$$

(iv) converges in L^p, and write $X_n \xrightarrow{L^p} X$ if

$$E[|X_n|^p + |X|] < \infty \text{ and } E[|X_n - X|^p] \to 0, \ n \to \infty. \tag{2.24}$$

The L^2-convergence is also referred as *mean square convergence* in the literature. Convergence with probability 1 implies convergence in probability which implies weak convergence. Furthermore, convergence in L^p implies convergence in probability. The converse implications are not in general true.

Two results on the convergence of random variables are next stated.

Theorem 2.11. *(Law of large numbers) Let $\{X_n\}_{n\in\mathbb{N}}$ be an independent, identical distributed (iid) sequence of random variables on a common probability space $(\Omega, \mathcal{A}, P)$. Then, $E[X_1] < \infty$ if and only if*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \overset{wp1}{\to} E[X_1],\ n \to \infty. \tag{2.25}$$

Theorem 2.12. *(Continuous mapping theorem) Let $\{X_n\}_{n\in\mathbb{N}}, X$ be random variables on a probability space $(\Omega, \mathcal{A}, P)$ with values in $\mathbb{R}^d$ such that $X_n \overset{wp1}{\to} X$ and let $g\colon \mathbb{R}^d \to \mathbb{R}^d$ be a continuous function. Then $g(X_n) \overset{wp1}{\to} g(X)$.*

Note that Theorem 2.12 holds also in the case of convergence in probability and in distribution, since these are implied from the the almost sure convergence.

Two types of convergence of measures are also here recalled.

Definition 2.13. *A sequence of measures $\{\rho_n\}_{n\in\mathbb{N}}$ defined on a measurable space $(\Omega, \mathcal{A})$*

(i) converges strongly to a measure ρ, denoted as $\rho_n \to \rho$, if

$$\rho_n(A) \to \rho(A),\ n \to \infty \tag{2.26}$$

for all events $A \in \mathcal{A}$ such that $\rho(\partial A) = 0$. Here, ∂A stands for the boundary of the set A.

(ii) converges weakly to a measure ρ, denoted as $\rho_n \rightharpoonup \rho$, if

$$\int_\Omega f(\omega) d\rho_n(\omega) \to \int_\Omega f(\omega) d\rho(\omega),\ n \to \infty \tag{2.27}$$

for all bounded continuous functions $f\colon \Omega \to \mathbb{R}$.

Finally, note that a property $\mathcal{E}$ will be said to hold μ-a.e. for a measure μ, if the property $\mathcal{E}$ is true for all $\omega \in \Omega \setminus A$ and $\mu(A) = 0$.

2.2 Generalized polynomial chaos (PC) expansions

The material in this section is based on the books [86,131]. For more details and theoretical results on orthogonal polynomials see in [35,117].

Polynomial chaos (PC) expansions are spectral expansions of random variables with finite second moments. The term dates back to the work of Wiener [126] who used the term *polynomial chaos* to define polynomial spaces over Gaussian random variables. Cameron and Martin [27] proved later that there exists an expansion of Hermite polynomials in Gaussian random variables for a certain class of functionals. The theory gained much attention in the engineering community with the work of Ghanem and Spanos [61] who

used expansions of Hermite polynomials in Gaussian random variables in a stochastic finite element framework. Later, Xiu and Karniadakis [128] proposed the use of polynomial expansions in non-Gaussian random variables to speed up the convergence rate of the approximation. These expansions are known also as *generalized* polynomial chaos expansions. Their idea is based on the fact that the density functions of the most common probability distributions are the same as the weighting functions used to define the most common sequences of orthogonal polynomials. In Table 2.1 (as in [128]) the most important distributions and the corresponding polynomial systems are summarized. The conditions under which the generalized polynomial chaos expansions actually converge were examined in [52].

Table 2.1: The Askey scheme of orthogonal polynomials

	probability distribution	orthogonal polynomials
continuous	Gaussian	Hermite
	Beta	Jacobi
	Gamma	Laguerre
	Uniform	Legendre
discrete	Poisson	Charlier
	Binomial	Krawtchouk
	Negative Binomial	Meixner
	Hypergeometric	Hahn

Here the theory is presented in the general case. For more details on the Gaussian case see in [72]. Moreover, only the case of expansions in global polynomials is considered. Other expansions, for example in wavelet or local bases in the stochastic space, have been considered and are an active research area [84, 125].

Let $(\Omega, \mathcal{A}, P)$ be a probability space and assume that there exist independent random variables

$$\Xi_i \colon (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R})),\ i = 1, \dots, d \tag{2.28}$$

such that, first of all,

$$\Xi_i \in L^p(\Omega, \mathcal{A}, P)\ \forall\, i = 1, \dots, d \text{ and } \forall\, 1 \leq p < \infty, \tag{2.29}$$

and secondly, the support $\mathrm{supp}\mu = S_i$ of the image measure $\mu_i = (\Xi_i)_{\#}P$ of Ξ_i is of infinite cardinality for any i. Note that the random variables need not be identically distributed. Then for any i there exists a sequence of orthogonal polynomials $\{P_{i;n}\}_{n=0}^{\infty}$ such that $P_{i;n}$ is a polynomial of degree n and

$$\int_{\mathbb{R}} P_{i;n}(x) P_{i;m}(x) d\mu_i(x) = \frac{1}{h_{i;n}} \delta_{n,m},\ \forall\, n, m \in \mathbb{N}_0 \tag{2.30}$$

with $h_{i;n} > 0$. The quantities $h_{i;n}$ are called *Haar weights*. Different normalizations exist for each sequence of orthogonal polynomials. For simplicity, it is assumed in this thesis that $P_{i;0}(x) = 1$ and thus $h_{i;0} = 1$ for any $i = 1, \dots, d$. Every sequence of orthogonal polynomials satisfies a 3-term recurrence relation as follows

$$x P_{i;n}(x) = \gamma_{i;n} P_{i;n+1}(x) + \beta_{i;n} P_{i;n}(x) + \alpha_{i;n} P_{i;n-1}(x),\ \forall\, i = 1, \dots, d,\ \forall\, n \in \mathbb{N}_0, \tag{2.31}$$

where $P_{i;-1}(x) = 0$, for all $x \in \mathbb{R}$ and for all $i = 1, \ldots, d$ and $\{\alpha_{i;n}\}_{n=0}^{\infty}$, $\{\beta_{i;n}\}_{n=0}^{\infty}$ and $\{\gamma_{i;n}\}_{n=0}^{\infty}$ are real sequences. One can further show that

$$a_{i;n+1}h_{i;n+1} = \gamma_{i;n}h_{i;n},\ \forall\, i = 1, \ldots, d,\ \forall\, n \in \mathbb{N}_0. \tag{2.32}$$

To each sequence of orthogonal polynomials one can associate its *linearization coefficients* $c_{i;m,n,k}$ which are determined by

$$P_{i;m}(x)P_{i;n}(x) = \sum_{k=0}^{m+n} c_{i;m,n,k}P_{i;k}(x) = \sum_{k=|m-n|}^{m+n} c_{i;m,n,k}P_{i;k}(x). \tag{2.33}$$

Consider now the random vector

$$\boldsymbol{\Xi} = (\Xi_1, \ldots, \Xi_d)\colon (\Omega, \mathcal{A}) \to (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)). \tag{2.34}$$

Due to independence, the image measure of the random vector $\boldsymbol{\Xi}$ is determined by $\boldsymbol{\mu} = \mu_1 \times \cdots \times \mu_d$ with support $\text{supp}\boldsymbol{\mu} = \boldsymbol{S} = S_1 \times \cdots \times S_d$. The set of multivariate polynomials $\{P_{\boldsymbol{n}} : \boldsymbol{n} \in \mathbb{N}_0^d\}$ with

$$P_{\boldsymbol{n}}(\boldsymbol{x}) = \prod_{i=1}^{d} P_{i;n_i}(x_i),\ \forall\, \boldsymbol{n} = (n_1, \ldots, n_d) \in \mathbb{N}_0^d,\ \boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d \tag{2.35}$$

is an orthogonal set with respect to $\boldsymbol{\mu}$, that is

$$\int_{\mathbb{R}^d} P_{\boldsymbol{n}}(\boldsymbol{x})P_{\boldsymbol{m}}(\boldsymbol{x})d\boldsymbol{\mu}(\boldsymbol{x}) = \frac{1}{h_{\boldsymbol{n}}}\delta_{\boldsymbol{n},\boldsymbol{m}},\ \forall\, \boldsymbol{n}, \boldsymbol{m} \in \mathbb{N}_0^d, \tag{2.36}$$

where $h_{\boldsymbol{n}} = h_{1,n_1} \cdots h_{d,n_d}$. In what follows $\mathcal{P} = \text{lin}\{P_{\boldsymbol{n}}\colon \boldsymbol{n} \in \mathbb{N}_0^d\}$ will denote the linear space spanned by the sequence of orthogonal polynomials.

The following theorem [52] states the conditions under which there exists a polynomial representation for square integrable random variables.

Theorem 2.14. *Let*

$$X\colon (\Omega, \sigma(\boldsymbol{\Xi})) \to (\mathbb{R}, \mathcal{B}(\mathbb{R})) \tag{2.37}$$

be a $\sigma(\boldsymbol{\Xi}) - \mathcal{B}(\mathbb{R})$ random variable such that $X \in L^2(\Omega, \sigma(\boldsymbol{\Xi}), P)$. Then, there is an expansion

$$X = g_X(\boldsymbol{\Xi}) = \sum_{\boldsymbol{n} \in \mathbb{N}_0^d} \widehat{g_X}_{\boldsymbol{n}} P_{\boldsymbol{n}}(\boldsymbol{\Xi})h_{\boldsymbol{n}}, \tag{2.38}$$

where

$$\widehat{g_X}_{\boldsymbol{n}} = \int_{\Omega} g_X(\boldsymbol{\Xi}(\omega))P_{\boldsymbol{n}}(\boldsymbol{\Xi}(\omega))dP(\omega) = \int_{\mathbb{R}^d} g_X(\boldsymbol{\xi})P_{\boldsymbol{n}}(\boldsymbol{\xi})d\boldsymbol{\mu}(\boldsymbol{\xi}), \tag{2.39}$$

if and only if $\{P_{\boldsymbol{n}} : \boldsymbol{n} \in \mathbb{N}_0^d\}$ is a dense (complete) orthogonal set in $L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \boldsymbol{\mu})$.

As mentioned in [52] a case when the polynomial system $\{P_{\boldsymbol{n}}\colon \boldsymbol{n} \in \mathbb{N}_0^d\}$ is complete is when the support $\boldsymbol{S}$ of $\boldsymbol{\mu}$ is compact. This result will be useful in the sequence.

The existence of the function g_X is ensured by Lemma 2.6. The convergence is to be interpreted in the mean square sense, i.e. if

$$X_N = \sum_{|\boldsymbol{n}|=0}^{N} \widehat{g_X}_{\boldsymbol{n}} P_{\boldsymbol{n}}(\boldsymbol{\Xi}) h_{\boldsymbol{n}},\ |\boldsymbol{n}| = n_1 + \ldots + n_d,\ N \in \mathbb{N}_0 \tag{2.40}$$

denotes the N-th order approximation of X, then one has

$$\|X - X_N\|_2 \ \to 0,\ N \to \infty. \tag{2.41}$$

In the case one deals with a random vector $\boldsymbol{X}$, then one has to consider a polynomial expansion for each component $X_i, i = 1, \ldots, d$. In practice, one works with the N-th order truncated PC expansion defined in (2.40). This is the best approximation of X in the subspace

$$\mathcal{P}^N = \text{lin}\{P_{\boldsymbol{n}} \colon \boldsymbol{n} \in \mathbb{N}_0^d, |\boldsymbol{n}| \leq N\} \tag{2.42}$$

of $\mathcal{P}$ in the sense that

$$\|X - X_N\|_2 = \inf_{Q \in \mathcal{P}^N} \|X - Q\|_2. \tag{2.43}$$

For a given order N and dimension d of the *basis* random vector $\boldsymbol{\Xi}$, the dimension of the space $\mathcal{P}^N$ is

$$\dim \mathcal{P}^N + 1 = \binom{N+d}{N}. \tag{2.44}$$

Usually, the *lexicographical order* is used in practice to order the terms in (2.40), i.e. the multi-indices $\boldsymbol{n} \in \mathbb{N}_0^d$ are ordered such that $\boldsymbol{n} > \boldsymbol{m}$ if and only if $|\boldsymbol{n}| > |\boldsymbol{m}|$ and the first non-zero coordinate of the multi-index $\boldsymbol{n} - \boldsymbol{m}$ is positive. A single index notation is then also used after the multi-indices are ordered in ascending order according to the above rule. An example of the lexicographical order and the single index notation for the case $d = 2$ and $N = 2$ is given in Table 2.2.

Table 2.2: Lexicographical order and single index notation

$\|\boldsymbol{n}\|$	multi-index $\boldsymbol{n}$	single index n	polynomials
0	(0,0)	0	$P_0(\boldsymbol{\xi}) = P_0(\xi_1)P_0(\xi_2)$
1	(1,0)	1	$P_1(\boldsymbol{\xi}) = P_1(\xi_1)P_0(\xi_2)$
1	(0,1)	2	$P_2(\boldsymbol{\xi}) = P_0(\xi_1)P_1(\xi_2)$
2	(2,0)	3	$P_3(\boldsymbol{\xi}) = P_2(\xi_1)P_0(\xi_2)$
2	(1,1)	4	$P_4(\boldsymbol{\xi}) = P_1(\xi_1)P_1(\xi_2)$
2	(0,2)	5	$P_5(\boldsymbol{\xi}) = P_0(\xi_1)P_2(\xi_2)$

Theorem 2.14 is stated here in the case of finitely many basis random variables Ξ_i. In the literature, this is called the *finite-dimensional noise assumption*. Convergence results in the case of countably many random variables are given in [52]. The infinite-dimensional situation arises for example in the approximation of stochastic processes by their *Karhunen-Loève (KL) expansion*. In this thesis only the finite dimensional case is considered. For more information on KL expansions see for example in [88].

When a polynomial expansion for a random variable X is available in the form (2.38), then its statistical quantities can be easily computed from this expansion by using the orthogonality of the polynomials. The expectation can be expressed by the 0-th order coefficient

$$E[X] = \int_\Omega X(\omega) dP(\omega) = \int_\Omega \left(\sum_{\boldsymbol{n} \in \mathbb{N}_0^d} \widehat{g_X}_{\boldsymbol{n}} P_{\boldsymbol{n}}(\boldsymbol{\Xi}(\omega)) h_{\boldsymbol{n}} \right) dP(\omega) = \widehat{g_X}_{\boldsymbol{0}}, \tag{2.45}$$

and the variance by the sum of the squared coefficients

$$\begin{aligned} \mathrm{Var}[X] &= \int_\Omega (X(\omega) - E[X])^2 dP(\omega) \\ &= \int_\Omega \left(\sum_{\boldsymbol{n} \in \mathbb{N}_0^d} \widehat{g_X}_{\boldsymbol{n}} P_{\boldsymbol{n}}(\boldsymbol{\Xi}(\omega)) h_{\boldsymbol{n}} - \widehat{g_X}_{\boldsymbol{0}} \right)^2 dP(\omega) \\ &= \sum_{|\boldsymbol{n}| \geq 1} (\widehat{g_X}_{\boldsymbol{n}})^2 h_{\boldsymbol{n}}. \end{aligned} \tag{2.46}$$

In the case of two random variables X_1, X_2 their covariance is given by

$$\begin{aligned} \mathrm{Cov}[X_1, X_2] &= \int_\Omega (X_1(\omega) - E[X_1])(X_2(\omega) - E[X_2]) dP(\omega) \\ &= \int_\Omega \left(\sum_{\boldsymbol{n} \in \mathbb{N}_0^d} \widehat{g_{X_1}}_{\boldsymbol{n}} P_{\boldsymbol{n}}(\boldsymbol{\Xi}(\omega)) h_{\boldsymbol{n}} - \widehat{g_{X_1}}_{\boldsymbol{0}} \right) \left(\sum_{\boldsymbol{n} \in \mathbb{N}_0^d} \widehat{g_{X_2}}_{\boldsymbol{n}} P_{\boldsymbol{n}}(\boldsymbol{\Xi}(\omega)) h_{\boldsymbol{n}} - \widehat{g_{X_2}}_{\boldsymbol{0}} \right) \\ &= \sum_{|\boldsymbol{n}| \geq 1} \widehat{g_{X_1}}_{\boldsymbol{n}} \widehat{g_{X_2}}_{\boldsymbol{n}} h_{\boldsymbol{n}}. \end{aligned} \tag{2.47}$$

2.3 PC expansions and differential equations with parametric uncertainty

In this section, it is described how polynomial chaos expansions are used for the propagation of parametric uncertainty through dynamical systems. The theory is presented on the example of ordinary differential equations (ODE) although it can be in the same way applied for example to partial differential and differential algebraic equations.

Let $\boldsymbol{x}$ denote a quantity of interest which evolves in time according to the following dynamics

$$\dot{\boldsymbol{x}}(t, \boldsymbol{\Theta}) = \boldsymbol{f}(t, \boldsymbol{x}, \boldsymbol{\Theta}), \ \boldsymbol{x}(t_0, \boldsymbol{\Theta}) = \boldsymbol{x}^0(\boldsymbol{\Theta}), \tag{2.48}$$

where $\boldsymbol{\Theta} = (\Theta_1, \ldots, \Theta_d)$ denotes a d-dimensional vector of unknown parameters, $\boldsymbol{x}(t, \boldsymbol{\Theta}) = (x_1(t, \boldsymbol{\Theta}), \ldots, x_r(t, \boldsymbol{\Theta}))$ and $\boldsymbol{f}(t, \boldsymbol{x}, \boldsymbol{\Theta}) = (f_1(t, \boldsymbol{x}, \boldsymbol{\Theta}), \ldots, f_r(t, \boldsymbol{x}, \boldsymbol{\Theta}))$.

The parameter d is called the *stochastic dimension* of the problem and r its *deterministic dimension*. Let $J \subseteq \mathbb{R}$, $U \subseteq \mathbb{R}^r$ and $D \subseteq \mathbb{R}^d$ be open sets with $(t_0, \boldsymbol{x}^0) \in J \times U$ such that the function $\boldsymbol{f}$ is smooth on $J \times U \times D$. Then, there is a unique smooth solution of (2.48), possibly defined on subsets of J, U and D [34].

Let $\boldsymbol{\Xi}$ be a d-dimensional vector of basis random variables defined on a probability space $(\Omega, \mathcal{A}, P)$ with image measure $\boldsymbol{\mu}$ such that the corresponding orthogonal set of multivariate polynomials $\{P_{\boldsymbol{n}} : \boldsymbol{n} \in \mathbb{N}_0^d\}$ with respect to $\boldsymbol{\mu}$ (as in (2.35)) is complete. Assume that $\boldsymbol{\Theta} \in L^2(\Omega, \sigma(\boldsymbol{\Xi}), P)$. Then, there exists a function $g_{\boldsymbol{\Theta}} = (g_{\Theta_1}, \ldots, g_{\Theta_d}) : \mathbb{R}^d \to \mathbb{R}^d$, so that each component of $\boldsymbol{\Theta}$ has due to (2.38) an expansion of the form

$$\Theta_i = g_{\Theta_i}(\boldsymbol{\Xi}) = \sum_{\boldsymbol{n} \in \mathbb{N}_0^d} \lambda_{i;\boldsymbol{n}} P_{\boldsymbol{n}}(\boldsymbol{\Xi}) h_{\boldsymbol{n}}, \ i = 1, \ldots, d. \tag{2.49}$$

The N-th order truncated expansion for $\Theta_i, \ i = 1, \ldots, d$ will be denoted as $(\Theta_i)_N$

$$(\Theta_i)_N = \sum_{|\boldsymbol{n}|=0}^{N} \lambda_{i;\boldsymbol{n}} P_{\boldsymbol{n}}(\boldsymbol{\Xi}) h_{\boldsymbol{n}}, \ i = 1, \ldots, d, \ N \in \mathbb{N}_0. \tag{2.50}$$

An important step in the formulation of stochastic systems is the parametrization of the unknown parameters $\boldsymbol{\Theta}$ in terms of the basis random vector. In the one dimensional case, i.e. d=1, the *isoprobabilistic transformation* used in random number generation provides such a parametrization. Its multi-dimensional generalization is the *Rosenblatt transformation* [110]. Both transformations are based on the distribution functions of the involved random variables.

Theorem 2.15. *(Isoprobabilistic transformation) If Θ is a continuous random variable with values in $\mathbb{R}$, distribution function F_Θ and inverse F_Θ^{-1}, then the random variable $U = F_\Theta(\Theta)$ is uniformly distributed on $[0, 1]$. If $X = F_\Theta^{-1}(U)$, then the distribution of X is F_Θ.*

Combining the above, one easily sees that the random variable $F_\Theta^{-1}(F_\Xi(\Xi))$ has the distribution function F_Θ, and is thus a copy of Θ.

Theorem 2.16. *(Rosenblatt) Let $\boldsymbol{X} = (X_1, \ldots, X_d)$ be a random vector with distribution function $F : \mathbb{R}^d \to \mathbb{R}$ and let $\boldsymbol{U} = (U_1, \ldots, U_d)$ be a vector of independent and uniformly distributed random variables. Define the random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_d) = T(\boldsymbol{U})$ recursively by the equations*

$$\begin{aligned} Y_1 &= F_1^{-1}(U_1) \\ Y_i &= F_{i|1,\ldots,i-1}^{-1}(U_i \mid U_1, \ldots, U_{i-1}), \ 2 \le i \le d. \end{aligned} \tag{2.51}$$

Then, the distribution function of $\boldsymbol{Y}$ is F.

In practice, the Rosenblatt transformation is not analytically tractable as it relies on the conditional inverse distribution functions, which are in general not known analytically. In chapter 3 a method to overcome this difficulty is presented.

Return now to the quantity $\boldsymbol{x}$. Assume from now on that $r = 1$ and drop the bold notation of $\boldsymbol{x}$ for simplicity. If $r > 1$, what follows should be applied to each component of $\boldsymbol{x}$. Under the previous assumptions, the solution $x(t, \boldsymbol{\Theta})$ of (2.48) is a stochastic process measurable with respect to $\sigma(\boldsymbol{\Xi})$ for each fixed time $t \in J$. If we assume that it belongs to the space $L^2(\Omega, \sigma(\boldsymbol{\Xi}), P)$ for all times $t \in J$, then again from (2.38) it admits the following polynomial expansion

$$x(t, \boldsymbol{\Theta}) = x(t, g_{\boldsymbol{\Theta}}(\boldsymbol{\Xi})) \equiv x(t, \boldsymbol{\Xi}) = \sum_{\boldsymbol{n} \in \mathbb{N}_0^d} q_{\boldsymbol{n}}(t) P_{\boldsymbol{n}}(\boldsymbol{\Xi}) h_{\boldsymbol{n}}, \tag{2.52}$$

where the coefficients are defined by (2.39)

$$q_{\boldsymbol{n}}(t) = \int_{\mathbb{R}^d} x(t, \boldsymbol{\xi}) P_{\boldsymbol{n}}(\boldsymbol{\xi}) d\boldsymbol{\mu}(\boldsymbol{\xi}),\ \boldsymbol{n} \in \mathbb{N}_0^d. \tag{2.53}$$

Equation (2.53) is of no practical use as it includes the (in general) unknown solution. The two main numerical approaches to compute these coefficients are *intrusive* or *spectral Galerkin* [61] and *non-intrusive* methods [129,130]. In the next subsections, these approaches are reviewed.

2.3.1 Stochastic Galerkin methods

Intrusive methods are based on a weak formulation of the original stochastic problem.

Assume that one looks for an approximation $x_N(t, \boldsymbol{\Xi})$ of $x(t, \boldsymbol{\Xi})$ in the finite dimensional polynomial subspace $\mathcal{P}^N$ defined in (2.42), where

$$x_N(t, \boldsymbol{\Xi}) = \sum_{|\boldsymbol{n}|=0}^{N} q_{\boldsymbol{n}}(t) P_{\boldsymbol{n}}(\boldsymbol{\Xi}) h_{\boldsymbol{n}},\ N \in \mathbb{N}_0. \tag{2.54}$$

The method consists of two main steps: one first introduces the truncated polynomial expansions (2.54) and (2.50) in the governing equations (2.48) and then projects the resulting residual on each basis polynomial in $\mathcal{P}^N$. This process leads to a deterministic coupled differential equation system satisfied by the PC coefficients $\{q_{\boldsymbol{n}}(t) \colon |\boldsymbol{n}| \leq N\}$. Standard numerical methods such as Runge-Kutta methods [116] can be then employed for its numerical integration with initial conditions that are determined by the initial conditions of the original ODE system (2.48). More precisely one requires that

$$E[(\dot{x}_N(t, \boldsymbol{\Xi}) - f(t, x_N, \boldsymbol{\Xi}))\, P_{\boldsymbol{n}}] = 0,\ \forall\, \boldsymbol{n} \in \mathbb{N}_0^d,\ |\boldsymbol{n}| \leq N. \tag{2.55}$$

The stochasticity in these equations will be integrated out and the resulting deterministic ODE system will be of dimension $\dim \mathcal{P}^N$.

The Galerkin approach is optimal in the mean square sense as the only error introduced is that resulting from truncating the infinite series expansion of the solution. A difficulty associated with this approach is that the derivation of the Galerkin system can be nontrivial in situations in which the original system is nonlinear. Furthermore, the method results in a system of higher dimension than the initial deterministic one and to determine the PC coefficients one has to solve a coupled system of equations. In the next subsection, it will be shown that in the case of stochastic non-intrusive methods each coefficient can be determined independently of the others.

2.3.2 Non-intrusive methods

Methods that fall in the class of non-intrusive methods include non-intrusive spectral projections, least squares approximation and collocation methods. The common feature underlying all these methods is that one solves the deterministic system for a set of realizations of the basis random vector and infers the PC expansion of the solution from the ensemble of the corresponding solution realizations.

Non-intrusive spectral projections (NISP)

The basic idea of NISP approaches is to numerically approximate the multidimensional integrals in the definition (2.53) of the PC coefficients. Existing methods can be categorized in stochastic methods which include Monte Carlo integration and similar techniques and deterministic methods which are based on numerical cubature. Once again, assume that one is looking for an approximation x_N of x in $\mathcal{P}^N$ as in (2.54).

In Monte Carlo integration, one generates a random sample $\{\boldsymbol{\xi}^m\}_{m=1,\dots,M}$ from the distribution of the basis random vector $\boldsymbol{\Xi}$ and then the PC coefficients of the solution are approximated by the sums

$$q_{\boldsymbol{n}}(t) \approx \frac{1}{M} \sum_{m=1}^{M} x(t, \boldsymbol{\xi}^m) P_{\boldsymbol{n}}(\boldsymbol{\xi}^m), \ |\boldsymbol{n}| \leq N, \ N \in \mathbb{N}_0. \tag{2.56}$$

The convergence rate of the Monte Carlo integration is $O(M^{-\frac{1}{2}})$ [30]. The main advantage of this approach is that this rate does not depend on the stochastic dimension of the problem. This makes the method appealing for high-dimensional integrals. Methods which improve the above rate of convergence include quasi Monte Carlo and Latin Hypercube methods. The reader is referred to [30, 95] for more details.

Alternatively and when the stochastic dimension d is not too large, deterministic integration methods can be used. The integrals in (2.53) can be approximated by the sums

$$q_{\boldsymbol{n}}(t) \approx \sum_{k=1}^{K} x(t, \boldsymbol{\xi}^k) P_{\boldsymbol{n}}(\boldsymbol{\xi}^k) w_k, \ |\boldsymbol{n}| \leq N, \ N \in \mathbb{N}_0, \tag{2.57}$$

where $\{\boldsymbol{\xi}^k\}_{k=1,\dots,K}$, $w_k > 0$ are the nodes and weights respectively of a d-dimensional cubature rule. Cubature rules can be constructed by the tensorization of one-dimensional quadrature rules. For more on numerical quadrature, see for example [116]. Compared to Monte Carlo methods, quadrature formulas suffer from the *curse of dimensionality*, i.e. the number of nodes on which the solution has to be evaluated grows very rapidly with the stochastic dimension. To overcome this difficulty, sparse cubature rules and adaptive sparse cubature rules have been developed [58, 113].

Least squares estimation

Another way to estimate the PC coefficients is via the solution of a least squares problem. Assume that one has realizations of the basis random variables $\{\boldsymbol{\xi}^l\}_{l=1,\dots,L}$ and the corresponding realizations of the solution $\{x(t, \boldsymbol{\xi}^l)\}_{l=1,\dots,L}$. One can estimate the PC coefficients by minimizing over the real numbers the residual

$$\sum_{l=1}^{L} \left| x(t, \boldsymbol{\xi}^l) - \sum_{|\boldsymbol{n}|=0}^{N} q_{\boldsymbol{n}}(t) P_{\boldsymbol{n}}(\boldsymbol{\xi}^l) h_{\boldsymbol{n}} \right|^2 \tag{2.58}$$

for each fixed time $t \in J$. As in the classical regression, the solution to this optimization problem for fixed time t is given by

$$\tilde{q}_N(t) = (D^T D)^{-1} D^T x^L(t, \boldsymbol{\xi}) \tag{2.59}$$

where $\tilde{q}_N(t) = (\tilde{q}_0(t), \ldots, \tilde{q}_N(t))^T$ is the vector of the estimated PC coefficients $\{q_{\boldsymbol{n}}(t) \colon \boldsymbol{n} \leq N\}$ in single index notation, $x^L(t, \boldsymbol{\xi}) = (x(t, \boldsymbol{\xi}^1), \ldots, x(t, \boldsymbol{\xi}^L))^T$ is the vector of the solution realizations at time t and the matrix $D \in \mathbb{R}^{L \times \dim \mathcal{P}^N}$ is the design matrix of the regression problem with elements $D_{i,j} = P_j(\boldsymbol{\xi}_i), i = 1, \ldots, L, j = 1, \ldots, \dim \mathcal{P}^N$, where again the single index notation is used. The residual

$$\text{res}(t, \boldsymbol{\xi}) = x(t, \boldsymbol{\xi}) - \sum_{|\boldsymbol{n}|=0}^{N} \tilde{q}_{\boldsymbol{n}}(t) P_{\boldsymbol{n}}(\boldsymbol{\xi}) h_{\boldsymbol{n}} \tag{2.60}$$

of the least squares problem will be orthogonal to $\mathcal{P}^N$ only in the limit $L \to \infty$. The choice of the points $\{\boldsymbol{\xi}^l\}_{l=1,\ldots,L}$ in the stochastic space is important for this method. In [20, 22] tensored Gaussian points were selected, while in [36] the points were selected at random based on the distribution of the basis random vector. More sophisticated techniques for selecting these points have been developed. They rely on experimental design to optimize certain properties of the design matrix [67, 105].

Collocation methods

In these methods, one seeks a polynomial approximation $\hat{x}_N$ of x, which interpolates the solution for a given time t at given points $(\boldsymbol{\xi}^i, x(t, \boldsymbol{\xi}^i))_{i=1,\ldots,I}$, i.e. one requires that

$$\hat{x}_N(t, \boldsymbol{\xi}^i) = x(t, \boldsymbol{\xi}^i), \ i = 1, \ldots, I. \tag{2.61}$$

A difference with the previous methods is that in this case the polynomial basis does not need to be pre-described but it can be defined by the points $\{\boldsymbol{\xi}^i\}_{i=1,\ldots,I}$. This can be achieved for example if one assumes the following form for the polynomial $\hat{x}_N$

$$\hat{x}_N(t, \boldsymbol{\xi}) = \sum_{|\boldsymbol{n}|=0}^{N} \hat{q}_{\boldsymbol{n}}(t) L_{\boldsymbol{n}}(\boldsymbol{\xi}), \ N \in \mathbb{N}_0, \tag{2.62}$$

where $L_{\boldsymbol{n}}(\boldsymbol{\xi})$ are the multi-dimensional Lagrange polynomials. These are the tensor products of the one-dimensional Lagrange polynomials $\{L_i(\xi)\}_{i=1,\ldots,I}$, which in classical approximation theory are defined for a given set of nodes $\{\xi^i\}_{i=1,\ldots,I}$ by the equations

$$L_j(\xi) = \prod_{\substack{i=1 \\ i \neq j}}^{I} \frac{\xi - \xi^i}{\xi^j - \xi^i}, \ j = 1, \ldots, I. \tag{2.63}$$

High-dimensional interpolation suffers like high-dimensional integration from the curse of dimensionality. Other methods, such as sparse collocation and adaptive collocation methods can also be applied as examined in [8, 98, 129].

Finally, it is noted that the main advantage of non-intrusive methods compared to Galerkin approaches is their ease of implementation: in the former, a deterministic code that solves the initial problem can be used, whereas in the latter, the equations of the Galerkin

system have to be formulated and new code for the coupled system is needed. Furthermore, the non-intrusive methods can easily deal with nonlinearities in the initial system. Despite this relative advantage, these methods are not always preferred as they are not optimal: apart from the truncation error introduced by the approximation of the quantity of interest in a finite dimensional space, another error, the *aliasing error* is introduced by the numerical integration or interpolation used [65,70].

3 Optimal maps and polynomial chaos expansions

This chapter deals with the problem of the representation of parametric uncertainty in terms of a finite number of independent basic random variables. Firstly, the problem is formulated using the notation introduced in the previous chapter. After a short review of previous work on the problem and basic tools to be used in the sequence, results are stated for the one-dimensional and the general case.

Problem formulation Let $(\Omega, \mathcal{A}, P)$ be a probability space and let $\boldsymbol{\Xi}\colon (\Omega, \mathcal{A}) \to (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ be a basic random vector as in (2.34) and $\boldsymbol{X}\colon (\Omega, \sigma(\boldsymbol{\Xi})) \to (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ another random vector of the same dimension as $\boldsymbol{\Xi}$ with finite variance. Given a sample $\{\boldsymbol{X}^m\}_{m=1,\dots,M}$ from the distribution of $\boldsymbol{X}$, estimate the coefficients $\widehat{g_X}_{\boldsymbol{n}}$ in the PC expansion

$$\boldsymbol{X} = g_X(\boldsymbol{\Xi}) = \sum_{\boldsymbol{n}\in\mathbb{N}_0^d} \widehat{g_X}_{\boldsymbol{n}} P_{\boldsymbol{n}}(\boldsymbol{\Xi}) h_{\boldsymbol{n}}, \tag{3.1}$$

up to a given order N, i.e. for all multi-indices $\boldsymbol{n} \in \mathbb{N}_0^d$ such that $|\boldsymbol{n}| \leq N$.

In the general multivariate case, it will be assumed that the random vectors $\boldsymbol{\Xi}$ and $\boldsymbol{X}$ are continuous and with compact supports. In the one-dimensional case, it will be further assumed that the distribution function F_X of X is twice differentiable and its density f_X is bounded away from zero on its support S_X, i.e. it is assumed that

$$\exists\, c \in \mathbb{R}\colon f_X(x) > c > 0, \forall\, x \in S_X. \tag{3.2}$$

For simplicity, the same polynomial system will be used in each stochastic dimension.

3.1 Previous results

As stated already in chapter 2, a basic result related to this problem is the Rosenblatt transformation [110]. It provides an explicit form for the function g_X which appears in the definition of the PC coefficients. This transformation is defined by conditional probability distribution functions. These functions are in general not known analytically and the transformation has to be approximated numerically based on a sample from the random variable $\boldsymbol{X}$. This approach was followed in [39], where the joint probability distribution function of $\boldsymbol{X}$ was estimated by employing the maximum entropy principle and a nonlinear least squares method. In [40], the same transformation was again used but two different approaches were there presented. In the first one, the joint probability distribution function was estimated by linear interpolation of the histogram of the observations and in the second approach, the authors estimated the marginal probability density functions and then the Spearman's rank correlation coefficient was used to capture the dependencies between the components of the vector $\boldsymbol{X}$. As mentioned in [39], one has to note that

the transformation depends on the ordering of the random variables. Thus, Rosenblatt actually provides with $d!$ transformations corresponding to all possible permutations of the components of $\boldsymbol{X}$. An obstacle of this method is that it breaks down for high dimensions, as conditioning on a discrete sample becomes numerically unstable.

Other existing approaches to the same problem are based on maximum likelihood estimation. In [45, 46] a surrogate model for the likelihood of the observations was used to simplify the computations and in [60] the authors used the method of simulated annealing to maximize this surrogate likelihood function. In [115] the likelihood was maximized by using a random search algorithm. Methods based on Bayesian estimation are stated in [2, 59]. In both papers, a kernel density estimation method is used to estimate the likelihood function of the coefficients given the data. Note that all the above authors are dealing with the estimation of PC coefficients for the random variables appearing in the KL expansion of a random field. The special properties of these random variables and especially the fact that they are uncorrelated form a central ingredient of these approaches.

The method proposed here relies on the matching of two samples from the distribution functions of $\boldsymbol{\Xi}$ and $\boldsymbol{X}$ by using a discrete transformation which converges to a continuous transformation g_X and such that it respects the underlying image measures. The discrete transformation is combined with a regression approach in order to estimate the coefficients. It is shown that these estimates are asymptotically consistent for the coefficients in the series expansion of a specific copy of $\boldsymbol{X}$. Recall that an estimate θ_M based on a sample of size M of a quantity θ is called *consistent*, if $\theta_M \xrightarrow{P} \theta$ as $M \to \infty$. Here one is making no assumption on the correlation of the random vectors. Thus, it can be used in both the cases of correlated and uncorrelated random vectors.

For the one-dimensional case the method relies on the theory of order statistics. As there is no total order in $\mathbb{R}^d$ for $d \geq 2$ and no appropriate definition of multivariate order statistics, the multi-dimensional case will be handled with results from optimal transportation (OT) theory. This theory has been recently applied independently by other authors in different problems related to uncertainty quantification. In [97] it was used to provide an alternative method for sampling from a posterior distribution function in a Bayesian framework and in [107] it was used as an alternative to importance sampling and related methods.

3.2 Order statistics

This section includes basic definitions and results from the theory of order statistics. For more details see in [111].

Let X be a continuous random variable with density f_X and distribution function F_X. For $0 < p < 1$, the *p-th quantile* of F_X is defined as

$$F_X^{-1}(p) = \zeta_p = \inf\{x \in \mathbb{R} \colon F_X(x) \geq p\}. \tag{3.3}$$

Let $\{X^m\}_{m=1,\dots,M}$ be an independent sample from the distribution function F_X. The *empirical distribution function* of X is then defined as the stochastic process

$$F_{X,M}(x) = \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}\{X^m \leq x\},\ x \in \mathbb{R}, \tag{3.4}$$

and the *sample p-th quantile* as

$$\hat{\zeta}_{p,M} = F^{-1}_{X,M}(p),\ 0 < p < 1. \tag{3.5}$$

A measure of closeness of $F_{X,M}$ to F_X is the *Kolmogorov-Smirnov distance* defined by

$$D_M = \sup_{x\in\mathbb{R}} |F_{X,M}(x) - F_X(x)|. \tag{3.6}$$

It can be shown that as $M \to \infty$, $D_M \overset{\text{wp1}}{\to} 0$ [121]. The following theorem due to Dvoretzky, Kiefer and Wolfowitz provides the rate of this convergence [51]. The proof for the sharp constant $K = 1$ can be found in [94].

Theorem 3.1. *Let F_X be a distribution function defined on $\mathbb{R}$. Then, for all $M \in \mathbb{N}$ and all $d > 0$*

$$P(D_M > d) \leq Ke^{-2Md^2}. \tag{3.7}$$

An equivalent formulation to the empirical distribution function are the *order statistics*. These are defined as the ordered sample values

$$X^{(1M)} \leq X^{(2M)} \leq \ldots \leq X^{(MM)}. \tag{3.8}$$

One has the relation

$$X^{(mM)} = \hat{\zeta}_{m/M,M},\ m = 1, \ldots, M. \tag{3.9}$$

The following theorem is proved in [9].

Theorem 3.2. *(Bahadur) Let $0 < p < 1$. Suppose that F_X is twice differentiable at ζ_p and $F'_X(\zeta_p) = f(\zeta_p) > 0$. Then*

$$\hat{\zeta}_{p,M} = \zeta_p + \frac{p - F_{X,M}(\zeta_p)}{f_X(\zeta_p)} + R_M, \tag{3.10}$$

where

$$R_M = O_{wp1}\left(M^{-3/4}(logM)^{3/4}\right),\ M \to \infty. \tag{3.11}$$

REMARK One says that a sequence of random variables $(X_n)_{n\in\mathbb{N}}$ is $O_{wp1}(a_n)$ as $n \to \infty$ for a real sequence $(a_n)_{n\in\mathbb{N}}$, if the sequence X_n/a_n is bounded with probability 1.

3.3 Optimal transportation

The theory of optimal transportation is dealing with the existence and characterization of maps $T\colon \mathbb{R}^d \to \mathbb{R}^d$ such that $\boldsymbol{X} = T(\boldsymbol{\Xi})$. These maps are the solution to a given minimization problem, which involves the image measures $\boldsymbol{\mu} = (\boldsymbol{\Xi})_{\#}P$ and $\boldsymbol{\nu} = (\boldsymbol{X})_{\#}P$ of $\boldsymbol{\Xi}$ and $\boldsymbol{X}$ respectively. Some results from the theory of optimal transportation on Euclidean spaces used below are summarized in the present section. See in [123,124] for more details on the topic and for the proofs of the stated results.

3.3.1 The continuous case on Euclidean spaces

Given two probability measures $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ on the measurable space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, denote by $\Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ the set of probability measures defined on $(\mathbb{R}^d \times \mathbb{R}^d, \mathcal{B}(\mathbb{R}^d) \times \mathcal{B}(\mathbb{R}^d))$ so that each $\boldsymbol{\pi} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ has marginal measures $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$. Given a measurable *cost* function $c\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, the *Kantorovich* optimal transportation problem reads

$$I[\boldsymbol{\pi}] = \int_{\mathbb{R}^d \times \mathbb{R}^d} c(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{\pi}(\boldsymbol{x}, \boldsymbol{y}) \to \inf_{\boldsymbol{\pi} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})}, \tag{3.12}$$

where $I[\boldsymbol{\pi}]$ is called the *total transportation cost* of $\boldsymbol{\pi}$. The measures $\boldsymbol{\pi}$ for which the infimum is attained are called *optimal transference plans*. If $\boldsymbol{\pi}$ can be represented as

$$d\boldsymbol{\pi}(\boldsymbol{x}, \boldsymbol{y}) = d\boldsymbol{\mu}(\boldsymbol{x}) \delta[\boldsymbol{y} = T(\boldsymbol{x})],\ \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d, \tag{3.13}$$

where $T\colon \mathbb{R}^d \to \mathbb{R}^d$ denotes a measurable function such that $\boldsymbol{\nu} = (T)_{\#}\boldsymbol{\mu}$, then T is called a *transport map* and $\boldsymbol{\pi}$ will be denoted in this case also by $\boldsymbol{\pi} = (Id, T)_{\#}\boldsymbol{\mu}$. The *Monge* formulation of the optimal transportation problem reads

$$I[T] = \int_{\mathbb{R}^d} c(\boldsymbol{x}, T(\boldsymbol{x})) d\boldsymbol{\mu}(\boldsymbol{x}) \to \inf_{T\colon\, \boldsymbol{\nu} = T_{\#}\boldsymbol{\mu}}. \tag{3.14}$$

The difference between transference plans and transport maps is that the former allow mass located in a point $\boldsymbol{x}$ to be split and distributed over several $\boldsymbol{y}$ locations, while the latter requires that the whole mass in $\boldsymbol{x}$ is transported to a unique location $\boldsymbol{y}$. This fact indicates that the Kantorovich problem is a relaxed version of the Monge problem. In contrast to transference plans, transport maps do not always exist. Consider for example the case where $\boldsymbol{\mu}$ is a Dirac delta and $\boldsymbol{\nu}$ a measure continuous with respect to the Lebesgue measure. Then, no transport map exists: the only way of transporting the mass concentrated on the Dirac delta is to split it and distribute it over the support of $\boldsymbol{\nu}$.

The existence of optimal maps depends on the cost function and on the regularity of the measures $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$. For example, consider cost functions of the form $c(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_E^p$, with $0 < p < \infty$, where $\|\cdot\|_E$ stands for the Euclidean distance in $\mathbb{R}^d$. One can then show that if $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are continuous with respect to the Lebesgue measure, then there is a unique optimal transference plan and a unique optimal map if $p > 1$, whereas for $p < 1$, there is in general no optimal transport map, although there is an optimal transference plan.

For the special case $p = 2$ and for the quadratic cost function

$$c(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_E^2,\ \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d, \tag{3.15}$$

the following results on existence, uniqueness and stability of optimal transference plans and transport maps hold.

Proposition 3.3. *The minimization problem (3.12) admits a minimizer. This means that there exists an optimal transference plan.*

A central result in OT theory is the theorem of Brenier, which gives a characterization of optimal maps for the quadratic cost function in $\mathbb{R}^d$. Before the main theorem is stated, the notion of a small set is defined.

Definition 3.4. *A measurable set $A \subset \mathbb{R}^d$ is a small set if it has Hausdorff dimension at most $d-1$.*

Note, that a measure $\boldsymbol{\mu}$ does not give mass to small sets if it is for example continuous with respect to the Lebesgue measure.

Theorem 3.5. *(Brenier) Assume that the measures $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ have finite second order moments. If $\boldsymbol{\mu}$ does not give mass to small sets, then the optimal transference plan $\boldsymbol{\pi}$ for the cost function in (3.15) is unique and $\boldsymbol{\pi} = (Id \times \nabla\phi)_{\#}\boldsymbol{\mu}$, where $\boldsymbol{\nu} = (\nabla\phi)_{\#}\boldsymbol{\mu}$ and $\nabla\phi$ is the unique $\boldsymbol{\mu}$-a.e. gradient of a convex function $\phi\colon \mathbb{R}^d \to \mathbb{R}^d$. The function $\nabla\phi$ is the unique optimal transport map.*

Note that in general there is no closed form solution for the optimal transport map in Theorem 3.5.

Theorem 3.6. *Let $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ be two probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Consider two sequences of probability measures $(\boldsymbol{\mu}_n)_{n\in\mathbb{N}}$ and $(\boldsymbol{\nu}_n)_{n\in\mathbb{N}}$ such that $\boldsymbol{\mu}_n \rightharpoonup \boldsymbol{\mu}$ and $\boldsymbol{\nu}_n \rightharpoonup \boldsymbol{\nu}$, as $n \to \infty$. Denote by $\boldsymbol{\pi}_n$ an optimal transference plan between $\boldsymbol{\mu}_n$ and $\boldsymbol{\nu}_n$ for the cost in (3.15). If*

$$\forall\, n \in \mathbb{N},\ I[\boldsymbol{\pi}_n] < +\infty \text{ and } \liminf_{n\to\infty} I[\boldsymbol{\pi}_n] < \infty, \tag{3.16}$$

then there exists a transference plan $\boldsymbol{\pi} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$, which is optimal for the measures $\boldsymbol{\mu}$, $\boldsymbol{\nu}$ and the quadratic cost. Furthermore, there exists a subsequence $(\pi_{n_k})_{k\in\mathbb{N}}$ of $(\pi_n)_{n\in\mathbb{N}}$ such that $\boldsymbol{\pi}_{n_k} \rightharpoonup \boldsymbol{\pi}$, as $k \to \infty$.

As for the transport maps the following result holds.

Corollary 3.7. *Let $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ be two probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ such that the support of $\boldsymbol{\nu}$ is a closed subset of $\mathbb{R}^d$. With the notation and assumptions of the previous theorem, assume furthermore that there exist measurable maps $T_n, T\colon \mathbb{R}^d \to \mathbb{R}^d$ such that*

$$\forall\, n \in \mathbb{N},\ \boldsymbol{\pi}_n = (Id, T_n)_{\#}\boldsymbol{\mu}_n \text{ and } \boldsymbol{\pi} = (Id, T)_{\#}\boldsymbol{\mu},$$

and that $\boldsymbol{\pi}$ is unique. Then,

$$\forall\, \varepsilon > 0,\ \lim_{n\to\infty} \boldsymbol{\mu}_n \Big[\{\boldsymbol{x} \in \mathbb{R}^d\colon \|T_n(\boldsymbol{x}) - T(\boldsymbol{x})\|_E > \varepsilon\}\Big] = 0.$$

3.3.2 The discrete case on Euclidean spaces

Consider next the transportation problem for the case of two discrete and equally weighted measures $\boldsymbol{\mu}_n$ and $\boldsymbol{\nu}_n$,

$$\boldsymbol{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\boldsymbol{x}_i} \text{ and } \boldsymbol{\nu}_n = \frac{1}{n}\sum_{j=1}^{n} \delta_{\boldsymbol{y}_j},\ \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \in \mathbb{R}^d. \tag{3.17}$$

The joint measures $\boldsymbol{\pi} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ are now $n \times n$ matrices A with nonnegative entries and such that the sum of their rows and columns are equal to $1/n$. Each element a_{ij} of such a matrix A gives the probability that the joint measure $\boldsymbol{\pi}$ assigns to the point $(\boldsymbol{x}_i, \boldsymbol{y}_j)$. By rescaling the matrices A, the set of joint measures $\boldsymbol{\pi} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ can be represented by the set $\mathbf{B}$ of bistochastic $n \times n$ matrices such that each matrix $B \in \mathbf{B}$ has real elements $b_{ij} \in [0, 1]$.

Recall that a *bistochastic* matrix is a matrix with nonnegative elements and such that its rows and columns sum up to 1.

In the discrete setting considered here, minimizing the total transportation cost $I[\boldsymbol{\pi}]$ in the Kantorovich formulation, equation (3.12), is equivalent to minimizing the following cost

$$I[B] = \frac{1}{n} \sum_{i,j=1}^{n} b_{ij} c(\boldsymbol{x}_i, \boldsymbol{y}_j) \to \min_{B \in \mathbf{B}}. \tag{3.18}$$

It can be shown that the minimizers of this linear optimization problem defined on the bounded, convex set $\mathbf{B}$ are permutation matrices, i.e. stochastic matrices with $b_{ij} = \delta_{i\sigma(i)}$, where $\sigma \in \mathcal{S}_n$, the set of all permutations of $\{1, \ldots, n\}$ [21]. Therefore, in the case of equally weighted discrete measures, every solution of the Kantorovich problem corresponds to a solution of the following problem

$$I[\sigma] = \frac{1}{n} \sum_{i=1}^{n} c(\boldsymbol{x}_i, \boldsymbol{y}_{\sigma(i)}) \to \min_{\sigma \in \mathcal{S}_n}. \tag{3.19}$$

The permutation $\overline{\sigma}$ which minimizes $I[\sigma]$ defines a map $T\colon \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \to \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ by

$$T(\boldsymbol{x}_i) = \boldsymbol{y}_{\overline{\sigma}(i)},\ i = 1, \ldots, n, \tag{3.20}$$

which is the optimal map for the Monge problem involving the measures $\boldsymbol{\mu}_n$ and $\boldsymbol{\nu}_n$.

The discrete transportation problem is a special type of a network optimization problem with linear cost function and it is called the *assignment problem* in the related literature. In this context, the locations $\boldsymbol{x}_i, i = 1, \ldots, n$ are considered as *persons* that one wishes to match with n *objects* $\boldsymbol{y}_j, j = 1, \ldots, n$, such that the total cost

$$J[\sigma] = \sum_{i=1}^{n} s(\boldsymbol{x}_i, \boldsymbol{y}_{\sigma(i)}) \to \max_{\sigma \in \mathcal{S}_n}, \tag{3.21}$$

is maximized. Here, $s(\boldsymbol{x}_i, \boldsymbol{y}_j)$ is the *value* (or *benefit*) for matching person $\boldsymbol{x}_i$ with the object $\boldsymbol{y}_j$. This equation is of the same type as the discrete Monge problem in (3.19). For the case of the quadratic cost function considered here, minimizing $I[\sigma]$ for the cost $c(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_E^2$ is the same as maximizing $J[\sigma]$ for the cost $s(\boldsymbol{x}, \boldsymbol{y}) = \langle \boldsymbol{x}, \boldsymbol{y} \rangle_E$, where $\langle \cdot, \cdot \rangle_E$ is the Euclidean inner product in $\mathbb{R}^d$.

An efficient algorithm to solve the discrete optimal transportation problem is the *auction algorithm* first proposed by Bertsekas in [17]. Of course, the linear optimization problem in (3.18) can be also solved by linear programming techniques such as simplex methods, see for example in [120]. The auction algorithm is more efficient as it takes into account the network structure of the problem.

3.3.3 The auction algorithm

The auction algorithm is here shortly described. It is one of the three main types of algorithms for linear and network flow optimization problems, see in [18, 26] for more details.

It is related to the *dual* of the assignment problem: here, one seeks for an optimal *price* vector $\boldsymbol{p} = (p_1, \ldots, p_n)$ such that the total cost

$$\sum_{i=1}^{n} \max_{j=1,\ldots,n} \{s(\boldsymbol{x}_i, \boldsymbol{y}_j) - p_j\} + \sum_{i=1}^{n} p_j \tag{3.22}$$

is minimized. If the following condition

$$s(\boldsymbol{x}_i, \boldsymbol{y}_{\sigma(i)}) - p_{\sigma(i)} = \max_{j=1,\ldots,n} \{s(\boldsymbol{x}_i, \boldsymbol{y}_j) - p_j\} \tag{3.23}$$

is satisfied by a permutation σ for all $i = 1, \ldots, n$, then one says that the assignment σ and the prices $\boldsymbol{n}$ satisfy *complementary slackness (CS)*. In this case it can be shown that this permutation is optimal for the assignment (discrete transportation) problem and the corresponding prices are optimal for the dual problem.

The auction algorithm is based on a related property called *ε-complementary slackness (ε-CS)* and on *partial* assignments, which are assignments in which only a subset of the n persons is assigned to objects. One says that a partial assignment and a vector of prices $\boldsymbol{p}$ satisfies ε-CS if

$$s(\boldsymbol{x}_i, \boldsymbol{y}_{\sigma(i)}) - p_{\sigma(i)} \geq \max_{j=1,\ldots,n} \{s(\boldsymbol{x}_i, \boldsymbol{y}_j) - p_j\} - \varepsilon, \ \varepsilon > 0. \tag{3.24}$$

The algorithm runs as follows: one starts with a price vector and a partial assignment satisfying ε-CS. One then chooses a person $\boldsymbol{x}_i$ which is still unassigned and assigns him to the object $\boldsymbol{y}_k$ where

$$k = \arg\max_{j=1,\ldots,n} \{s(\boldsymbol{x}_i, \boldsymbol{y}_j) - p_j\}. \tag{3.25}$$

If any person was already assigned to the object $\boldsymbol{y}_k$, he becomes unassigned. The price vector is changed by augmenting the element p_k by the factor $\gamma_i + \varepsilon$, where

$$\begin{aligned} \gamma_i &= v_i - w_i, \\ v_i &= \max_{j=1,\ldots,n} \{s(\boldsymbol{x}_i, \boldsymbol{y}_j) - p_j\}, \\ w_i &= \max_{j=1,\ldots,n, j \neq k} \{s(\boldsymbol{x}_i, \boldsymbol{y}_j) - p_j\}. \end{aligned} \tag{3.26}$$

The iterations continue until all persons are assigned. It can be shown that the algorithm ends with a permutation which total transportation cost is within $n\varepsilon$ from being optimal. In the special case where the values $s(\boldsymbol{x}_i, \boldsymbol{y}_j)$ are integers for all $i, j = 1, \ldots, n$ and $\varepsilon < \frac{1}{n}$, the auction algorithm yields the optimal transportation cost. The running time of the algorithm depends on the initial price vector, on the precision ε and on the maximal value $\max_{i,j=1,\ldots,n} |s(\boldsymbol{x}_i, \boldsymbol{y}_j)|$. It is noted finally that the algorithm is offered for parallel implementation and has a time complexity of $O(n^2 log n)$. For more details on computational issues, the reader is referred to [32].

REMARK

(i) In the one-dimensional case the permutation which solves the assignment problem and thus the discrete transportation problem corresponds to the matching of the ordered sample values and is the discrete analog of the integral transformation, see for example in [38]. This result justifies the choice of the optimal maps as the generalization of the order statistics in the multi-dimensional case.

(ii) An interesting result in the optimal transportation theory is that the Rosenblatt (or Knothe-Rosenblatt as is known in this field) transformation turns out to be the limit of the optimal transportation maps, each one occurring as the solution of a transportation problem with a special quadratic cost function [31].

(iii) Note that no transport map exists in the case that the discrete measures are not equally weighted, i.e. if

$$\boldsymbol{\mu}_n = \sum_{i=1}^{n} w_i \delta_{\boldsymbol{x}_i} \text{ and } \boldsymbol{\nu}_n = \sum_{j=1}^{n} \tilde{w}_j \delta_{\boldsymbol{y}_j},\ \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \in \mathbb{R}^d, \tag{3.27}$$

with $w_i, w_j \in (0, 1]$ and $w_i \neq \tilde{w}_j, i, j = 1, \ldots, n$. This is again because mass in points $\boldsymbol{x}_i, i = 1, \ldots, n$ will have to be split in order to be transported to mass in the points $\boldsymbol{y}_j, j = 1, \ldots, n$.

3.4 Estimation of PC coefficients

3.4.1 One-dimensional case

Assume in this section that $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are measures on the Euclidean space $\mathbb{R}$.

Linear case

Consider first the simple case where the random variables Ξ and X have the same distribution function F with density f and support S. Assume two independent samples $\{\Xi^m\}_{m=1,\ldots,M}$ and $\{X^m\}_{m=1,\ldots,M}$ from F are given. Denote by $\boldsymbol{\Xi}^{(M)} = (\Xi^{(1M)}, \ldots, \Xi^{(MM)})$ and $\boldsymbol{X}^{(M)} = (X^{(1M)}, \ldots, X^{(MM)})$ the vectors of the corresponding order statistics. Assume the following truncated polynomial regression model

$$X^{(mM)} = \sum_{n=0}^{N} q_{n,M} P_n(\Xi^{(mM)}) h_n,\ m = 1, \ldots, M,\ N \in \mathbb{N}_0. \tag{3.28}$$

This means that it is assumed that the underlying discrete map $T^M \colon \{X^1, \ldots, X^M\} \to \{\Xi^1, \ldots, \Xi^M\}$ is defined by the equations

$$T^M(\Xi^{(mM)}) = X^{(mM)},\ m = 1, \ldots, M. \tag{3.29}$$

The vector of the unknown coefficients $q_M = (q_{0,M}, \ldots, q_{N,M})$ is to be estimated by minimizing the residual

$$\sum_{m=1}^{M} |X^{(mM)} - \sum_{n=0}^{N} q_{n,M} P_n(\Xi^{(mM)}) h_n|^2 \to \min_{q_M \in \mathbb{R}^{N+1}}. \tag{3.30}$$

The design matrix of the regression model reads

$$D(\boldsymbol{\Xi}^{(M)}) = \begin{pmatrix} P_0(\Xi^{(1M)})h_0 & P_1(\Xi^{(1M)})h_1 & \ldots & P_N(\Xi^{(1M)})h_N \\ P_0(\Xi^{(2M)})h_0 & P_1(\Xi^{(2M)})h_1 & \ldots & P_N(\Xi^{(2M)})h_N \\ & & & \vdots \\ P_0(\Xi^{(MM)})h_0 & P_1(\Xi^{(MM)})h_1 & \ldots & P_N(\Xi^{(MM)})h_N \end{pmatrix} \tag{3.31}$$

The following proposition shows that the discrete transformation T^M defined in equation (3.29) converges to the identity operator as the sample size grows. Thus, the matching of the corresponding ordered statistics yields in the limit the expected result, as in this case one can assume that $X = g_X(\Xi) = \Xi$.

Proposition 3.8. *Assume that the distribution function F is twice differentiable and that the density f is bounded away from zero on the support S. Furthermore, assume that S is a compact subset of $\mathbb{R}$. Let $\tilde{q}_M$ be the solution to the least squares problem in (3.30)*

$$\tilde{q}_M = \left(D(\boldsymbol{\Xi}^{(M)})^T D(\boldsymbol{\Xi}^{(M)})\right)^{-1} D(\boldsymbol{\Xi}^{(M)})^T \boldsymbol{X}^{(M)}. \tag{3.32}$$

Then, in the limit $M \to \infty$, the quantities $\tilde{q}_M$ are well defined, the inverse $\left(D(\boldsymbol{\Xi}^{(M)})^T D(\boldsymbol{\Xi}^{(M)})\right)^{-1}$ exists and the underlying transformation g_X such that $X = g_X(\Xi)$ is the identity.

The proof is based on the following two lemmata.

Lemma 3.9. *Under the assumptions of Proposition 3.8, consider the random variable*

$$Z_M = \frac{1}{M} \sum_{m=1}^{M} \left(X^{(mM)} - \zeta_{\frac{m}{M}} \right), \tag{3.33}$$

where $\zeta_{\frac{m}{M}}$ is the $\frac{m}{M}$-th quantile of the distribution function F of X. Then $Z_M \xrightarrow{P} 0$, as $M \to \infty$.

Before the proof is given, some more notation is introduced. Denote by $[x]$ the *floor function*,

$$\forall\, x \in \mathbb{R},\ [x] = \max\{k \in \mathbb{Z} \colon k \leq x\}. \tag{3.34}$$

The following properties can be easily verified

$$[x] \leq x \leq [x] + 1 \text{ and } [x + k] = [x] + k,\ \forall\, k \in \mathbb{Z},\ x \in \mathbb{R}. \tag{3.35}$$

Proof. One has to show that $\forall\ \varepsilon > 0, \lim_{M\to\infty} P(|Z_M| > \varepsilon) = 0$. This is equivalent to showing that

$$\forall\, \varepsilon > 0\ \forall\, \delta > 0\ \exists\, M' \in \mathbb{N} \colon \forall\, M \geq M'\ P(|Z_M| > \varepsilon) < \delta.$$

Fix $\varepsilon > 0$, $\delta > 0$ and choose $p \in (0,1)$ such that $0 < p < \frac{\varepsilon}{8 \sup S}$. It holds,

$$\begin{aligned}
P(|Z_M| > \varepsilon) &= P\left(\left|\frac{1}{M}\sum_{m=1}^{M}\left(X^{(mM)} - \zeta_{\frac{m}{M}}\right)\right| > \varepsilon\right) \\
&= P\left(\left|\frac{1}{M}\sum_{m=1}^{M}\left(\hat{\zeta}_{\frac{m}{M},M} - \zeta_{\frac{m}{M}}\right)\right| > \varepsilon\right) \\
&\leq P\left(\frac{1}{M}\sum_{m=1}^{M}|\hat{\zeta}_{\frac{m}{M},M} - \zeta_{\frac{m}{M}}| > \varepsilon\right) \\
&\leq P\left(\frac{1}{M}\left[\sum_{m=1}^{[Mp]}|\hat{\zeta}_{\frac{m}{M},M} - \zeta_{\frac{m}{M}}| + \sum_{m=[Mp]+1}^{[M(1-p)]}|\hat{\zeta}_{\frac{m}{M},M} - \zeta_{\frac{m}{M}}| + \sum_{m=[M(1-p)]+1}^{M}|\hat{\zeta}_{\frac{m}{M},M} - \zeta_{\frac{m}{M}}|\right] > \varepsilon\right) \\
&\leq P\left(\frac{1}{M}2Mp\sup S + \frac{1}{M}\sum_{m=[Mp]+1}^{[M(1-p)]}|\hat{\zeta}_{\frac{m}{M},M} - \zeta_{\frac{m}{M}}| + \frac{1}{M}2Mp\sup S > \varepsilon\right) \\
&\leq P\left(\frac{1}{M}\sum_{m=[Mp]+1}^{[M(1-p)]}|\hat{\zeta}_{\frac{m}{M},M} - \zeta_{\frac{m}{M}}| > \varepsilon - 4p\sup S\right).
\end{aligned}$$

Let $\varepsilon' = \varepsilon - 4p\sup S > 0$. Then, it follows from Theorem 3.2 and equation (3.10) that

$$\begin{aligned}
P(|Z_M| > \varepsilon) &\leq P\left(\frac{1}{M}\sum_{m=[Mp]+1}^{[M(1-p)]}|\hat{\zeta}_{\frac{m}{M},M} - \zeta_{\frac{m}{M}}| > \varepsilon'\right) \\
&= P\left(\frac{1}{M}\sum_{m=[Mp]+1}^{[M(1-p)]}|\frac{m/M - F_M(\zeta_{\frac{m}{M}})}{f(\zeta_{\frac{m}{M}})} + R_M| > \varepsilon'\right).
\end{aligned}$$

From (3.35), it follows that

$$\frac{1}{M}([M(1-p)] - [Mp]) \leq \frac{1}{M}(M - 2Mp) = 1 - 2p.$$

Therefore,

$$\begin{aligned}
P(|Z_M| > \varepsilon) &\leq P\left(\frac{1}{M}\sum_{m=[Mp]+1}^{[M(1-p)]}\frac{|F(\zeta_{\frac{m}{M}}) - F_M(\zeta_{\frac{m}{M}})|}{f(\zeta_{\frac{m}{M}})} + (1-2p)|R_M| > \varepsilon'\right) \\
&\leq P\left(\frac{1}{M}\sum_{m=[Mp]+1}^{[M(1-p)]}\frac{|F(\zeta_{\frac{m}{M}}) - F_M(\zeta_{\frac{m}{M}})|}{f(\zeta_{\frac{m}{M}})} + (1-2p)C_R M^{-3/4}(logM)^{3/4} > \varepsilon'\right) \\
&\leq P\left(\sup_{\xi\in S}|F(\xi) - F_M(\xi)|\frac{1}{M}\sum_{m=[Mp]+1}^{[M(1-p)]}\frac{1}{f(\zeta_{\frac{m}{M}})} > \varepsilon' - (1-2p)C_R M^{-3/4}(logM)^{3/4}\right).
\end{aligned}$$

Consider now the term

$$\frac{1}{M}\sum_{m=[Mp]+1}^{[M(1-p)]}\frac{1}{f(\zeta_{\frac{m}{M}})} \tag{3.36}$$

and the interval $I_p = [\zeta_p, \zeta_{1-p}]$. In the sum given by (3.36), there are

$$[M(1-p)] - [Mp] \leq M - Mp - Mp = M - 2Mp < M$$

terms. Choose randomly $M - ([M(1-p)] - [Mp])$ points from the distribution of F conditioned on I_p and denote by $\zeta_k, k = 1, \ldots, M$ the set of points consisting of the points in I_p appearing in (3.36) along with the extra chosen points. Then, one has

$$\frac{1}{M} \sum_{m=[Mp]+1}^{[M(1-p)]} \frac{1}{f(\zeta_{\frac{m}{M}})} \leq \frac{1}{M} \sum_{k=1}^{M} \frac{1}{f(\zeta_k)},$$

as the density function f is assumed to be strictly positive. By the law of large numbers (Theorem 2.11),

$$\frac{1}{M} \sum_{k=1}^{M} \frac{1}{f(\zeta_k)} \overset{\text{wp1}}{\to} \int_{\zeta_p}^{\zeta_{1-p}} \frac{1}{f(x)} dx, \text{ as } M \to \infty.$$

Define

$$\int_{\zeta_p}^{\zeta_{1-p}} \frac{1}{f(x)} dx = C_{p,f}.$$

This integral exists, as it is assumed that f is bounded away from zero in the support S. Therefore, for $\varepsilon' > 0$, $\exists\ M'$ such that with probability 1 and for all $M \geq M'$, it holds

$$\frac{1}{M} \sum_{k=1}^{M} \frac{1}{f(\zeta_k)} \leq \int_{\zeta_p}^{\zeta_{1-p}} \frac{1}{f(x)} dx + \varepsilon' = C_{p,f} + \varepsilon'.$$

Thus, for all $M \geq M'$,

$$\begin{aligned} P(|Z_M| > \varepsilon) &\leq P\left(\sup_{\xi \in S} |F(\xi) - F_M(\xi)|(C_{p,f} + \varepsilon') > \varepsilon' - (1-2p) C_R M^{-3/4} (logM)^{3/4} \right) \\ &= P\left(\sup_{\xi \in S} |F(\xi) - F_M(\xi)| > \frac{\varepsilon' - (1-2p) C_R M^{-3/4} (logM)^{3/4}}{C_{p,f} + \varepsilon'} \right) \\ &\leq \exp\left(-2M \left(\frac{\varepsilon' - (1-2p) C_R M^{-3/4} (logM)^{3/4}}{C_{p,f} + \varepsilon'} \right)^2 \right), \end{aligned}$$

where Theorem 3.1 has been used. Here, C_R and $C_{p,f}$ are positive real numbers. The term in the exponential converges to 0 as $M \to \infty$. Therefore, for all $\delta > 0$, there exists a $M'' \in \mathbb{N}$ such that $\forall\ M > max(M', M'')$, $P(|Z_M| > \varepsilon) < \delta$ holds true. ■

Lemma 3.10. *Under the assumptions of Proposition 3.8, it holds that the random variable e_M defined as*

$$e_M = \frac{1}{M} \sum_{m=1}^{M} P_{k-1}(\Xi^{(mM)}) h_{k-1} \left(X^{(mM)} - \Xi^{(mM)} \right) \tag{3.37}$$

converges in probability to 0 as $M \to \infty$.

Proof. Let $\varepsilon > 0$. Then,

$$\begin{aligned}
P(|e_M| > \varepsilon) &= P\left(\left|\frac{1}{M}\sum_{m=1}^{M} P_{k-1}(\Xi^{(mM)})h_{k-1}\left(X^{(mM)} - \Xi^{(mM)}\right)\right| > \varepsilon\right) \\
&\leq P\left(\frac{1}{M}\sum_{m=1}^{M} |P_{k-1}(\Xi^{(mM)})h_{k-1}|\,|X^{(mM)} - \Xi^{(mM)}| > \varepsilon\right) \\
&\leq P\left(\|P_{k-1}\|_\infty h_{k-1}\frac{1}{M}\sum_{m=1}^{M} |X^{(mM)} - \Xi^{(mM)}| > \varepsilon\right) \\
&\leq P\left(\|P_{k-1}\|_\infty h_{k-1}\frac{1}{M}\sum_{m=1}^{M} |X^{(mM)} - \zeta_{\frac{m}{M},M} + \zeta_{\frac{m}{M},M} - \Xi^{(mM)}| > \varepsilon\right) \\
&\leq P\left(\|P_{k-1}\|_\infty h_{k-1}\frac{1}{M}\sum_{m=1}^{M} \left(|X^{(mM)} - \zeta_{\frac{m}{M},M}| + |\zeta_{\frac{m}{M},M} - \Xi^{(mM)}|\right) > \varepsilon\right) \\
&\leq P\left(\|P_{k-1}\|_\infty h_{k-1}\left(Z_M^1 + Z_M^2\right) > \varepsilon\right),
\end{aligned}$$

where $Z_M^1 = \frac{1}{M}\sum_{m=1}^{M}|X^{(mM)} - \zeta_{\frac{m}{M},M}|$ and $Z_M^2 = \frac{1}{M}\sum_{m=1}^{M}|\zeta_{\frac{m}{M},M} - \Xi^{(mM)}|$ are copies of the random variable Z_M appearing in Lemma 3.9, from which it follows that $Z_M^i \xrightarrow{P} 0, i = 1, 2$. It holds $\|P_{k-1}\|_\infty < \infty$ due to the assumed compactness of the support. Then, by Theorem 2.12, one has that $\|P_{k-1}\|_\infty h_{k-1}(Z_M^1 + Z_M^2) \xrightarrow{P} 0$, which completes the proof. ■

The proof of Proposition 3.8 is now straight-forward.

Proof. (Proposition 3.8) Rewrite equation (3.32) in the form

$$\frac{1}{M}\left(D(\boldsymbol{\Xi}^{(M)})^T D(\boldsymbol{\Xi}^{(M)})\right)\tilde{q}_M = \frac{1}{M} D(\boldsymbol{\Xi}^{(M)})^T \boldsymbol{X}^{(M)}, \tag{3.38}$$

and define the random matrix, respectively random vector

$$A_M = \frac{1}{M}\left(D(\boldsymbol{\Xi}^{(M)})^T D(\boldsymbol{\Xi}^{(M)})\right),\ z_M = \frac{1}{M} D(\boldsymbol{\Xi}^{(M)})^T \boldsymbol{X}^{(M)}.$$

Equation (3.38) can be thus written in the compact form

$$A_M \tilde{q}_M = z_M.$$

For the elements $(A_M)_{k,l}$ of the matrix A_M one has

$$(A_M)_{k,l} = \frac{1}{M}\sum_{m=1}^{M} P_{k-1}(\Xi^{(mM)})h_{k-1}P_{l-1}(\Xi^{(mM)})h_{l-1},\ k, l = 1, \ldots, N+1.$$

From Theorem 2.11 one has in the limit $M \to \infty$

$$(A_M)_{k,l} \overset{\text{wp1}}{\to} h_{k-1}h_{l-1}E[P_{k-1}(\Xi)P_{l-1}(\Xi)] = \delta_{kl}h_{k-1},\ k, l = 1, \ldots, N+1, \tag{3.39}$$

due to the orthogonality of the polynomials, see equation (2.30). For the elements of the vector z_M it holds

$$\begin{aligned}(z_M)_k &= \frac{1}{M}\sum_{m=1}^{M} P_{k-1}(\Xi^{(mM)})h_{k-1}X^{(mM)} \\ &= \frac{1}{M}\sum_{m=1}^{M} P_{k-1}(\Xi^{(mM)})h_{k-1}\left(X^{(mM)} - \Xi^{(mM)} + \Xi^{(mM)}\right) \\ &= \frac{1}{M}\sum_{m=1}^{M} h_{k-1}P_{k-1}(\Xi^{(mM)})\Xi^{(mM)} + \frac{1}{M}\sum_{m=1}^{M} P_{k-1}(\Xi^{(mM)})h_{k-1}\left(X^{(mM)} - \Xi^{(mM)}\right).\end{aligned} \tag{3.40}$$

The first term converges to $h_{k-1}E[P_{k-1}(\Xi)\Xi]$ with probability 1, as $M \to \infty$. From the 3-term recurrence relation in equation (2.31), it follows

$$P_{k-1}(\Xi)\Xi = \gamma_{k-1}P_k(\Xi) + \beta_{k-1}P_{k-1}(\Xi) + \alpha_{k-1}P_{k-2}(\Xi),\ k \geq 1.$$

Using the linearity of the expectation, it follows for all $k \geq 1$

$$\begin{aligned}E[P_{k-1}(\Xi)\Xi] &= \gamma_{k-1}E[P_k(\Xi)] + \beta_{k-1}E[P_{k-1}(\Xi)] + \alpha_{k-1}E[P_{k-2}(\Xi)] \\ &= \beta_{k-1}\delta_{k1}\frac{1}{h_0} + \alpha_{k-1}\delta_{k2}\frac{1}{h_0} \\ &= \beta_0\frac{1}{h_0} + \alpha_1\frac{1}{h_0},\end{aligned}$$

because $E[P_k(\Xi)] = 0$, for all $k \geq 1$, $E[P_{k-1}(\Xi)] = \delta_{k1}\frac{1}{h_0}$ and $E[P_{k-2}(\Xi)] = \delta_{k2}\frac{1}{h_0}$. The second term in equation (3.40) converges in probability to 0 as follows from Lemma 3.10. Thus, the Gauss estimator $\tilde{q}_M$ satisfies in the limit the following linear system

$$\begin{pmatrix} h_0 & 0 & \dots & 0 \\ 0 & h_1 & \dots & 0 \\ & & & \vdots \\ 0 & 0 & \dots & h_N \end{pmatrix}\begin{pmatrix}\tilde{q}_0 \\ \tilde{q}_1 \\ \vdots \\ \tilde{q}_N\end{pmatrix} = \begin{pmatrix} h_0\beta_0\frac{1}{h_0} \\ h_1\alpha_1\frac{1}{h_0} \\ \vdots \\ 0\end{pmatrix} \tag{3.41}$$

The matrix in (3.41) is invertible as the Haar weights are positive real numbers. It follows that $\lim_{M\to\infty}\tilde{q}_M = \tilde{q} = (\frac{\beta_0}{h_0}, \frac{\alpha_1}{h_0}, 0, \dots, 0)$, where the limit is to be understood as convergence in probability. All together for $M \to \infty$,

$$X = \frac{\beta_0}{h_0}P_0(\Xi)h_0 + \frac{\alpha_1}{h_0}P_1(\Xi)h_1.$$

In chapter 2, it was assumed $P_0(\xi) = 1, \forall\ \xi \in \mathbb{R}$. Under this assumption and by using again the 3-term recurrence relation, it follows that $P_1(\xi) = \frac{\xi-\beta_0}{\gamma_0}, \forall\ \xi \in \mathbb{R}$. By equation (2.32), it holds $\alpha_1 h_1 = \gamma_0 h_0$. Therefore, as the sample size grows to infinity

$$X = \beta_0 + \frac{\alpha_1}{h_0}\frac{\Xi - \beta_0}{\gamma_0}h_1 = \beta_0 + \Xi - \beta_0 = \Xi,$$

and the proof is completed. ■

Non-linear case

The general case where the samples $\{\Xi^m\}_{m=1,\dots,M}$ and $\{X^m\}_{m=1,\dots,M}$ come from two different distributions F_Ξ and F_X is now considered. Assume as in the linear case the following regression model

$$X^{(mM)} = \sum_{n=0}^{N} q_{n,M} P_n(\Xi^{(mM)}) h_n,\ m = 1, \dots, M. \tag{3.42}$$

Theorem 3.11. *Assume that the distribution function F_X is twice differentiable and that the density f_X is bounded away from zero on the support S_X. Furthermore, assume that S_X is a compact subset of $\mathbb{R}$. The solution $\tilde{q}_M$ to the least squares problem as in (3.30) converges in probability to the vector $q = (q_0, \dots, q_N)$ of the coefficients in the PC expansion of a copy of X, as the sample size M grows to infinity.*

Proof. Consider the Gauss estimator satisfying $A_M \tilde{q}_M = z_M$, as in the linear case. Again equation (3.39) holds. For the components of the vector z_M, one has

$$\begin{aligned}
(z_M)_k &= \frac{1}{M} \sum_{m=1}^{M} P_{k-1}(\Xi^{(mM)}) h_{k-1} X^{(mM)} \\
&= \frac{1}{M} \sum_{m=1}^{M} P_{k-1}(\Xi^{(mM)}) h_{k-1} \left(g_X(\Xi^{(mM)}) - g_X(\Xi^{(mM)}) + X^{(mM)} \right) \\
&= \frac{1}{M} \sum_{m=1}^{M} P_{k-1}(\Xi^{(mM)}) h_{k-1} g_X(\Xi^{(mM)}) + \frac{1}{M} \sum_{m=1}^{M} P_{k-1}(\Xi^{(mM)}) h_{k-1} \left(X^{(mM)} - g_X(\Xi^{(mM)}) \right).
\end{aligned}$$

For the first term it holds

$$\frac{1}{M} \sum_{m=1}^{M} P_{k-1}(\Xi^{(mM)}) h_{k-1} g_X(\Xi^{(mM)}) \overset{\text{wp1}}{\to} E[h_{k-1} P_{k-1}(\Xi) g_X(\Xi)] = h_{k-1} q_{k-1},\ M \to \infty.$$

Consider the second term

$$e_M = \frac{1}{M} \sum_{m=1}^{M} P_{k-1}(\Xi^{(mM)}) h_{k-1} \left(X^{(mM)} - g_X(\Xi^{(mM)}) \right). \tag{3.43}$$

As in the linear case, one has to show that $e_M \overset{P}{\to} 0$, with $M \to \infty$. Note that the match of the ordered values in the two samples is the discrete version of the isoprobabilistic transformation: ordering the values $\{\Xi^m\}_{m=1,\dots,M}$ and taking their image under $F_X^{-1} \circ F_\Xi$ results in an ordered sample in the X-space. This is because the transformation $G = F_X^{-1} \circ F_\Xi$ is non-decreasing. So, the coefficients one is actually computing with the proposed method are the coefficients of the random variable $X = F_X^{-1} \circ F_\Xi(\Xi)$. Assume therefore that $g_X \equiv G$. This means that the difference $X^{(mM)} - g_X(\Xi^{(mM)})$ in equation (3.43) can be assumed to be the difference of two ordered samples from the same distribution. It follows immediately from Lemma 3.10 that also in the nonlinear case

$$e_M = \frac{1}{M} \sum_{m=1}^{M} P_{k-1}(\Xi^{(mM)}) h_{k-1} \left(X^{(mM)} - g_X(\Xi^{(mM)}) \right) \overset{P}{\to} 0,\ M \to \infty.$$

By building again the limit linear system for the Gauss estimator, the conclusion of the proposition follows easily. ■

REMARK

(i) The coefficients q_n and $\tilde{q}_{n,M}$, $n \in \{0, \dots, N\}$ do not depend on N. The truncation error of the PC expansion of X plays here no role, as the orthogonal projection to the space $\text{lin}\{P_n \colon n = 0, \dots, N\}$ is equivalent with the minimization of the quadratic distance to this space.

(ii) As explained in the proof of Theorem 3.11, the transformation corresponding to the matching of the ordered values is the discrete version of the one-dimensional Rosenblatt transformation.

(iii) One could possibly estimate the coefficients q_n direct by their definition as

$$\hat{q}_{n,M} = \frac{1}{M} \sum_{m=1}^{M} X^{(mM)} P_n(\Xi^{(mM)}), \ n = 0, \dots, N. \tag{3.44}$$

In the next section, it will be shown how this approach can be mathematically justified.

3.4.2 Multi-dimensional case

Consider next two random vectors $\boldsymbol{\Xi} = (\Xi_1, \dots, \Xi_d)$ and $\boldsymbol{X} = (X_1, \dots, X_d)$ on a probability space $(\Omega, \mathcal{A}, P)$. Assume that the supports $S_{\boldsymbol{\Xi}}$ and $S_{\boldsymbol{X}}$ of their corresponding image measures $\boldsymbol{\mu} = (\boldsymbol{\Xi})_{\#}P$ and $\boldsymbol{\nu} = (\boldsymbol{X})_{\#}P$ are compact subsets of $\mathbb{R}^d$. Let $\{\boldsymbol{\Xi}^m = (\Xi_1^m, \dots, \Xi_d^m)\}_{m=1,\dots,M}$ and $\{\boldsymbol{X}^m = (X_1^m, \dots, X_d^m)\}_{m=1,\dots,M}$ be two independent random samples from their distributions $F_{\boldsymbol{\Xi}}$ and $F_{\boldsymbol{X}}$ respectively and assume that these distribution functions admit densities with respect to the Lebesgue measure. Denote by

$$\boldsymbol{\mu}_M = \frac{1}{M} \sum_{m=1}^{M} \delta_{\boldsymbol{\xi}^m} \text{ and } \boldsymbol{\nu}_M = \frac{1}{M} \sum_{m=1}^{M} \delta_{\boldsymbol{x}^m} \tag{3.45}$$

the discrete empirical (random) measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ corresponding to these random samples.

Let $T = (T_1, \dots, T_d) \colon S_{\boldsymbol{\Xi}} \to S_{\boldsymbol{X}}$ be the optimal transportation map between the measures $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ for the quadratic cost, so that $\boldsymbol{X} = T(\boldsymbol{\Xi})$. From Theorem 3.5 and the assumptions made here, this map exists and is unique and so is also the corresponding optimal transference plan $\boldsymbol{\pi}$. One wishes again to estimate the PC coefficients in the truncated expansions

$$(X_1, \dots, X_d) \simeq \left(\sum_{|\boldsymbol{n}|=0}^{N} q_{\boldsymbol{n}}^1 P_{\boldsymbol{n}}(\boldsymbol{\Xi}) h_{\boldsymbol{n}}, \dots, \sum_{|\boldsymbol{n}|=0}^{N} q_{\boldsymbol{n}}^d P_{\boldsymbol{n}}(\boldsymbol{\Xi}) h_{\boldsymbol{n}} \right), \ \boldsymbol{n} \in \mathbb{N}_0^d, \tag{3.46}$$

for a given approximation order $N \in \mathbb{N}_0$. Analogously to the one-dimensional case, a matching between the samples $\{\boldsymbol{\Xi}^m\}_{m=1,\dots,M}$ and $\{\boldsymbol{X}^m\}_{m=1,\dots,M}$ has to be first build. As

mentioned earlier, there is no total order in the Euclidean space $\mathbb{R}^d$ and the samples will be matched by solving the corresponding assignment problem and constructing the discrete analog of T. Therefore, denote by $\overline{\sigma}^M$ the optimal permutation which solves the assignment problem for the cost $s(\boldsymbol{x}, \boldsymbol{y}) = \langle \boldsymbol{x}, \boldsymbol{y} \rangle_E$ and the measures $\boldsymbol{\mu}_M$ and $\boldsymbol{\nu}_M$. This permutation defines the discrete optimal map $T^M = (T_1^M, \dots, T_d^M)$ between the measures $\boldsymbol{\mu}_M$ and $\boldsymbol{\nu}_M$ by the equation

$$T^M(\boldsymbol{\Xi}^m) = \boldsymbol{X}^{\overline{\sigma}^M(m)},\ m = 1, \dots, M, \tag{3.47}$$

and corresponds to an optimal transference plan $\boldsymbol{\pi}_M$. As explained in sections 3.3.2 and 3.3.3 this permutation can be computed via the auction algorithm.

Regression approach

The goal is to estimate the coefficients in (3.46) from the given samples by minimizing the residual

$$\sum_{m=1}^{M} |X_1^{\overline{\sigma}^M(m)} - \sum_{|\boldsymbol{n}|=0}^{N} q_{\boldsymbol{n},M}^1 P_{\boldsymbol{n}}(\boldsymbol{\Xi}^m) h_{\boldsymbol{n}}|^2 + \dots + \sum_{m=1}^{M} |X_d^{\overline{\sigma}^M(m)} - \sum_{|\boldsymbol{n}|=0}^{N} q_{\boldsymbol{n},M}^d P_{\boldsymbol{n}}(\boldsymbol{\Xi}^m) h_{\boldsymbol{n}}|^2.$$

Analogously as in the one-dimensional case, the design matrix of this least-squares problem reads

$$D(\boldsymbol{\Xi}) = \begin{pmatrix} P_0(\boldsymbol{\Xi}^1)h_0 & P_1(\boldsymbol{\Xi}^1)h_1 & \dots & P_{d_P}(\boldsymbol{\Xi}^1)h_{d_P} \\ P_0(\boldsymbol{\Xi}^2)h_0 & P_1(\boldsymbol{\Xi}^2)h_1 & \dots & P_{d_P}(\boldsymbol{\Xi}^2)h_{d_P} \\ & & & \vdots \\ P_0(\boldsymbol{\Xi}^M)h_0 & P_1(\boldsymbol{\Xi}^M)h_1 & \dots & P_{d_P}(\boldsymbol{\Xi}^M)h_{d_P} \end{pmatrix}$$

Here the single index notation is used for simplicity and $d_P + 1$ denotes the dimension of the polynomial subspace $\mathcal{P}^N$ defined in equation (2.42). One has

$$\tilde{q}_{\boldsymbol{n},M}^j = \left(D(\boldsymbol{\Xi})^T D(\boldsymbol{\Xi})\right)^{-1} D(\boldsymbol{\Xi})^T \boldsymbol{X}_j^{(M)},\ j = 1, \dots, d, \tag{3.48}$$

where $\boldsymbol{X}_j^{(M)} = (X_j^{\overline{\sigma}^M(1)}, \dots, X_j^{\overline{\sigma}^M(M)})^T$, $j = 1, \dots, d$.

Theorem 3.12. *Assume that $\boldsymbol{\Xi}$ and $\boldsymbol{X}$ are two continuous random vectors with compact supports. Let $\tilde{q}_{\boldsymbol{n},M}^j$, $j = 1, \dots, d$ be the solutions of the least-squares problems as in (3.48). Then,*

$$\tilde{q}_{\boldsymbol{n},M}^j \overset{wp1}{\to} q_{\boldsymbol{n}}^j, \text{ for } M \to \infty,\ j = 1, \dots, d, \tag{3.49}$$

where $\{q_{\boldsymbol{n}}^j : |\boldsymbol{n}| \leq N\}_{j=1,\dots,d}$ are the PC coefficients of the random vector $\boldsymbol{X} = T(\boldsymbol{\Xi})$ in equation (3.46).

Proof. The proof is similar to the proof of Theorem 3.8. Define as before

$$A_M = \frac{1}{M} D(\boldsymbol{\Xi})^T D(\boldsymbol{\Xi}),\ z_M^j = \frac{1}{M} D(\boldsymbol{\Xi})^T \boldsymbol{X}_j^{(M)},\ j = 1, \dots, d.$$

The elements $(A_M)_{k,l}, k, l = 1 \ldots, K+1$ of the matrix A_M converge with probability 1 to $\delta_{k,l} h_{k-1}$ as before. For the elements of z_M^j, $j = 1, \ldots, d$ we have

$$\begin{aligned}(z_M^j)_k &= \frac{1}{M} \sum_{m=1}^{M} P_{k-1}(\boldsymbol{\Xi}^m) h_{k-1} X_j^{\overline{\sigma}^M(m)} \\ &= \frac{1}{M} \sum_{m=1}^{M} P_{k-1}(\boldsymbol{\Xi}^m) h_{k-1} \left(X_j^{\overline{\sigma}^M(m)} - T_j(\boldsymbol{\Xi}^m) + T_j(\boldsymbol{\Xi}^m) \right) \\ &= \frac{1}{M} \sum_{m=1}^{M} P_{k-1}(\boldsymbol{\Xi}^m) h_{k-1} T_j(\boldsymbol{\Xi}^m) + \frac{1}{M} \sum_{m=1}^{M} P_{k-1}(\boldsymbol{\Xi}^m) h_{k-1} \left(X_j^{\overline{\sigma}^M(m)} - T_j(\boldsymbol{\Xi}^m) \right).\end{aligned}$$

The first term converges with probability 1 to $E[P_{k-1} h_{k-1} T_j(\boldsymbol{\Xi})] = h_{k-1} q_{k-1}^j$ for $M \to \infty$. It remains to examine the terms

$$e_M^j = \frac{1}{M} \sum_{m=1}^{M} P_{k-1}(\boldsymbol{\Xi}^m) h_{k-1} \left(X_j^{\overline{\sigma}^M(m)} - T_j(\boldsymbol{\Xi}^m) \right), \; j = 1, \ldots, d.$$

One has

$$\begin{aligned}|e_M^j| &= \left| \frac{1}{M} \sum_{m=1}^{M} P_{k-1}(\boldsymbol{\Xi}^m) h_{k-1} \left(X_j^{\overline{\sigma}^M(m)} - T_j(\boldsymbol{\Xi}^m) \right) \right| \\ &= \left| \frac{1}{M} \sum_{m=1}^{M} P_{k-1}(\boldsymbol{\Xi}^m) h_{k-1} \left(T_j^M(\boldsymbol{\Xi}^m) - T_j(\boldsymbol{\Xi}^m) \right) \right| \\ &\leq \|P_{k-1}\|_\infty h_{k-1} \frac{1}{M} \sum_{m=1}^{M} |T_j(\boldsymbol{\Xi}^m) - T_j^M(\boldsymbol{\Xi}^m)|.\end{aligned}$$

Again $\|P_{k-1}\|_\infty < \infty$, due to compactness of $S_{\boldsymbol{\Xi}}$. Define $C = \|P_{k-1}\|_\infty h_{k-1}$. Then,

$$\begin{aligned}|e_M^j| &\leq C \frac{1}{M} \sum_{m=1}^{M} |T_j(\boldsymbol{\Xi}^m) - T_j^M(\boldsymbol{\Xi}^m)| \\ &= C \int_{S_{\boldsymbol{\Xi}}} |T_j(\boldsymbol{\xi}) - T_j^M(\boldsymbol{\xi})| d\boldsymbol{\mu}_M(\boldsymbol{\xi}) \\ &\leq C \int_{S_{\boldsymbol{\Xi}}} \|T(\boldsymbol{\xi}) - T^M(\boldsymbol{\xi})\|_E d\boldsymbol{\mu}_M(\boldsymbol{\xi}).\end{aligned}$$

Choose an $\epsilon > 0$ and rewrite the above integral

$$\begin{aligned}
|e_M^j| &\leq C \left[\int_{S_{\boldsymbol{\Xi}}} \mathbb{1}\{\|T(\boldsymbol{\xi}) - T^M(\boldsymbol{\xi})\|_E > \epsilon\} \|T(\boldsymbol{\xi}) - T^M(\boldsymbol{\xi})\|_E d\boldsymbol{\mu}_M(\boldsymbol{\xi}) \right. \\
&\qquad \left. + \int_{S_{\boldsymbol{\Xi}}} \mathbb{1}\{\|T(\boldsymbol{\xi}) - T^M(\boldsymbol{\xi})\|_E \leq \epsilon\} \|T(\boldsymbol{\xi}) - T^M(\boldsymbol{\xi})\|_E d\boldsymbol{\mu}_M(\boldsymbol{\xi}) \right] \\
&\leq C \left[\int_{S_{\boldsymbol{\Xi}}} \mathbb{1}\{\|T(\boldsymbol{\xi}) - T^M(\boldsymbol{\xi})\|_E > \epsilon\} \|T(\boldsymbol{\xi}) - T^M(\boldsymbol{\xi})\|_E d\boldsymbol{\mu}_M(\boldsymbol{\xi}) + \epsilon \boldsymbol{\mu}_M(S_{\boldsymbol{\Xi}}) \right] \\
&\leq C \left[\text{diam}(S_{\boldsymbol{X}}) \int_{S_{\boldsymbol{\Xi}}} \mathbb{1}\{\|T(\boldsymbol{\xi}) - T^M(\boldsymbol{\xi})\|_E > \epsilon\} d\boldsymbol{\mu}_M(\boldsymbol{\xi}) + \epsilon \boldsymbol{\mu}_M(S_{\boldsymbol{\Xi}}) \right] \\
&= C \left[\text{diam}(S_{\boldsymbol{X}}) \boldsymbol{\mu}_M(\|T(\boldsymbol{\xi}) - T^M(\boldsymbol{\xi})\|_E > \epsilon) + \epsilon \right]
\end{aligned}$$

All together one has that

$$|e_M^j| \leq C \text{diam}(S_{\boldsymbol{X}}) \boldsymbol{\mu}_M \left(\|T(\boldsymbol{\xi}) - T^M(\boldsymbol{\xi})\|_E > \epsilon \right) + C\epsilon, \ j = 1, \ldots, d.$$

The empirical measure $\boldsymbol{\mu}_M$ converges strongly to $\boldsymbol{\mu}$ by the law of large numbers and with probability 1, as $M \to \infty$. Thus, the weak convergence $\boldsymbol{\mu}_M \rightharpoonup \boldsymbol{\mu}$ holds also with probability 1. By using the uniqueness of the optimal map and the optimal transference plan for the quadratic cost, Corollary 3.7 then guarantees that with probability 1 it holds

$$\forall\, \epsilon > 0, \ \boldsymbol{\mu}_M \left(\|T(\boldsymbol{\xi}) - T^M(\boldsymbol{\xi})\|_E > \epsilon \right) \to 0, \ M \to \infty.$$

This means that $|e_M^j| \overset{\text{wp1}}{\to} 0$, when the sample size M grows to infinity. Returning again to the Gauss estimators $\tilde{q}^j_{\boldsymbol{n},M}$, $j = 1, \ldots, d$ in (3.48), it follows that in the limit $M \to \infty$ they satisfy P-a.s the equation

$$\begin{pmatrix} h_0 & 0 & \ldots & 0 \\ 0 & h_1 & \ldots & 0 \\ & & & \vdots \\ 0 & 0 & \ldots & h_{d_P} \end{pmatrix} \begin{pmatrix} \tilde{q}_0^j \\ \tilde{q}_1^j \\ \vdots \\ \tilde{q}_{d_P}^j \end{pmatrix} = \begin{pmatrix} h_0 q_0^j \\ h_1 q_1^j \\ \vdots \\ h_{d_P} q_{d_P}^j \end{pmatrix}, \tag{3.50}$$

where the single index notation is again used. The matrix in (3.50) is the same invertible matrix which appeared in the limit of the one-dimensional problem and is invertible. Now, the conclusion of the proposition follows immediately. ■

Non-intrusive approach

Another way to compute the PC coefficients using the theory of optimal transportation is directly through their definition in equation (2.39), where instead of $f_{\boldsymbol{X}}$ one can substitute the optimal map T from the previous subsection. In this case

$$q_{\boldsymbol{n}}^j = \int_\Omega T_j(\boldsymbol{\Xi}(\omega)) P_{\boldsymbol{n}}(\boldsymbol{\Xi}(\omega)) dP(\omega) = \int_{S_{\boldsymbol{\Xi}}} T_j(\boldsymbol{\xi}) P_{\boldsymbol{n}}(\boldsymbol{\xi}) d\boldsymbol{\mu}(\boldsymbol{\xi}), \ j = 1, \ldots, d. \tag{3.51}$$

Assume again that finite samples of size M from $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are available. Then, one can approximate $q_{\boldsymbol{n}}^j$ by

$$\begin{aligned}\hat{q}_{\boldsymbol{n},M}^j &= \int_{S_{\boldsymbol{\Xi}}} T_j^M(\boldsymbol{\xi}) P_{\boldsymbol{n}}(\boldsymbol{\xi}) d\boldsymbol{\mu}_M(\boldsymbol{\xi}) \\ &= \frac{1}{M}\sum_{m=1}^{M} T_j^M(\boldsymbol{\xi}^m) P_{\boldsymbol{n}}(\boldsymbol{\xi}^m) = \frac{1}{M}\sum_{m=1}^{M} x_j^{\bar{\sigma}^M(m)} P_{\boldsymbol{n}}(\boldsymbol{\xi}^m).\end{aligned} \tag{3.52}$$

Proposition 3.13. *Under the assumptions of Theorem 3.12, it holds that there exists a subsequence $(\hat{q}_{\boldsymbol{n},M_L})_{L\in\mathbb{N}}$ of $(\hat{q}_{\boldsymbol{n},M})_{M\in\mathbb{N}}$ such that*

$$\hat{q}_{\boldsymbol{n},M_L}^j \overset{wp1}{\to} q_{\boldsymbol{n}}^j,\ L\to\infty, \text{ for } j=1,\dots,d, \tag{3.53}$$

where $\hat{q}_{\boldsymbol{n},M}$ are the estimated PC coefficients defined by equation (3.52).

Proof. Fix $\varepsilon > 0$ and $\boldsymbol{n} \in \mathbb{N}_0^d$.

Then,

$$\begin{aligned}|\hat{q}_{\boldsymbol{n},M}^j - q_{\boldsymbol{n}}^j| &= \left| \int_{S_{\boldsymbol{\Xi}}} T_j^M(\boldsymbol{\xi}) P_{\boldsymbol{n}}(\boldsymbol{\xi}) d\boldsymbol{\mu}_M(\boldsymbol{\xi}) - \int_{S_{\boldsymbol{\Xi}}} T_j(\boldsymbol{\xi}) P_{\boldsymbol{n}}(\boldsymbol{\xi}) d\boldsymbol{\mu}(\boldsymbol{\xi}) \right| \\ &= \left| \int_{S_{\boldsymbol{\Xi}}\times S_{\boldsymbol{X}}} x_j P_{\boldsymbol{n}}(\boldsymbol{\xi}) d\boldsymbol{\pi}_M(\boldsymbol{\xi},\boldsymbol{x}) - \int_{S_{\boldsymbol{\Xi}}\times S_{\boldsymbol{X}}} x_j P_{\boldsymbol{n}}(\boldsymbol{\xi}) d\boldsymbol{\pi}(\boldsymbol{\xi},\boldsymbol{x}) \right|.\end{aligned}$$

Theorem 3.6 assures that as $M\to\infty$, there exists a subsequence $(\boldsymbol{\pi}_{M_L})_{L\in\mathbb{N}}$ of $(\boldsymbol{\pi}_M)_{M\in\mathbb{N}}$ which converges weakly to the measure $\boldsymbol{\pi}$ with probability 1. The real functions $r_j(\boldsymbol{x},\boldsymbol{\xi}) = x_j P_{\boldsymbol{n}}(\boldsymbol{\xi}),\ j=1,\dots,d$ belong to the class of continuous bounded functions on $S_{\boldsymbol{\Xi}}\times S_{\boldsymbol{X}}$ because of the assumed compactness of the supports. Therefore, by Definition 2.13, there is a $L'\in\mathbb{N}$, so that $\forall\ L\geq L'$,

$$\left| \int_{S_{\boldsymbol{\Xi}}\times S_{\boldsymbol{X}}} x_j P_{\boldsymbol{n}}(\boldsymbol{\xi}) d\boldsymbol{\pi}_{M_L}(\boldsymbol{\xi},\boldsymbol{x}) - \int_{S_{\boldsymbol{\Xi}}\times S_{\boldsymbol{X}}} x_j P_{\boldsymbol{n}}(\boldsymbol{\xi}) d\boldsymbol{\pi}(\boldsymbol{\xi},\boldsymbol{x}) \right| < \varepsilon,$$

thus also

$$|\hat{q}_{\boldsymbol{n},M_L}^j - q_{\boldsymbol{n}}^j| < \varepsilon,\ \forall\, j=1,\dots,d.$$

■

REMARK

(i) In the one-dimensional case, the solution to the discrete assignment problem is the matching of the ordered statistics. Thus, OT theory allows to prove formula (3.44) in the one-dimensional case. This was not possible with the theory of ordered statistics as in general $F_X^{-1}(F_\Xi(\Xi^{(mM)})) \neq X^{(mM)}$.

(ii) When working with the definition of the polynomial chaos coefficients, it is here proved that there is only a subsequence which converges to the exact coefficients. This property expresses itself in rather bad approximations, compared with the results obtained by the method that uses the linear regression, as it will be shown in the next section via numerical experiments.

(iii) The approach using the optimal transportation theory allows to prove convergence with probability 1. Thus, a stronger result is obtained compared to the order statistics approach which resulted only in convergence in probability.

3.5 Numerical simulations

The theoretical results obtained in the previous sections are verified here numerically.

3.5.1 One-dimensional case

Let Ξ be a random variable distributed uniformly on $[-1, 1]$. Then the sequence of the corresponding orthogonal polynomials are the Legendre polynomials. Let X be a random variable distributed uniformly on the interval $[a, b]$. Then, the underlying transformation can be assumed to be

$$X = g_X(U) = \frac{a+b}{2} + \frac{b-a}{2}U, \tag{3.54}$$

so that the coefficients in the PC expansion of X

$$X = \sum_{n=0}^{\infty} q_n P_n(\Xi) h_n \tag{3.55}$$

in terms of the Legendre polynomials are

$$q_0 = \frac{a+b}{2},\ q_1 = \frac{b-a}{2h_1},\ \text{and } q_n = 0,\ \forall\, n \geq 2. \tag{3.56}$$

The PC coefficients are estimated from given samples from the distribution of X of different sizes M by using the regression and the non-intrusive approach. For the simulations it was assumed that $a = 0$ and $b = 20$. In Figure 3.1, density estimators from samples generated by the estimated PC expansions are shown. In Table 3.1, the estimated coefficients up to order $N = 4$ are summarized for different sample sizes and for the two different methods.

3.5.2 Multi-dimensional case

Let now $\boldsymbol{X} = (X_1, \ldots, X_d)$ be a random vector distributed according to a Dirichlet distribution parametrized by a vector $\boldsymbol{r}$. The Dirichlet distribution is the conjugate prior of the categorical and the multinomial distribution in Bayesian statistics [57]. Its support is the standard $(d-1)$ - dimensional simplex, this means the compact set

$$S_{\boldsymbol{X}} = \{\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d \colon \sum_{i=1}^{d} x_i = 1 \text{ and } x_i \in [0, 1], \forall\, i = 1, \ldots, d\}. \tag{3.57}$$

Let $\boldsymbol{\Xi} = (\Xi_1, \ldots, \Xi_d)$ be a vector of independent random variables each distributed uniformly on $[-1, 1]$. The multivariate Legendre polynomials are the corresponding orthogonal polynomials. Assume one has a sample of size M from the Dirichlet distribution and wishes to estimate the PC expansion of the associated random variable $\boldsymbol{X}$ with respect to

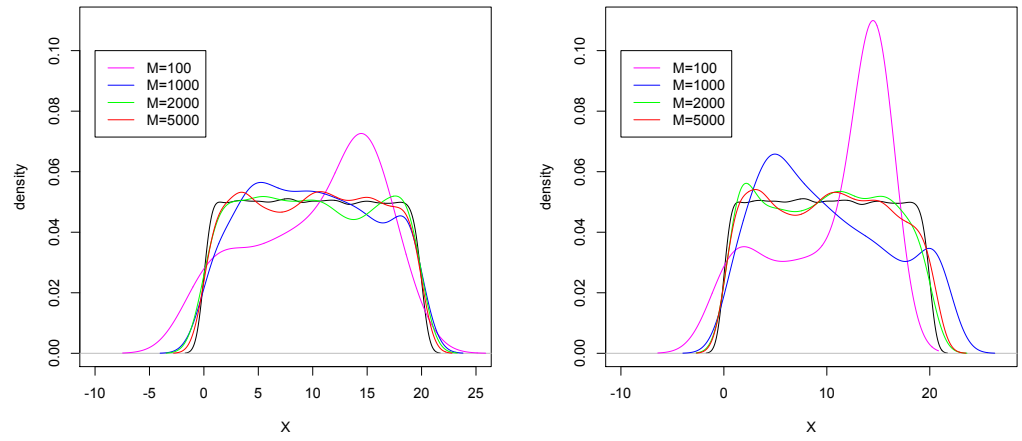

Figure 3.1: Density estimates from samples generated by the estimated PC expansion of $X \sim U[0, 20]$: regression (left) and non-intrusive (right) approach. M denotes the size of the given sample. The black line corresponds to the density estimate from a sample of size 60000 from the uniform distribution.

Ξ. The performance of the two different approaches is next numerically demonstrated. The parameter $\boldsymbol{r}$ of the Dirichlet distribution was fixed to $\boldsymbol{r} = (0.5, 1, 1.5, 2, 2.5)$. The order of the approximation was fixed in all cases to $N = 5$.

In Figures 3.2 and 3.3, the regression approach was followed when it was assumed that samples of two different sizes $M = 1000$ and $M = 5000$ respectively from the distribution of $\boldsymbol{X}$ are given. The parameter ε of the auction algorithm was set in both cases to $\varepsilon = 0.001$. The implementation of the auction algorithm in MATLAB®was used for the simulations. It can be seen that already with the smaller sample size the dependencies and the marginal densities of the 5-dimensional random vector $\boldsymbol{X}$ are captured well. When the sample sizes increases, the estimated samples tend to lie better on the simplex.

In Figure 3.4, the PC coefficients were estimated directly via the definition and by equation (3.52) when a set of observations of size $M = 5000$ from the distribution of $\boldsymbol{X}$ was given. This method fails to capture the underlying distribution and verifies the theoretical obtained results of poor convergence.

Table 3.1: Estimated PC coefficients for $X \sim U[0, 20]$

	q_0	q_1	q_2	q_3	q_4
true	10	10/3	0	0	0
			regression		
$M = 100$	9.96492051	3.18918970	-0.20330119	0.14687919	0.01739027
$M = 1000$	9.90444163	3.22706395	0.07341866	0.06003626	-0.04232979
$M = 2000$	9.936422608	3.377922081	0.017718355	-0.027664351	-0.014423813
$M = 5000$	10.078532222	3.346777517	-0.023336428	-0.009852080	0.004974737
			non-intrusive		
$M = 100$	10.17417866	2.77339313	-0.71158370	-0.05911298	0.05289290
$M = 1000$	9.88902374	3.45150115	0.36403198	0.09614726	-0.06000841
$M = 2000$	9.892016654	3.308645462	0.006242148	-0.020342941	0.059028209
$M = 5000$	10.12343	3.403760	0.01609069	0.00004540619	0.02416661

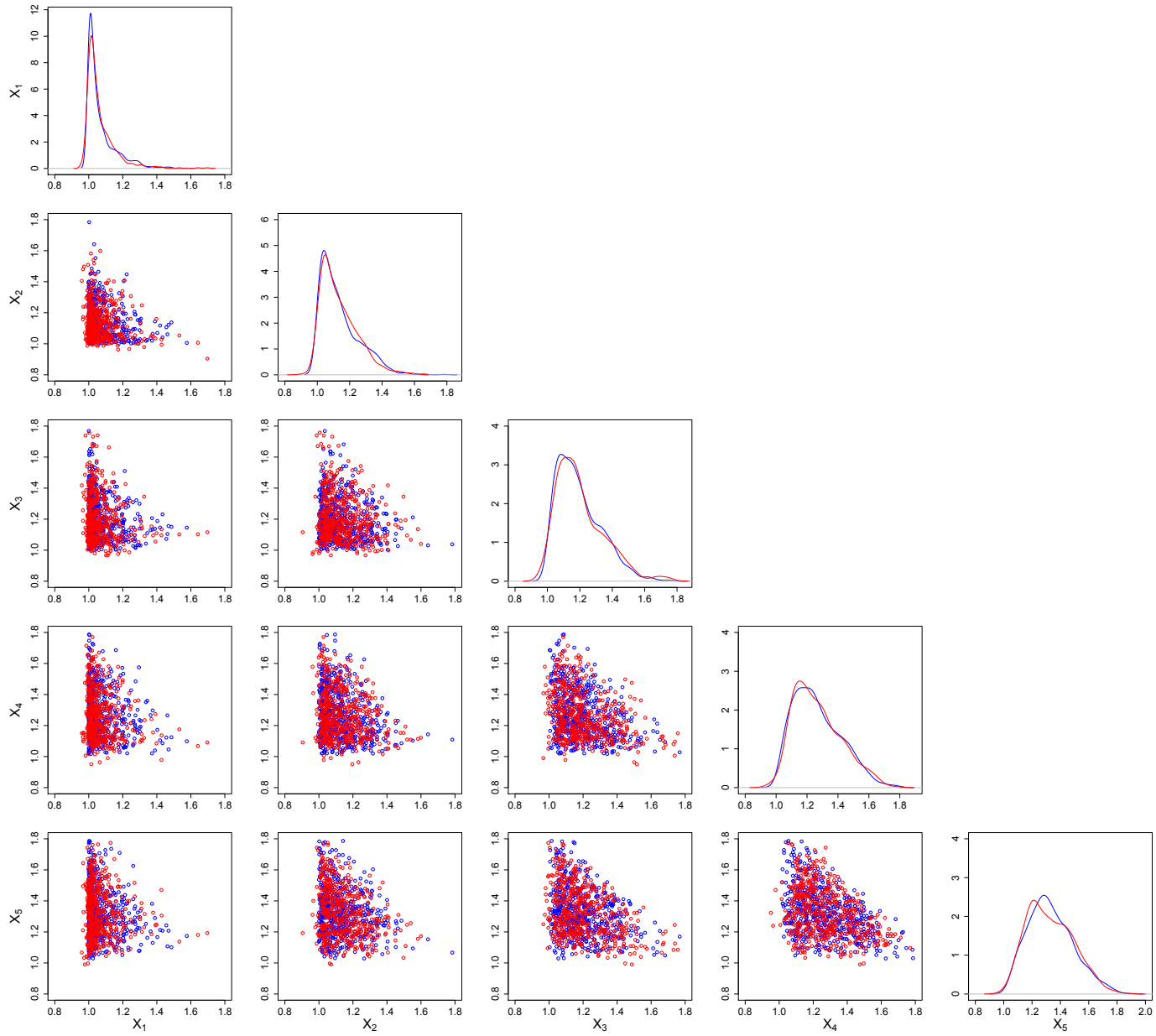

Figure 3.2: Univariate marginal estimates and bivariate sample plots. Blue curves and points are generated from the Dirichlet $D(0.5, 1, 1.5, 2, 2.5)$ and red curves and points from its estimated PC expansion given a sample of size $M = 1000$ and for $\varepsilon = 0.001$ via regression.

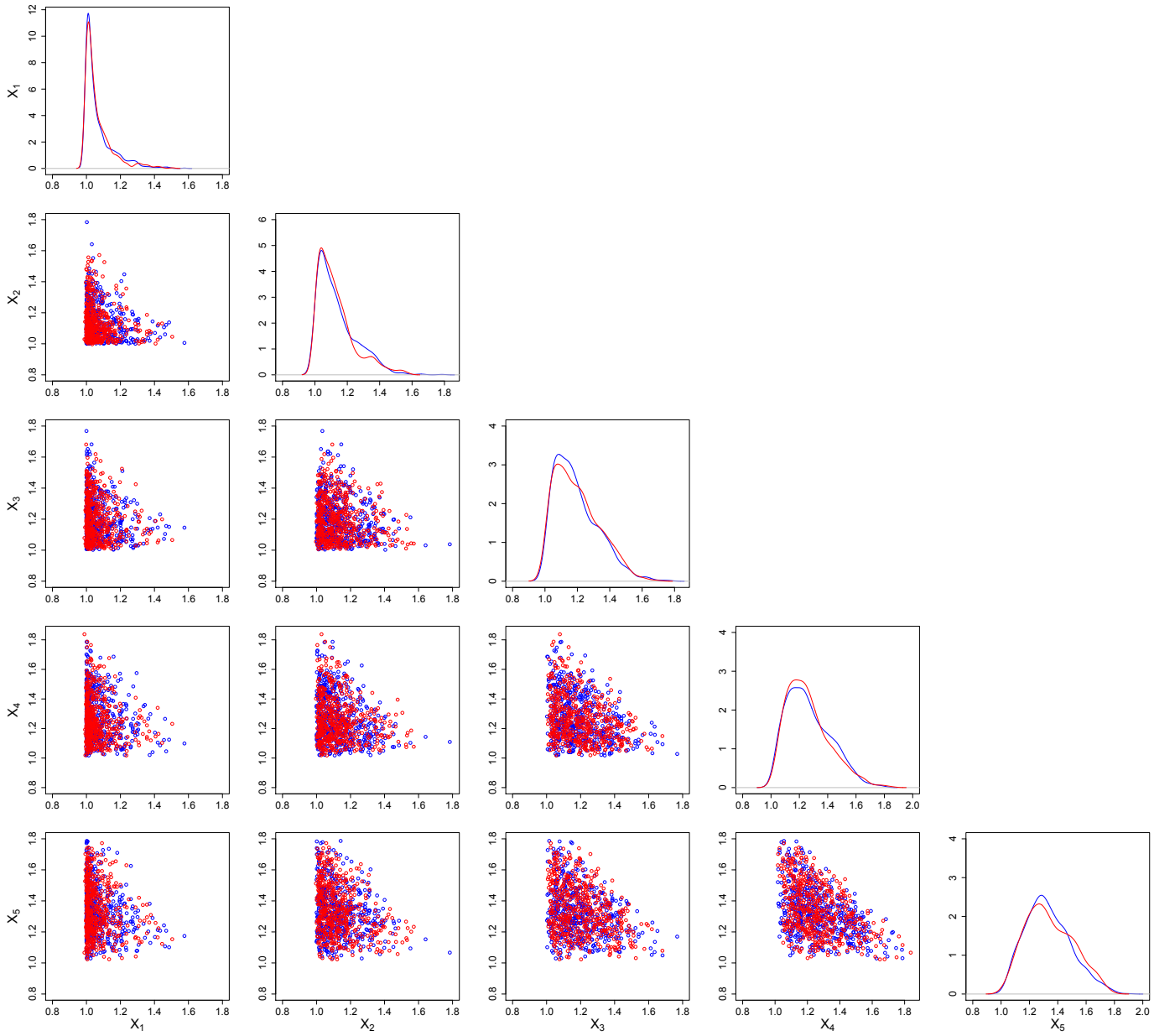

Figure 3.3: Univariate marginal estimates and bivariate sample plots. Blue curves and points are generated from the Dirichlet $D(0.5, 1, 1.5, 2, 2.5)$ and red curves and points from the estimated PC expansion given a sample of size $M = 5000$ and for $\varepsilon = 0.001$ via regression.

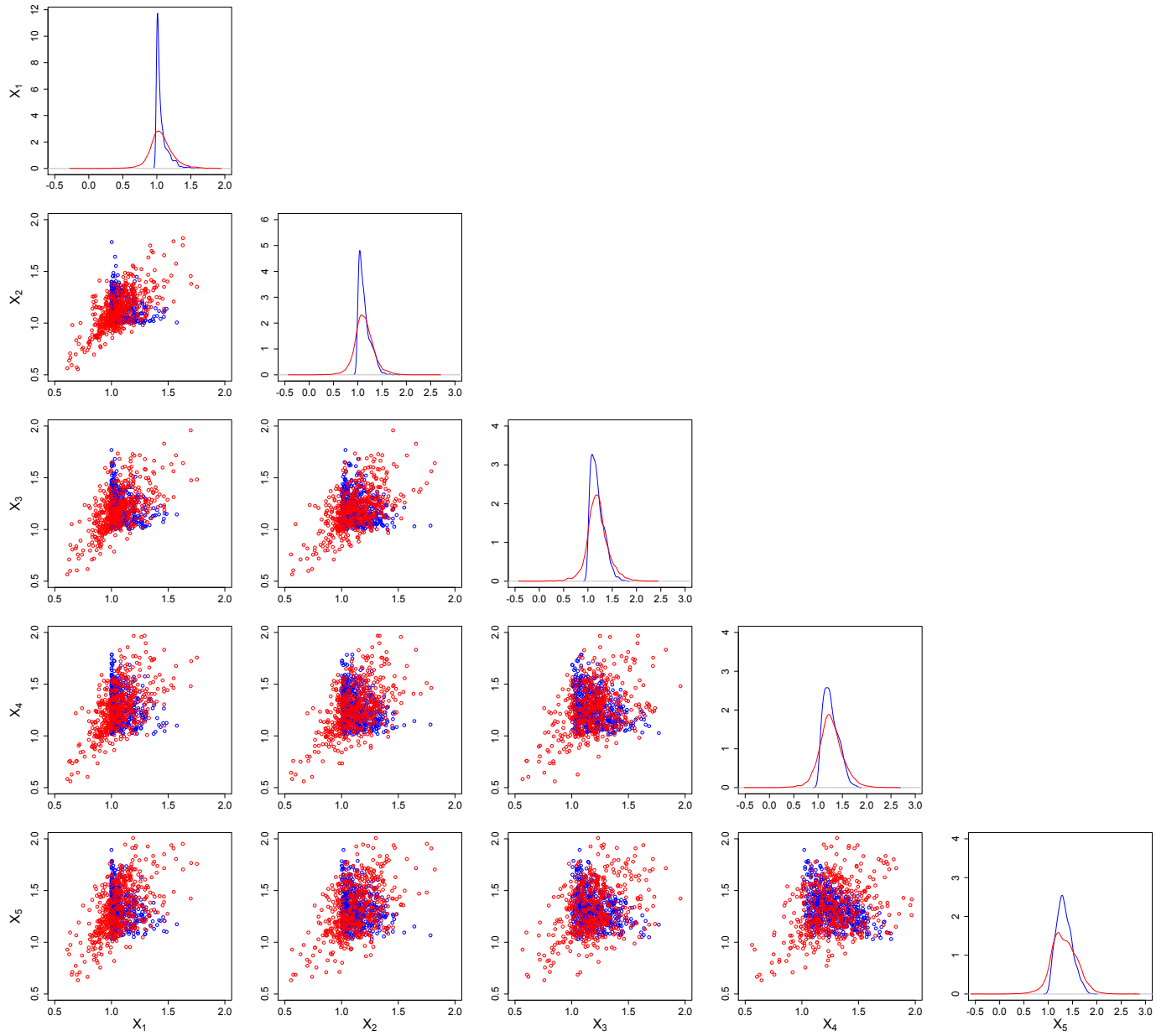

Figure 3.4: Univariate marginal estimates and bivariate sample plots. Blue curves and points are generated from the Dirichlet $D(0.5, 1, 1.5, 2, 2.5)$ and red curves and points from the estimated PC expansion given a sample of size $M = 5000$ and for $\varepsilon = 0.001$ via the non-intrusive approach.

4 Weighted polynomial chaos expansions

As discussed in chapter 2, truncated polynomial chaos expansions are being used for the approximation of the solution of dynamical systems with parametric uncertainty, and their coefficients are determined by stochastic Galerkin or non-intrusive methods. In many applications, the solution $x(t, \mathbf{\Theta})$ of (2.48) describes a quantity such as a chemical concentration or a population density. Thus, positivity (or better non-negativity) is a natural property to require for the solution and the initial condition of the dynamical system.

Even when the given dynamical system *preserves positivity*, i.e. solutions starting from non-negative initial data remain non-negative in their existence interval, it cannot be assured that any finite polynomial approximation $x_N(t, \mathbf{\Xi})$ as in (2.54) remains positive for all realizations of $\mathbf{\Xi}$ and all times t.

The problem of positivity in truncated expansions was addressed in [106], where it was analyzed that positivity is related to stability problems. It was there shown that increasing the approximation order may not solve this problem, as one would expect. In [41, 106] it was further stressed out that representing random variables which are positive but have a small mean and a large variance can lead to positive probabilities of negative values and therefore to instability. One way to overcome this problem is by using local expansions in a multi-wavelet basis [85]. Another possible solution is to transform the random variable first by using a strictly positive function such as the exponential, i.e. to find first the PC expansion of the random variable $Y = e^X$ and then infer the PC coefficients of X from the coefficients of the random variable Y. This step will introduce a high degree of nonlinearity along with the issues that its existence implies [41]. Furthermore, in this way one is not able to represent quantities that can be also zero, which is usually the case in practice.

In this chapter it is shown how the positivity of the solution can be preserved by applying proper summability methods. This is equivalent to introducing weights in the truncated polynomial approximation. The same methods allow to gain expansions also in the case $X \in L^p(\Omega, \sigma(\mathbf{\Xi}), P)$, $p \neq 2$.

4.1 Basics from functional analysis

In this section, results from functional analysis are summarized which will be used as tools in the sequence. More details can be found for example in [68] or other textbooks on functional analysis.

Definition 4.1. *A linear operator* $F\colon L^p(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \boldsymbol{\mu}) \to L^p(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \boldsymbol{\mu})$ *is called bounded if there exists a constant* $M \geq 0$*, such that*

$$\|F(f)\|_p \leq M\|f\|_p, \ \forall\, f \in L^p(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \boldsymbol{\mu}). \tag{4.1}$$

One can show that F is a bounded operator if and only if it is continuous. The *norm* $\|F\|^{L^p}$ of the bounded linear operator F is defined as

$$\|F\|^{L^p} = \sup_{f \in S_{L^p}} \|F(f)\|_p,\ 1 \leq p \leq \infty, \tag{4.2}$$

where

$$S_{L^p} = \{f \in L^p(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \boldsymbol{\mu})\colon\ \|f\|_p = 1\} \tag{4.3}$$

denotes the unit sphere in $L^p(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \boldsymbol{\mu})$. One can show that

$$\|F\|^{L^p} = \inf\{M \geq 0\colon \|F(f)\|_p \leq M\|f\|_p,\ \forall\, f \in L^p(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \boldsymbol{\mu})\}. \tag{4.4}$$

Definition 4.2. *An operator $F\colon L^p(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \boldsymbol{\mu}) \to L^p(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \boldsymbol{\mu})$ is called positive if and only if for all $f \in L^p(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \boldsymbol{\mu})$ such that $f(\boldsymbol{x}) \geq 0$ $\boldsymbol{\mu}$-a.e. it holds that also $F(f)(\boldsymbol{x}) \geq 0$ $\boldsymbol{\mu}$-a.e.*

Known theorems are next recalled.

Theorem 4.3. *(Hölder's inequality) Let $1 \leq p \leq q \leq \infty$ with $\frac{1}{p} + \frac{1}{q} = 1$. If $f \in L^p(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \boldsymbol{\mu})$ and $g \in L^q(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \boldsymbol{\mu})$, then*

$$\int_{\mathbb{R}^d} |f(\boldsymbol{x})g(\boldsymbol{x})| d\boldsymbol{\mu}(\boldsymbol{x}) \leq \|f\|_p \|g\|_q. \tag{4.5}$$

Lemma 4.4. *(Urysohn) Let $A \subseteq \mathbb{R}^d$ be a compact set and $U \subseteq \mathbb{R}^d$ an open set such that $A \subset U$. Then, there exists a continuous function $g\colon \mathbb{R}^d \to [0,1]$ such that $g(\boldsymbol{x}) = 1$, for all $\boldsymbol{x} \in A$ and $g(\boldsymbol{x}) = 0$, for all $\boldsymbol{x} \in \mathbb{R}^d \setminus U$.*

Theorem 4.5. *(Fubini) Let $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \boldsymbol{\mu})$ and $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \boldsymbol{\nu})$ be two probability spaces and let $\mathcal{W} = (\mathbb{R}^d \times \mathbb{R}^d, \mathcal{B}(\mathbb{R}^d) \times \mathcal{B}(\mathbb{R}^d), \boldsymbol{\mu} \times \boldsymbol{\nu})$ be their product space. Let f be a measurable function on $\mathcal{W}$ and assume that one of the following integrals is finite*

$$\begin{aligned} &\int_{\mathbb{R}^d \times \mathbb{R}^d} |f(\boldsymbol{x}, \boldsymbol{y})| d\boldsymbol{\mu}(\boldsymbol{x}) \times \boldsymbol{\nu}(\boldsymbol{y}), \\ &\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |f(\boldsymbol{x}, \boldsymbol{y})| d\boldsymbol{\mu}(\boldsymbol{x}) d\boldsymbol{\nu}(\boldsymbol{y}), \\ &\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |f(\boldsymbol{x}, \boldsymbol{y})| d\boldsymbol{\nu}(\boldsymbol{y}) d\boldsymbol{\mu}(\boldsymbol{x}). \end{aligned} \tag{4.6}$$

Then, it holds

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} f(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{\mu}(\boldsymbol{x}) \times \boldsymbol{\nu}(\boldsymbol{y}) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} f(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{\mu}(\boldsymbol{x}) d\boldsymbol{\nu}(\boldsymbol{y}) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} f(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{\nu}(\boldsymbol{y}) d\boldsymbol{\mu}(\boldsymbol{x}). \tag{4.7}$$

Definition 4.6. *Let $\boldsymbol{\mu}$ be a measure on the measurable space $(\Omega, \mathcal{A})$. The total variation $|\boldsymbol{\mu}|\colon \mathcal{A} \to \mathbb{R}$ is defined as*

$$|\boldsymbol{\mu}|(A) = \sup_{A \in \mathcal{A}} \{\sum_{n=1}^{N} |\boldsymbol{\mu}(A_n)|\colon \{A_1, \ldots, A_N\} \text{ is a measurable partition of } A\}. \tag{4.8}$$

The measure norm $\|\boldsymbol{\mu}\|$ is defined as $|\boldsymbol{\mu}|(\Omega)$.

Theorem 4.7. *(Riesz's representation theorem) Let $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \boldsymbol{\mu})$ be a probability space and let $\boldsymbol{\sigma}$ be another measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Then, the functional $Q\colon L^\infty(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \boldsymbol{\mu}) \to \mathbb{R}$ defined by*

$$Q(f) = \int_{\mathbb{R}^d} f(\boldsymbol{x}) d\boldsymbol{\sigma}(\boldsymbol{x}) \tag{4.9}$$

is a bounded linear functional and

$$\|Q\|^{L^\infty} = \|\boldsymbol{\sigma}\|. \tag{4.10}$$

Theorem 4.8. *(Banach-Steinhaus) Let X, Y be two Banach spaces and let $\{F_i\}_{i\in I}$ be a family of bounded linear operators from X into Y such that*

$$\sup_{i\in I}\{\|F_i(x)\|_Y\} < \infty,\ \forall\, x \in X. \tag{4.11}$$

Then, the operator norms are also bounded,

$$\sup_{i\in I}\{\|F_i\|\} < \infty. \tag{4.12}$$

Finally, Sobolev spaces are here defined (as in [4]). They will be used later in the error analysis of positive summability methods.

Definition 4.9. *Let $1 \le p < \infty$ and $l \in \mathbb{N}_0$. The Sobolev space $H_p^l(\mathbb{R}^d)$ is defined as*

$$H_p^l(\mathbb{R}^d) = \{f \in L^p(\mathbb{R}^d)\colon D^{(\boldsymbol{n})} f \text{ exists and } D^{(\boldsymbol{n})} f \in L^p(\mathbb{R}^d),\ \forall\, \boldsymbol{n} \text{ such that } |\boldsymbol{n}| \le l\}, \tag{4.13}$$

where $D^{(\boldsymbol{n})} f$ is the $\boldsymbol{n}^{th}$-weak derivative of f. If equipped with the norm

$$\|f\|_{H_p^l} = \left(\sum_{|\boldsymbol{n}|=0}^{l} \|D^{(\boldsymbol{n})} f\|_p^p\right)^{1/p} \tag{4.14}$$

it becomes a Banach space.

4.2 Summability methods based on kernels

4.2.1 General summability methods

Following the notation from chapter 2, $\boldsymbol{\mu}$ is the image measure of the basis random vector $\boldsymbol{\Xi} = (\Xi_1, \ldots, \Xi_d)$ and denote by $\boldsymbol{S} = \text{supp}\boldsymbol{\mu}$ its support. Assume from now on that $\boldsymbol{S}$ is compact. In this case, it holds that

$$L^p(\boldsymbol{S}, \mathcal{B}(\boldsymbol{S}), \boldsymbol{\mu}) \subset L^q(\boldsymbol{S}, \mathcal{B}(\boldsymbol{S}), \boldsymbol{\mu}),\ \forall\, 1 \le q \le p \le \infty, \tag{4.15}$$

which can be proven by using the Hölder's inequality. Let now $f \in L^1(\boldsymbol{S}, \mathcal{B}(\boldsymbol{S}), \boldsymbol{\mu})$ and define its *Fourier coefficients* by

$$\hat{f}_{\boldsymbol{n}} = \int_{\boldsymbol{S}} f(\boldsymbol{x}) P_{\boldsymbol{n}}(\boldsymbol{x}) d\boldsymbol{\mu}(\boldsymbol{x}),\ \boldsymbol{n} \in \mathbb{N}_0^d. \tag{4.16}$$

Let $\tau_i \colon \mathbb{N}_0 \to \mathbb{N}_0$, $i = 1, \dots, d$, be sequences with $\lim_{N\to\infty} \tau_i(N) = \infty$, and $\omega_{N,\boldsymbol{n}}$ be complex numbers for all $N \in \mathbb{N}_0$, $\boldsymbol{n} \in \mathbb{N}_0^d$. Define for $N \in \mathbb{N}_0$ the linear operators

$$F_N \colon L^p(\boldsymbol{S}, \mathcal{B}(\boldsymbol{S}), \boldsymbol{\mu}) \to L^p(\boldsymbol{S}, \mathcal{B}(\boldsymbol{S}), \boldsymbol{\mu}),\ 1 \leq p \leq \infty \tag{4.17}$$

by

$$F_N(f) = \sum_{n_1=0}^{\tau_1(N)} \cdots \sum_{n_d=0}^{\tau_d(N)} \omega_{N,\boldsymbol{n}} \hat{f}_{\boldsymbol{n}} P_{\boldsymbol{n}} h_{\boldsymbol{n}}. \tag{4.18}$$

Lemma 4.10. *The operators F_N as defined above are continuous.*

Proof. Let $f \in L^p(\boldsymbol{S}, \mathcal{B}(\boldsymbol{S}), \boldsymbol{\mu}), 1 \leq p \leq \infty$. Then, using the triangle inequality for the norm yields

$$\|F_N(f)\|_p \leq \sum_{n_1=0}^{\tau_1(N)} \cdots \sum_{n_d=0}^{\tau_d(N)} \|\omega_{N,\boldsymbol{n}} \hat{f}_{\boldsymbol{n}} P_{\boldsymbol{n}} h_{\boldsymbol{n}}\|_p = \sum_{n_1=0}^{\tau_1(N)} \cdots \sum_{n_d=0}^{\tau_d(N)} |\omega_{N,\boldsymbol{n}}| |\hat{f}_{\boldsymbol{n}}| \|P_{\boldsymbol{n}}\|_p h_{\boldsymbol{n}}. \tag{4.19}$$

From the definition of the coefficients $\hat{f}_{\boldsymbol{n}}$ it follows

$$|\hat{f}_{\boldsymbol{n}}| \leq \int_{\boldsymbol{S}} |f(\boldsymbol{x})| |P_{\boldsymbol{n}}(\boldsymbol{x})| d\boldsymbol{\mu}(\boldsymbol{x}) \leq \sup_{\boldsymbol{x}\in\boldsymbol{S}} |P_{\boldsymbol{n}}(\boldsymbol{x})| \int_{\boldsymbol{S}} |f(\boldsymbol{x})| d\boldsymbol{\mu}(\boldsymbol{x}) = \|P_{\boldsymbol{n}}\|_\infty \|f\|_1,$$

so that equation (4.19) becomes

$$\begin{aligned}
\|F_N(f)\|_p &\leq \sum_{n_1=0}^{\tau_1(N)} \cdots \sum_{n_d=0}^{\tau_d(N)} |\omega_{N,\boldsymbol{n}}| \|P_{\boldsymbol{n}}\|_\infty \|f\|_1 \|P_{\boldsymbol{n}}\|_p h_{\boldsymbol{n}} \\
&= \left(\sum_{n_1=0}^{\tau_1(N)} \cdots \sum_{n_d=0}^{\tau_d(N)} |\omega_{N,\boldsymbol{n}}| \|P_{\boldsymbol{n}}\|_\infty \|P_{\boldsymbol{n}}\|_p h_{\boldsymbol{n}} \right) \|f\|_1.
\end{aligned}$$

From Hölder's inequality, it follows immediately

$$\|f\|_1 \leq \|f\|_p,\ \forall\, 1 \leq p \leq \infty,$$

which completes the proof. ∎

This lemma thus assures the existence of the operator norms $\|F_N\|$ for all $N \in \mathbb{N}_0$. The following theorem states the conditions under which the weighted truncated expansions in (4.18) converge to the actual function of interest f. When these conditions are satisfied, it is guaranteed that in the limit the weighted expansions converge to the correct quantity and thus the weights do not affect the approximation property of the polynomial expansions.

Theorem 4.11. *Let $1 \leq p < \infty$ and $\{F_N\}_{N=0}^\infty$ be a sequence of operators defined by (4.18). Then it holds that*

$$F_N(f) \xrightarrow{L^p} f,\ \forall\, f \in L^p(\boldsymbol{S}, \mathcal{B}(\boldsymbol{S}), \boldsymbol{\mu}) \tag{4.20}$$

if and only if

$$\lim_{N\to\infty} \omega_{N,\boldsymbol{n}} = 1,\ \forall\, \boldsymbol{n} \in \mathbb{N}_0^d, \tag{4.21}$$

and there exists a constant $C > 0$, independent of N such that

$$\|F_N\|^{L^p} < C,\ \forall\, N \in \mathbb{N}_0. \tag{4.22}$$

Proof. Before the proof is stated, it is noted that if $\lim_{N\to\infty} F_N(f) = f$ for all functions $f \in L^p(\boldsymbol{S}, \mathcal{B}(\boldsymbol{S}), \boldsymbol{\mu})$, then the following equivalence holds true

$$\left[\lim_{N\to\infty} F_N(P_{\boldsymbol{n}}) = P_{\boldsymbol{n}}\right] \Leftrightarrow \left[\lim_{N\to\infty} \omega_{N,\boldsymbol{n}} = 1\right]. \tag{4.23}$$

This can be easily seen by substituting $P_{\boldsymbol{n}}$ as the function f in (4.18). Using the orthogonality of the polynomials, it then follows that

$$F_N(P_{\boldsymbol{n}}) = \omega_{N,\boldsymbol{n}} P_{\boldsymbol{n}},\ \forall\, \boldsymbol{n} \in \mathbb{N}_0^d,\ N \in \mathbb{N}_0.$$

Consider firstly the forward direction for the proof of the theorem. Equation (4.2.3) follows then immediately from the remark above. Moreover, equation (4.20) is equivalent to

$$\forall\, \epsilon > 0, \exists\, N' \in \mathbb{N}_0 : \forall\, N \geq N', \|F_N(f) - f\|_p < \epsilon.$$

This means, that

$$\|F_N(f)\|_p \leq \max\{\|F_0(f)\|_p, \ldots \|F_{N'-1}(f)\|_p, \|f\|_p + \epsilon\}.$$

Thus, from Theorem 4.8, (4.22) follows immediately.

Conversely, assume that equations (4.2.3) and (4.22) hold. Let $f \in L^p(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \boldsymbol{\mu})$ and $\epsilon > 0$. Equations (4.2.3) and (4.23) imply that $\lim_{N\to\infty} F_N(Q) = Q$ for all polynomials $Q \in \mathcal{P}$. Since $\mathcal{P}$ is dense in $L^p(\boldsymbol{S}, \mathcal{B}(\boldsymbol{S}), \boldsymbol{\mu})$, for $1 \leq p < \infty$ [68] one may choose $Q \in \mathcal{P}$ with $\|Q - f\|_p < \epsilon$. Moreover, for any such Q, and any ϵ, there exists an $N' \in \mathbb{N}_0$, such that $\forall\, N \geq N',\ \|F_N(Q) - Q\|_p < \epsilon$. By the triangle inequality one has

$$\|f - F_N(f)\|_p \leq \|f - Q\|_p + \|Q - F_N(Q)\|_p + \|F_N(Q) - F_N(f)\|_p.$$

Using now the above and equations (4.4) and (4.22), one has

$$\begin{aligned} \|f - F_N(f)\|_p &\leq \|f - Q\|_p + \|Q - F_N(Q)\|_p + C\|Q - f\|_p \\ &\leq (2 + C)\epsilon \end{aligned}$$

for all $N \geq N'$ and for all $\epsilon > 0$, which completes the proof. ■

One wishes now to verify equation (4.22) for the operators under consideration, that is one wishes to find an upper bound for the operator norms $\|F_N\|^{L^p}$, which is independent of N. To this end, define for all $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{S}$ and for all $N \in \mathbb{N}_0$ the kernels

$$K_N(\boldsymbol{x}, \boldsymbol{y}) = \sum_{n_1=0}^{\tau_1(N)} \cdots \sum_{n_d=0}^{\tau_d(N)} \omega_{N,\boldsymbol{n}} P_{\boldsymbol{n}}(\boldsymbol{x}) P_{\boldsymbol{n}}(\boldsymbol{y}) h_{\boldsymbol{n}}. \tag{4.24}$$

Then, by substituting the definition of the Fourier coefficients (4.16) in (4.18) one gets the integral representation for the operators

$$F_N(f)(\boldsymbol{x}) = \int_{\boldsymbol{S}} f(\boldsymbol{y}) K_N(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{\mu}(\boldsymbol{y}). \tag{4.25}$$

The following lemma provides with upper bounds for the norm of the operators. These bounds depend on the kernels defined in (4.24).

Lemma 4.12. *Let $1 \leq p \leq \infty$ and F_N defined by (4.25). Then*

$$\|F_N\|^{L^p} \leq \|F_N\|^{L^1} = \|F_N\|^{L^\infty} = \sup_{\boldsymbol{x}\in\boldsymbol{S}} \int_{\boldsymbol{S}} |K_N(\boldsymbol{x},\boldsymbol{y})| d\boldsymbol{\mu}(\boldsymbol{y}). \tag{4.26}$$

Proof. **Case** $p = \infty$. Then for all $f \in L^\infty(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \boldsymbol{\mu})$, one has

$$\begin{aligned}
\|F_N(f)\|_\infty &= \sup_{\boldsymbol{x}\in\boldsymbol{S}} |F_N(f)(\boldsymbol{x})| \\
&\leq \sup_{\boldsymbol{x}\in\boldsymbol{S}} \int_{\boldsymbol{S}} |f(\boldsymbol{y}) K_N(\boldsymbol{x},\boldsymbol{y}) d\boldsymbol{\mu}(\boldsymbol{y})| \\
&= \sup_{\boldsymbol{x}\in\boldsymbol{S}} \int_{\boldsymbol{S}} |f(\boldsymbol{y})| |K_N(\boldsymbol{x},\boldsymbol{y})| d\boldsymbol{\mu}(\boldsymbol{y}) \\
&\leq \sup_{\boldsymbol{x}\in\boldsymbol{S}} \int_{\boldsymbol{S}} \|f\|_\infty |K_N(\boldsymbol{x},\boldsymbol{y})| d\boldsymbol{\mu}(\boldsymbol{y}) \\
&\leq \left(\sup_{\boldsymbol{x}\in\boldsymbol{S}} \int_{\boldsymbol{S}} |K_N(\boldsymbol{x},\boldsymbol{y})| d\boldsymbol{\mu}(\boldsymbol{y}) \right) \|f\|_\infty,
\end{aligned}$$

which means that

$$\|F_N\|^{L^\infty} \leq \left(\sup_{\boldsymbol{x}\in\boldsymbol{S}} \int_{\boldsymbol{S}} |K_N(\boldsymbol{x},\boldsymbol{y})| d\boldsymbol{\mu}(\boldsymbol{y}) \right). \tag{4.27}$$

To establish the equality, one can show the reverse inequality in (4.27). Let $\boldsymbol{x} \in \boldsymbol{S}$ be arbitrary, fixed. Consider the linear functional $Q_{\boldsymbol{x}} \colon L^\infty \to \mathbb{R}$ defined by

$$Q_{\boldsymbol{x}}(f) = \int_{\boldsymbol{S}} f(\boldsymbol{y}) K_N(\boldsymbol{x},\boldsymbol{y}) d\boldsymbol{\mu}(\boldsymbol{y}).$$

By Theorem 4.7, it holds for the norm of $Q_{\boldsymbol{x}}$

$$\|Q_{\boldsymbol{x}}\|^{L^\infty} = \int_{\boldsymbol{S}} |K_N(\boldsymbol{x},\boldsymbol{y})| d\boldsymbol{\mu}(\boldsymbol{y}).$$

On the other hand from the definition of the operator norm, one has

$$\|Q_{\boldsymbol{x}}\|^{L^\infty} = \sup_{f\in S_{L^\infty}} |Q_{\boldsymbol{x}}(f)|.$$

All together, for all $\boldsymbol{x} \in \boldsymbol{S}$ one has

$$\begin{aligned}
\|Q_{\boldsymbol{x}}\|^{L^\infty} &= \int_{\boldsymbol{S}} |K_N(\boldsymbol{x},\boldsymbol{y})| d\boldsymbol{\mu}(\boldsymbol{y}) \\
&= \sup_{f\in S_{L^\infty}} |Q_{\boldsymbol{x}}(f)| \\
&= \sup_{f\in S_{L^\infty}} \left| \int_{\boldsymbol{S}} f(\boldsymbol{y}) K_N(\boldsymbol{x},\boldsymbol{y}) d\boldsymbol{\mu}(\boldsymbol{y}) \right| \\
&= \sup_{f\in S_{L^\infty}} |F_N(f)(\boldsymbol{x})| \\
&\leq \sup_{f\in S_{L^\infty}} \sup_{\boldsymbol{x}\in\boldsymbol{S}} |F_N(f)(\boldsymbol{x})| \\
&= \sup_{f\in S_{L^\infty}} \| F_N(f) \|_\infty = \|F_N\|^{L^\infty}.
\end{aligned} \tag{4.28}$$

Thus, by (4.27) and by taking supremums on both sides of (4.28) it follows that

$$\|F_N\|^{L^\infty} = \sup_{\boldsymbol{x}\in\boldsymbol{S}} \int_{\boldsymbol{S}} |K_N(\boldsymbol{x},\boldsymbol{y})| d\boldsymbol{\mu}(\boldsymbol{y}).$$

Case $p = 1$. By using the definition of the L^1-operator norm, Fubini's theorem, and the symmetry property of the kernel K_N, namely

$$K_N(\boldsymbol{x},\boldsymbol{y}) = K_N(\boldsymbol{y},\boldsymbol{x}),\ \forall\, \boldsymbol{x},\boldsymbol{y} \in \boldsymbol{S},$$

one has

$$\begin{aligned}
\|F_N\|^{L^1} &= \sup_{g\in S_{L^1}} \|F_N(g)\|_1 \\
&= \sup_{g\in S_{L^1}} \sup_{f\in S_{L^\infty}} \left| \int_{\boldsymbol{S}} f(\boldsymbol{x}) \int_{\boldsymbol{S}} g(\boldsymbol{y}) K_N(\boldsymbol{x},\boldsymbol{y}) d\boldsymbol{\mu}(\boldsymbol{y}) d\boldsymbol{\mu}(\boldsymbol{x}) \right| \\
&= \sup_{g\in S_{L^1}} \sup_{f\in S_{L^\infty}} \left| \int_{\boldsymbol{S}} g(\boldsymbol{y}) \int_{\boldsymbol{S}} f(\boldsymbol{x}) K_N(\boldsymbol{y},\boldsymbol{x}) d\boldsymbol{\mu}(\boldsymbol{x}) d\boldsymbol{\mu}(\boldsymbol{y}) \right| \\
&= \sup_{f\in S_{L^\infty}} \sup_{g\in S_{L^1}} \left| \int_{\boldsymbol{S}} g(\boldsymbol{y}) F_N(f)(\boldsymbol{y}) d\boldsymbol{\mu}(\boldsymbol{y}) \right| \\
&= \sup_{f\in S_{L^\infty}} \|F_N(f)\|_\infty = \|F_N\|^{L^\infty}.
\end{aligned}$$

Case $1 < p < \infty$. Let q be such that $1/p + 1/q = 1$, and $f \in L^p(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \boldsymbol{\mu})$. Applying Hölder's inequality with respect to the measure $|K_N(\boldsymbol{x},\boldsymbol{y})| d\mu(\boldsymbol{y})$ yields

$$\left(\int_{\boldsymbol{S}} |f(\boldsymbol{y})| |K_N(\boldsymbol{x},\boldsymbol{y})| d\boldsymbol{\mu}(\boldsymbol{y}) \right)^p \le \left(\int_{\boldsymbol{S}} |K_N(\boldsymbol{x},\boldsymbol{y})| d\boldsymbol{\mu}(\boldsymbol{y}) \right)^{p/q} \int_{\boldsymbol{S}} |f(\boldsymbol{y})|^p |K_N(\boldsymbol{x},\boldsymbol{y})| d\boldsymbol{\mu}(\boldsymbol{y}).$$

Using the above equation and Fubini's theorem one gets

$$\begin{aligned}
\|F_N(f)\|_p^p &= \int_{\boldsymbol{S}} |F_N(f)(\boldsymbol{x})|^p d\boldsymbol{\mu}(\boldsymbol{x}) \\
&= \int_{\boldsymbol{S}} | \int_{\boldsymbol{S}} f(\boldsymbol{y}) K_N(\boldsymbol{x},\boldsymbol{y}) d\boldsymbol{\mu}(\boldsymbol{y})|^p d\boldsymbol{\mu}(\boldsymbol{x}) \\
&\le \left(\sup_{\boldsymbol{x}\in\boldsymbol{S}} \int_{\boldsymbol{S}} |K_N(\boldsymbol{x},\boldsymbol{y})| d\boldsymbol{\mu}(\boldsymbol{y}) \right)^{p/q} \int_{\boldsymbol{S}} \int_{\boldsymbol{S}} |f(\boldsymbol{y})|^p |K_N(\boldsymbol{y},\boldsymbol{x})| d\boldsymbol{\mu}(\boldsymbol{x}) d\boldsymbol{\mu}(\boldsymbol{y}) \\
&\le \left(\sup_{\boldsymbol{x}\in\boldsymbol{S}} \int_{\boldsymbol{S}} |K_N(\boldsymbol{x},\boldsymbol{y})| d\boldsymbol{\mu}(\boldsymbol{y}) \right)^{p/q} \left(\sup_{\boldsymbol{y}\in\boldsymbol{S}} \int_{\boldsymbol{S}} |K_N(\boldsymbol{y},\boldsymbol{x})| d\boldsymbol{\mu}(\boldsymbol{x}) \right) \int_{\boldsymbol{S}} |f(\boldsymbol{y})|^p d\boldsymbol{\mu}(\boldsymbol{y}) \\
&\le \left(\sup_{\boldsymbol{x}\in\boldsymbol{S}} \int_{\boldsymbol{S}} |K_N(\boldsymbol{x},\boldsymbol{y})| d\boldsymbol{\mu}(\boldsymbol{y}) \right)^{p/q} \left(\sup_{\boldsymbol{y}\in\boldsymbol{S}} \int_{\boldsymbol{S}} |K_N(\boldsymbol{y},\boldsymbol{x})| d\boldsymbol{\mu}(\boldsymbol{x}) \right) \|f\|_p^p.
\end{aligned}$$

Hence,

$$\|F_N\|^{L^p} \le \left(\sup_{\boldsymbol{x}\in\boldsymbol{S}} \int_{\boldsymbol{S}} |K_N(\boldsymbol{x},\boldsymbol{y})| d\boldsymbol{\mu}(\boldsymbol{y}) \right)^{1/q} \left(\sup_{\boldsymbol{y}\in\boldsymbol{S}} \int_{\boldsymbol{S}} |K_N(\boldsymbol{y},\boldsymbol{x})| d\boldsymbol{\mu}(\boldsymbol{x}) \right)^{1/p}.$$

By the symmetry of the kernel K_N, it follows

$$\|F_N\|^{L^p} \leq \sup_{\boldsymbol{x}\in\boldsymbol{S}} \int_{\boldsymbol{S}} |K_N(\boldsymbol{x},\boldsymbol{y})| d\mu(\boldsymbol{y}),\ \forall\, 1 < p < \infty,$$

and the proof is completed. ■

So far, the estimates for the operator norms depend on the index N. Next, it is shown that when one restricts itself on positive operators, one can raise this problem and thus fulfill the condition in equation (4.22).

4.2.2 Positive summability methods

Consider now the case where one is dealing with positive quantities f and wishes to establish the positivity and approximating property for the operators F_N. Firstly, a simple characterization of the positivity for the operators under consideration is stated.

Lemma 4.13. *The operator F_N is positive if and only if $K_N(\boldsymbol{x},\boldsymbol{y}) \geq 0$ for all $\boldsymbol{x},\boldsymbol{y} \in \boldsymbol{S}$.*

Proof. It is easily seen by (4.18) that $K_N(\boldsymbol{x},\boldsymbol{y}) \geq 0$ for all $\boldsymbol{x},\boldsymbol{y} \in \boldsymbol{S}$ implies the positivity of F_N.
Assume now that F_N is positive and that there exist $\boldsymbol{x}_0, \boldsymbol{y}_0 \in \boldsymbol{S}$ with $K_N(\boldsymbol{x}_0,\boldsymbol{y}_0) < 0$. Then, because of the continuity of the kernel K_N, there is a $\delta < 0$ and an open set $U \subset \boldsymbol{S}$ with $\boldsymbol{y}_0 \in U$ such that

$$K_N(\boldsymbol{x}_0,\boldsymbol{y}) < \delta < 0,\ \forall\, \boldsymbol{y} \in U.$$

There exists a compact set $V \subsetneq U$ with $\boldsymbol{\mu}(V) > 0$. Due to Lemma 4.4, there exists a continuous function $g\colon \boldsymbol{S} \to [0,1]$ with $g(\boldsymbol{y}) = 1$, for all $\boldsymbol{y} \in V$ and $g(\boldsymbol{y}) = 0$, for all $\boldsymbol{y} \in \boldsymbol{S} \setminus U$. Thus,

$$\begin{aligned} F_N(g)(\boldsymbol{x}_0) &= \int_{\boldsymbol{S}} g(\boldsymbol{y}) K_N(\boldsymbol{x}_0,\boldsymbol{y}) d\boldsymbol{\mu}(\boldsymbol{y}) \\ &= \int_{\boldsymbol{S}\setminus U} g(\boldsymbol{y}) K_N(\boldsymbol{x}_0,\boldsymbol{y}) d\boldsymbol{\mu}(\boldsymbol{y}) + \int_U g(\boldsymbol{y}) K_N(\boldsymbol{x}_0,\boldsymbol{y}) d\boldsymbol{\mu}(\boldsymbol{y}) \\ &< \delta \int_U g(\boldsymbol{y}) d\boldsymbol{\mu}(\boldsymbol{y}) < \delta \int_V g(\boldsymbol{y}) d\boldsymbol{\mu}(\boldsymbol{y}) = \delta \boldsymbol{\mu}(V) < 0 \end{aligned}$$

which contradicts the positivity of the operator F_N. ■

In the case of positive operators, the following results follow from Theorem 4.11 and Lemma 4.12. The first one gives a more concrete value to the norm of the positive operators and the second one simplifies the conditions under which the sequence of the positive operators converges.

Corollary 4.14. *If F_N is positive, then $\|F_N\|^{L^p} \leq \omega_{N,\mathbf{0}}$, $1 \leq p \leq \infty$, where $\mathbf{0} = (0,\ldots,0) \in \mathbb{N}_0^d$.*

Proof. If F_N is positive, then $K_N(\boldsymbol{x}, \boldsymbol{y}) \geq 0$ for all $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{S}$, see Lemma 4.13. Therefore, by orthogonality

$$\begin{aligned}
\|F_N\|^{L^\infty} &= \sup_{\boldsymbol{x}\in\boldsymbol{S}} \int_{\boldsymbol{S}} K_N(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{\mu}(\boldsymbol{y}) \\
&= \sup_{\boldsymbol{x}\in\boldsymbol{S}} \sum_{n_1=0}^{\tau_1(N)} \cdots \sum_{n_d=0}^{\tau_d(N)} \omega_{N,\boldsymbol{n}} P_{\boldsymbol{n}}(\boldsymbol{x}) h_{\boldsymbol{n}} \int_{\boldsymbol{S}} P_{\boldsymbol{n}}(\boldsymbol{y}) d\boldsymbol{\mu}(\boldsymbol{y}) \\
&= \sup_{\boldsymbol{x}\in\boldsymbol{S}} \sum_{n_1=0}^{\tau_1(N)} \cdots \sum_{n_d=0}^{\tau_d(N)} \omega_{N,\boldsymbol{n}} P_{\boldsymbol{n}}(\boldsymbol{x}) h_{\boldsymbol{n}} \delta_{\boldsymbol{0n}} \frac{1}{h_{\boldsymbol{n}}} \\
&= \sup_{\boldsymbol{x}\in\boldsymbol{S}} P_{\boldsymbol{0}}(\boldsymbol{x}) \omega_{N,\boldsymbol{0}} = \omega_{N,\boldsymbol{0}},
\end{aligned}$$

as it was assumed in Chapter 2 that $P_{\boldsymbol{0}}(\boldsymbol{x}) = 1$. By using the fact that $\|F_N\|^{L^p} \leq \|F_N\|^{L^\infty}$, as established in Lemma 4.12, the conclusion follows. ■

Finally, the central theorem for this chapter is given. It states that in the case of positive kernel operators, defined as in equations (4.24) and (4.25), only the limiting condition on the weights has to be verified. The condition of the uniform boundedness of the operator norms in Theorem 4.11 follows then immediately.

Theorem 4.15. *If $\{F_N\}_{N=0}^{\infty}$ is a sequence of positive operators, then $F_N(f) \overset{L^p}{\to} f$ for all $f \in L^p(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \boldsymbol{\mu}),\ 1 \leq p < \infty$ if and only if $\lim_{N\to\infty} \omega_{N,\boldsymbol{n}} = 1$ for all $\boldsymbol{n} \in \mathbb{N}_0^d$.*

Proof. The forward direction of the proof is obvious from Theorem 4.11. It remains to show that if $\lim_{N\to\infty} \omega_{N,\boldsymbol{n}} = 1$ for all $\boldsymbol{n} \in \mathbb{N}_0^d$, then the condition $\|F_N\|^{L^p} < C,\ \forall\, N \in \mathbb{N}_0$ holds true.

If $\{F_N\}_{N=0}^{\infty}$ is a sequence of positive operators, then Corollary 4.14 yields $\|F_N\|^{L^p} \leq \omega_{N,\boldsymbol{0}}$. Hence, if equation (4.2.3) is true, then especially for $\boldsymbol{n} = \boldsymbol{0}$, one has that $\lim_{N\to\infty} \omega_{N,\boldsymbol{0}} = 1$. This means that the sequence $\{\omega_{N,\boldsymbol{0}}\}_{N\in\mathbb{N}_0}$ is bounded, thus there exists a constant $C > 0$ such that $\omega_{N,\boldsymbol{0}} < C$, for all $N \in \mathbb{N}_0$ and the proof is completed. ■

More generally, the positive kernel operators allow to preserve boundedness, as shown in the following lemma.

Lemma 4.16. *Let $m, M \in \mathbb{R}$ be given real numbers and let $\{F_N\}_{N=0}^{\infty}$ be a sequence of positive operators. Then, $m \leq F_N(f)(\boldsymbol{x}) \leq M$, $\boldsymbol{\mu}$-a.e. for all $f \in L^p(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \boldsymbol{\mu}), 1 \leq p < \infty$ such that $m \leq f(\boldsymbol{x}) \leq M$, $\boldsymbol{\mu}$-a.e. if and only if $\omega_{N,\boldsymbol{0}} = 1$, for all $N \in \mathbb{N}_0$.*

Proof. Let $f \in L^p(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \boldsymbol{\mu}), 1 \leq p < \infty$ and assume that $f(\boldsymbol{x}) \geq m$, $\boldsymbol{\mu}$-a.e. Then, the function $g(\boldsymbol{x}) = f(\boldsymbol{x}) - m$ will be positive and thus $F_N(g)$ will be also positive for all

$N \in \mathbb{N}_0$, since $\{F_N\}_{N=0}^{\infty}$ is a assumed to be a sequence of positive operators. Then

$$\begin{aligned} 0 \leq F_N(g) &= \sum_{n_1=0}^{\tau(N)} \cdots \sum_{n_d=0}^{\tau(N)} \omega_{N,\boldsymbol{n}} \hat{g}_{\boldsymbol{n}} P_{\boldsymbol{n}} h_{\boldsymbol{n}} \\ &= \sum_{n_1=0}^{\tau(N)} \cdots \sum_{n_d=0}^{\tau(N)} \omega_{N,\boldsymbol{n}} (\hat{f}_{\boldsymbol{n}} - m\delta_{\boldsymbol{n}\mathbf{0}}) P_{\boldsymbol{n}} h_{\boldsymbol{n}} \\ &= \sum_{n_1=0}^{\tau(N)} \cdots \sum_{n_d=0}^{\tau(N)} \omega_{N,\boldsymbol{n}} \hat{f}_{\boldsymbol{n}} P_{\boldsymbol{n}} h_{\boldsymbol{n}} - \omega_{N,\mathbf{0}} m \\ &= F_N(f) - \omega_{N,\mathbf{0}} m \end{aligned}$$

from it follows that the condition $\omega_{N,\mathbf{0}} \geq 1, \forall N \in \mathbb{N}_0$ must hold. By bounding the truncated expansion from above, it follows that $\omega_{N,\mathbf{0}} \leq 1, \forall\, N \in \mathbb{N}_0$, and the proof is completed. ■

A method to construct multivariate kernels is by using univariate ones in the following way. Assume that one has the one-dimensional kernel

$$L_N(x,y) = \sum_{n=0}^{\tau(N)} \omega_{N,n} P_n(x) P_n(y) h_n,\ x, y \in \mathbb{R}. \tag{4.29}$$

Consider the kernel $K_N(\boldsymbol{x}, \boldsymbol{y})$ defined as the product

$$\begin{aligned} K_N(\boldsymbol{x}, \boldsymbol{y}) &= L_N(x_1, y_1) \cdots L_N(x_d, y_d) \\ &= \sum_{n_1=0}^{\tau(N)} \cdots \sum_{n_d=0}^{\tau(N)} \omega_{N,\boldsymbol{n}} P_{\boldsymbol{n}}(\boldsymbol{x}) P_{\boldsymbol{n}}(\boldsymbol{y}) h_{\boldsymbol{n}} \end{aligned} \tag{4.30}$$

with $\omega_{N,\boldsymbol{n}} = \omega_{N,n_1} \cdots \omega_{N,n_d}$. It is obvious that the kernel $K_N(\boldsymbol{x}, \boldsymbol{y})$ is positive if the one-dimensional kernel L_N is positive. Therefore, due to Corollary 4.15, it is sufficient to construct positive kernels L_N with $\lim_{N\to\infty} \omega_{N,n} = 1$ for all $n \in \mathbb{N}_0$.

REMARK Note here that for the special construction of the multivariate kernels as in (4.30), the weighted polynomial approximation is carried out in the polynomial space $\tilde{\mathcal{P}}^N = lin\{P_{\boldsymbol{n}} \colon \boldsymbol{n} \in \mathbb{N}_0^d,\ \max_{1\leq i\leq d} n_i \leq \tau(N)\}$, whose dimension is $dim\tilde{\mathcal{P}^N} = \tau(N)^d$. The dimension of this space grows much faster than the dimension of the polynomial space $\mathcal{P}^N = lin\{P_{\boldsymbol{n}} \colon \boldsymbol{n} \in \mathbb{N}_0^d,\ |\boldsymbol{n}| \leq N\}$ which is usually used in practice.

4.2.3 Approximation error

Let $\tilde{\mathcal{P}}^N = lin\{P_{\boldsymbol{n}} \colon \boldsymbol{n} \in \mathbb{N}_0^d,\ \max_{1\leq i\leq d} n_i \leq N\}$. For a function $f \in L^2(\boldsymbol{S}, \mathcal{B}(\boldsymbol{S}), \boldsymbol{\mu})$, the polynomial expansion

$$G_N(f) = \sum_{n_1=0}^{N} \cdots \sum_{n_d=0}^{N} \hat{f}_{\boldsymbol{n}} P_{\boldsymbol{n}} h_{\boldsymbol{n}} \tag{4.31}$$

is the best approximation of f in the polynomial space $\tilde{\mathcal{P}}^N$. Results on the rate of convergence of the approximation error

$$\varepsilon_N = \|f - G_N(f)\|_2,\ N \in \mathbb{N}_0 \tag{4.32}$$

to zero are usually stated in terms of the norm of the function f in a suitable Sobolev space $H_p^l, l \in \mathbb{N}_0$. This rate depends also on the polynomial system under consideration. Assume that for a polynomial system $\{P_{\boldsymbol{n}} \colon \boldsymbol{n} \in \mathbb{N}_0^d\}$, there exists a constant $\tilde{C} > 0$ and a real sequence $\{\kappa(N,l)\}_{N,l\in\mathbb{N}_0}$ such that

$$\|f - G_N(f)\|_2 \leq \tilde{C}\kappa(N,l)\|f\|_{H_p^l},\ N \in \mathbb{N}. \tag{4.33}$$

Estimations of the type (4.33) can be found for example in [5] for Hermite polynomials and in [28] for Legendre polynomials and Chebyshev polynomials of the first kind. More general results can be found in the books [4,47,70] and the references therein.

The weighted expansion will preserve the positivity of the approximation on the cost of the optimality. Theorem 4.15 does not give any information on the quality of the approximation, therefore the behavior of the approximation error e_N is here considered, where

$$e_N = \|f - F_N(f)\|_2,\ N \in \mathbb{N}_0. \tag{4.34}$$

It was seen before, that the sequence of weights $\omega_{N,\boldsymbol{n}}$ should satisfy the limiting condition

$$\lim_{N\to\infty} \omega_{N,\boldsymbol{n}} = 1,\ \forall\, \boldsymbol{n} \in \mathbb{N}_0^d,$$

in order to gain positive approximation operators. Assume that in the one-dimensional case one has that

$$1 - \omega_{N,n} = O(\gamma(N)),\ \forall\, n = 0,\ldots,N,$$

where $\gamma \colon \mathbb{R} \to \mathbb{R}$ is a function such that $\lim_{x\to\infty} \gamma(x) = 0$. Then, it holds also that

$$\omega_{N,n} = 1 + O(\gamma(N)),\ \forall\, n = 0,\ldots,N.$$

In the multi-dimensional case, the weight sequences were defined as products of the one-dimensional weights. Therefore, it holds for their convergence rate

$$\begin{aligned} 1 - \omega_{N,\boldsymbol{n}} &= 1 - \omega_{N,n_1}\cdots\omega_{N,n_d} \\ &= 1 - (1 + O(\gamma(N)))\cdots(1 + O(\gamma(N))) \\ &= O(\gamma(N)),\ \forall\, \boldsymbol{n} = (n_1,\ldots,n_d) \in \mathbb{N}_0^d,\ \max_{1\leq i\leq d} n_i \leq N. \end{aligned}$$

For the error e_N in (4.34) and for all $N \in \mathbb{N}_0$, it holds

$$e_N^2 \leq \|f - G_N(f)\|_2^2 + \|G_N(f) - F_N(f)\|_2^2,$$

which means that the approximation error can be decomposed in the classical truncation error and in the error due to the introduction of weights in the expansion. For the second

term in the above equation, one has

$$\begin{aligned}\|G_N(f) - F_N(f)\|_2^2 &= \| \sum_{n_1=0}^{N} \cdots \sum_{n_d=0}^{N} \hat{f}_{\boldsymbol{n}} P_{\boldsymbol{n}} h_{\boldsymbol{n}} - \sum_{n_1=0}^{N} \cdots \sum_{n_d=0}^{N} \omega_{N,\boldsymbol{n}} \hat{f}_{\boldsymbol{n}} P_{\boldsymbol{n}} h_{\boldsymbol{n}} \|_2^2 \\ &= \| \sum_{n_1=0}^{N} \cdots \sum_{n_d=0}^{N} (1 - \omega_{N,\boldsymbol{n}}) \hat{f}_{\boldsymbol{n}} P_{\boldsymbol{n}} h_{\boldsymbol{n}} \|_2^2 \\ &= \sum_{n_1=0}^{N} \cdots \sum_{n_d=0}^{N} (1 - \omega_{N,\boldsymbol{n}})^2 \hat{f}_{\boldsymbol{n}}^2 h_{\boldsymbol{n}} \\ &\leq C^2 \gamma(N)^2 \sum_{n_1=0}^{N} \cdots \sum_{n_d=0}^{N} \hat{f}_{\boldsymbol{n}}^2 h_{\boldsymbol{n}}.\end{aligned} \tag{4.35}$$

The expansion $G_N(f)$ is the orthogonal projection of the function f on the space $\tilde{\mathcal{P}}^N$, therefore it holds

$$\|f\|_2^2 = \|f - G_N(f)\|_2^2 + \|G_N(f)\|_2^2, \tag{4.36}$$

and thus $\|G_N(f)\|_2^2 = \|f\|_2^2 - \|f - G_N(f)\|_2^2$. On the other hand, due to the orthogonality of the polynomials, it holds

$$\|G_N(f)\|_2^2 = \| \sum_{n_1=0}^{N} \cdots \sum_{n_d=0}^{N} \hat{f}_{\boldsymbol{n}} P_{\boldsymbol{n}} h_{\boldsymbol{n}} \|_2^2 = \sum_{n_1=0}^{N} \cdots \sum_{n_d=0}^{N} \hat{f}_{\boldsymbol{n}}^2 h_{\boldsymbol{n}}. \tag{4.37}$$

Combining equations (4.36) and (4.37), it follows that

$$\sum_{n_1=0}^{N} \cdots \sum_{n_d=0}^{N} \hat{f}_{\boldsymbol{n}}^2 h_{\boldsymbol{n}} = \|f\|_2^2 - \|f - G_N(f)\|_2^2. \tag{4.38}$$

Substituting (4.38) in (4.35), one obtains

$$\|G_N(f) - F_N(f)\|_2^2 \leq C^2 \gamma(N)^2 (\|f\|_2^2 - \|f - G_N(f)\|_2^2) \leq C^2 \gamma(N)^2 \|f\|_2^2.$$

All together, it holds for the error of the weighted expansions

$$e_N^2 \leq \tilde{C}^2 \kappa(N,l)^2 \|f\|_{H_p^l}^2 + C^2 \gamma(N)^2 \|f\|_2^2.$$

The definition of the Sobolev norm in equation (4.14) implies that

$$\|f\|_2^2 \leq \|f\|_{H_p^l}^2.$$

All together one has

$$e_N^2 \leq (\tilde{C}^2 \kappa(N,l)^2 + C^2 \gamma(N)^2) \|f\|_{H_p^l}^2,$$

and thus

$$e_N \leq (\tilde{C}^2 \kappa(N,l)^2 + C^2 \gamma(N)^2) \|f\|_{H_p^l}.$$

This means, that although one has a control over the approximation error, this may be worse than the error resulting from the classical unweighted approximation, depending on the chosen weighted sequence and its properties.

4.3 Positive kernels for Jacobi polynomials

In this section, explicit examples for weight sequences are summarized for the case of approximations with Jacobi polynomials. The Jacobi polynomials are firstly introduced.

Let the basic random variable Ξ have a Beta distribution with parameters α and β on $[-1,1]$. Then the image measure μ has a density with respect to the Lebesgue measure given by

$$d\mu^{(\alpha,\beta)}(x) = \frac{\Gamma(\alpha+\beta+2)}{2^{\alpha+\beta+1}\Gamma(\alpha+1)\Gamma(\beta+1)}(1-x)^{\alpha}(1+x)^{\beta}dx,\ \alpha,\beta>-1.$$

The sequence of Jacobi polynomials $\{P_n^{(\alpha,\beta)}\}_{n=0}^{\infty}$ is orthogonal with respect to the measure $\mu^{(\alpha,\beta)}$ and assume from now on that the polynomials are normalized so that

$$P_n^{(\alpha,\beta)}(1) = 1,\ \forall\, n\in\mathbb{N}_0,\ \alpha,\beta>-1.$$

Then, the Haar weights are given by the following expression

$$h_0^{(\alpha,\beta)} = 1 \quad \text{and} \quad h_n^{(\alpha,\beta)} = \frac{(2n+\alpha+\beta+1)\Gamma(\beta+1)\Gamma(n+\alpha+1)\Gamma(n+\alpha+\beta+1)}{\Gamma(\alpha+1)\Gamma(\alpha+\beta+2)\Gamma(n+1)\Gamma(n+\beta+1)}$$

for all $n\in\mathbb{N}$, where Γ denotes the Gamma function. One can easily verify the symmetric relation

$$P_n^{(\alpha,\beta)}(x) = P_n^{(\beta,\alpha)}(-x)P_n^{(\alpha,\beta)}(-1),\ \forall\, x\in[-1,1]. \tag{4.39}$$

Furthermore, in the case $\alpha\geq\beta$ and $\alpha\geq -1/2$ it holds

$$|P_n^{(\alpha,\beta)}(x)| \leq P_n^{(\alpha,\beta)}(1) = 1,\ \forall\, x\in[-1,1], \tag{4.40}$$

see [117].

In the case $\alpha=\beta$, the Jacobi polynomials are called *ultraspherical polynomials* and there exists an explicit formula for their linearization coefficients defined in equation (2.33). If $k\in\{|m-n|,|m-n|+2,\ldots,m+n\}$, then

$$c_{m,n,k}^{(\alpha,\alpha)} = \frac{(k+\alpha+1/2)\Gamma(2\alpha+1)\Gamma(m+1)\Gamma(n+1)\Gamma((m+n-k+1)/2+\alpha)}{((m+n+k+1)/2+\alpha)\Gamma(\alpha+1/2)\Gamma(\alpha+1/2)\Gamma(m+2\alpha+1)\Gamma(n+2\alpha+1)} \times$$
$$\frac{\Gamma((m-n+k+1)/2+\alpha)\Gamma((n-m+k+1)/2+\alpha)\Gamma((m+n+k)/2+2\alpha+1)}{\Gamma((m+n-k)/2+1)\Gamma((m-n+k)/2+1)\Gamma((n-m+k)/2+1)\Gamma((n+m+k+1)/2+\alpha)},$$

and else $c_{m,n,k}^{(\alpha,\alpha)}=0$ [54,55].

An important role in the construction of positive kernels for Jacobi polynomials plays the following *product formula*, which was first proved in [56]. In the case $\alpha\geq\beta>-1$ and ($\beta\geq -1/2$ or $\alpha+\beta\geq 0$) and for all $x,y\in[-1,1]$ there exists a probability measure $\pi_{x,y}$ such that

$$P_n^{(\alpha,\beta)}(x)P_n^{(\alpha,\beta)}(y) = \int_{\mathbb{R}} P_n^{(\alpha,\beta)}(z)d\pi_{x,y}^{(\alpha,\beta)}(z),\ \forall\, n\in\mathbb{N}_0. \tag{4.41}$$

Examples of positive kernels for Jacobi polynomials are now stated. In all the following examples and in order to show the positivity of the kernel, the idea was to expand first a positive quantity in the polynomial system under consideration and to use the product formula (4.41) to imply the positivity of the corresponding kernel.

4.3.1 De la Vallée-Poussin kernel

In [99] the following kernel was studied

$$V_N^{(\alpha,\beta)}(x,y) = \sum_{n=0}^{N} \nu_{N,n}^{(\alpha,\beta)} P_n^{(\alpha,\beta)}(x) P_n^{(\alpha,\beta)}(y) h_n^{(\alpha,\beta)}, \tag{4.42}$$

where

$$\nu_{N,n}^{(\alpha,\beta)} = \frac{\Gamma(N+1)\Gamma(N+\alpha+\beta+2)}{\Gamma(N-n+1)\Gamma(N+n+\alpha+\beta+2)},\ n=0,\ldots,N. \tag{4.43}$$

To verify for which $\alpha, \beta > -1$ the above kernel is positive, define

$$F_N^{(\alpha,\beta)}(x) = \sum_{n=0}^{N} \nu_{N,n}^{(\alpha,\beta)} P_n^{(\alpha,\beta)}(x) h_n^{(\alpha,\beta)},\ x \in [-1,1],\ N \in \mathbb{N}_0.$$

In [3] it is shown that $F_N^{(\alpha,\beta)}(x)$ can be expressed as follows

$$F_N^{(\alpha,\beta)}(x) = \sum_{n=0}^{N} \nu_{N,n}^{(\alpha,\beta)} P_n^{(\alpha,\beta)}(x) h_n^{(\alpha,\beta)} = \frac{\Gamma(\beta+1)\Gamma(N+\alpha+\beta+2)}{\Gamma(N+\beta+1)\Gamma(\alpha+\beta+2)} \left(\frac{1+x}{2}\right)^N,$$

from which it follows that for all $x \in [-1,1]$ and for all $N \in \mathbb{N}_0$, the quantity $F_N^{(\alpha,\beta)}(x)$ is positive. By using now the product formula (4.41), one has that in the case $\alpha \geq \beta > -1$ and ($\beta \geq -1/2$ or $\alpha + \beta \geq 0$)

$$V_N^{(\alpha,\beta)}(x,y) = \sum_{n=0}^{N} \nu_{N,n}^{(\alpha,\beta)} P_n^{(\alpha,\beta)}(x) P_n^{(\alpha,\beta)}(y) h_n^{(\alpha,\beta)} = \int F_N^{(\alpha,\beta)}(z) d\pi_{x,y}^{(\alpha,\beta)}(z) \geq 0$$

for all $x, y \in [-1,1]$, $N \in \mathbb{N}_0$. By using the symmetric relation in equation (4.39) one can see that

$$V_N^{(\alpha,\beta)}(x,y) = V_N^{(\beta,\alpha)}(-x,-y),\ \forall x,y \in [-1,1],\ N \in \mathbb{N}_0.$$

This means that if $\alpha + \beta \geq 0$ or ($\alpha \geq -1/2$ and $\beta \geq -1/2$), then $V_N^{(\alpha,\beta)}(x,y)$ is a positive kernel on $[-1,1]^2$. Furthermore, it was shown that the asymptotic behavior of the Gamma function yields

$$\lim_{N\to\infty} \nu_{N,n}^{(\alpha,\beta)} = 1 \quad \text{for all } n \in \mathbb{N}_0$$

and that the convergence rate of the weights is

$$1 - \nu_{N,n} = O(\frac{1}{N}),\ n=0,\ldots,N. \tag{4.44}$$

4.3.2 Fejér kernel

In [81, 82] the following kernel was proposed

$$FJ_N^{(\alpha,\alpha)}(x,y) = \sum_{n=0}^{2N} \frac{\chi_{2N,n}^{(\alpha,\beta)}}{\chi_{2N,0}^{(\alpha,\beta)}} P_n^{(\alpha,\beta)}(x) P_n^{(\alpha,\beta)}(y) h_n^{(\alpha,\beta)}, \tag{4.45}$$

where

$$\chi_{2N,n}^{(\alpha,\beta)} = \sum_{m=0}^{N} \sum_{i=|n-j|}^{N} c_{n,m,i}^{(\alpha,\beta)} h_i^{(\alpha,\beta)}. \tag{4.46}$$

It was shown that the function

$$F_N^{(\alpha,\beta)}(x) = \sum_{n=0}^{2N} \frac{\chi_{2N,n}^{(\alpha,\beta)}}{\chi_{2N,0}^{(\alpha,\beta)}} P_n^{(\alpha,\beta)}(x) h_n^{(\alpha,\beta)},\ x \in [-1,1],\ N \in \mathbb{N}_0$$

can be expressed as follows

$$F_N^{(\alpha,\beta)}(x) = \frac{\Gamma(N+\alpha+\beta+2)\Gamma(\beta+1)\Gamma(N+\alpha+2)}{\Gamma(\alpha+\beta+2)\Gamma(N+\beta+1)\Gamma(\alpha+2)\Gamma(N+1)} \left(P_N^{(\alpha+1,\beta)}(x)\right)^2$$

and is thus positive for all $x \in [-1,1]$ and all $N \in \mathbb{N}_0$. The required convergence of the weights was also shown in the aforementioned works. By using the product formula one concludes that $FJ_N^{(\alpha,\beta)}(x,y)$ is a positive kernel for all $\alpha \geq \beta > -1$ and $\alpha+\beta \geq -1$ and for all $x, y \in [-1,1]$. The convergence rate of the Fejér weights is

$$1 - \chi_{2N,n} = O(\frac{1}{N}),\ n = 0, \ldots, 2N. \tag{4.47}$$

4.3.3 Modified Fejér kernel

A more recent example was considered in [100] for the ultraspherical polynomials. Define the kernel

$$MF_N^{(\alpha,\alpha)}(x,y) = \sum_{n=0}^{N} \frac{\varphi_{N,n}^{(\alpha,\alpha)}}{\varphi_{N,0}^{(\alpha,\alpha)}} P_n^{(\alpha,\alpha)}(x) P_n^{(\alpha,\alpha)}(y) h_n^{(\alpha,\alpha)}, \tag{4.48}$$

with

$$\varphi_{N,n}^{(\alpha,\alpha)} = \sum_{k=n}^{N} \frac{\Gamma(2\alpha+2)\Gamma(k+1)}{\Gamma(k+2\alpha+2)} \tag{4.49}$$

$$= \begin{cases} N+1-n & \text{if} \quad \alpha = -1/2, \\ \displaystyle\sum_{k=n}^{N} \frac{1}{1+k} & \text{if} \quad \alpha = 0, \\ \dfrac{\Gamma(2\alpha+2)}{-2\alpha} \left(\dfrac{\Gamma(N+2)}{\Gamma(N+2\alpha+2)} - \dfrac{\Gamma(n+1)}{\Gamma(n+2\alpha+1)} \right) & \text{if} \quad \alpha \neq 0, -1/2. \end{cases} \tag{4.50}$$

The above kernel is positive for $\alpha \geq -1/2$ because the quantity

$$\frac{1 - P_{N+1}^{(\alpha,\alpha)}(x)}{1 - P_1^{(\alpha,\alpha)}(x)} = \sum_{n=0}^{N} \varphi_{N,n}^{(\alpha,\alpha)} P_n^{(\alpha,\alpha)}(x) h_n^{(\alpha,\alpha)}$$

is positive for all $N \in \mathbb{N}_0$ and all $x \in [-1,1)$ as one can see by equation (4.40). It was in addition showed that the weights satisfy the limiting condition if and only if $\alpha \leq 0$. All

together, the kernel $MF_N^{(\alpha,\alpha)}(x,y)$ is positive for all $-1/2 \leq \alpha \leq 0$ and for all $x, y \in [-1,1]$. The convergence rate of the modified Fejér weights depends on the parameter α and is given by

$$1 - \frac{\varphi_{N,n}^{(\alpha,\alpha)}}{\varphi_{N,0}^{(\alpha,\alpha)}} = \begin{cases} O(\frac{1}{N}) & \text{if} \quad \alpha = -\frac{1}{2}, \\ O(\frac{1}{logN}) & \text{if} \quad \alpha = 0. \end{cases} \tag{4.51}$$

4.3.4 Modified Jackson kernel

In the same work, the following kernel was considered

$$MJ_N^{(\alpha,\alpha)}(x,y) = \sum_{n=0}^{2N} \frac{\iota_{N,n}^{(\alpha,\alpha)}}{\iota_{N,0}^{(\alpha,\alpha)}} P_n^{(\alpha,\alpha)}(x) P_n^{(\alpha,\alpha)}(y) h_n^{(\alpha,\alpha)}, \tag{4.52}$$

where

$$\iota_{N,n}^{(\alpha,\alpha)} = \sum_{m=0}^{N} \sum_{k=0}^{N} c_{m,n,k}^{(\alpha,\alpha)} \varphi_{N,k}^{(\alpha,\alpha)} \varphi_{N,m}^{(\alpha,\alpha)} h_m^{(\alpha,\alpha)}. \tag{4.53}$$

Based on the representation

$$\left(\frac{1 - P_{N+1}^{(\alpha,\alpha)}(x)}{1 - P_1^{(\alpha,\alpha)}(x)} \right)^2 = \left(\sum_{m=0}^{N} \varphi_{N,m}^{(\alpha,\alpha)} P_m^{(\alpha,\alpha)}(x) h_m^{(\alpha,\alpha)} \right)^2 = \sum_{n=0}^{2N} \iota_{N,n}^{(\alpha,\alpha)} P_n^{(\alpha,\alpha)}(x) h_n^{(\alpha,\alpha)},$$

for all $N \in \mathbb{N}_0$, $x \in [-1,1)$ one can show that the kernels $MJ_N^{(\alpha,\alpha)}(x,y)$ remain positive for all $\alpha \geq -1/2$ and for all $x, y \in [-1,1]$. It was in addition shown that

$$\lim_{N \to \infty} \frac{\iota_{N,n}^{(\alpha,\alpha)}}{\iota_{N,0}^{(\alpha,\alpha)}} = 1 \quad \text{for all } n \in \mathbb{N}_0,$$

if and only if $\alpha \leq 1$. All together, the kernel $MJ_N^{(\alpha,\alpha)}(x,y)$ is positive on $[-1,1]^2$ provided that $-1/2 \leq \alpha \leq 1$. The convergence rate depends on the parameter α and is given by

$$1 - \frac{\iota_{N,n}^{(\alpha,\alpha)}}{\iota_{N,0}^{(\alpha,\alpha)}} = \begin{cases} O(\frac{1}{N^2}) & \text{if} \quad \alpha = -\frac{1}{2}, \\ O(\frac{logN}{N^2}) & \text{if} \quad \alpha = 0, \\ O(\frac{1}{N}) & \text{if} \quad \alpha = \frac{1}{2}, \\ O(\frac{1}{logN}) & \text{if} \quad \alpha = 1. \end{cases} \tag{4.54}$$

Finally, one should mention that there is also a product formula for the systems of generalized Chebyshev polynomials $P_n^{(\alpha,\beta)}(x)$, $\alpha, \beta > -1$ [80]. These polynomials are orthogonal on $[-1,1]$ with respect to the measure with density $(1-x^2)^\alpha |x|^{2\beta+1} dx$. Examples of positive kernels for such systems are given for instance in [82, 83, 99].

4.4 Application of weighted expansions in dynamical systems

Returning now to the problem of positivity in polynomial chaos approximations, this can be solved if one assumes a weighted PC expansion related to a positive kernel. So, instead of the classical PC approximation one can work with

$$x_N(t,\boldsymbol{\Xi}) = \sum_{n_1=0}^{\tau(N)} \cdots \sum_{n_d=0}^{\tau(N)} \omega_{N,\boldsymbol{n}} q_{\boldsymbol{n}}(t) P_{\boldsymbol{n}}(\boldsymbol{\Xi}) h_{\boldsymbol{n}},\ N \in \mathbb{N}_0, \tag{4.55}$$

see (4.18). As demonstrated in the previous section, the weights are pre-computed real numbers. They depend neither on the solution nor on the coefficients $q_{\boldsymbol{n}}(t)$ in (4.55). The PC coefficients can be computed by the numerical methods summarized in chapter 2.

4.4.1 Example: the logistic equation

One of the first continuous models describing population dynamics was the logistic equation introduced by Verhulst [122]. Assume $x(t)$ describes the size of a population at time t growing according to the following law

$$\dot{x}(t) = rx(t)(1 - \frac{x(t)}{K}),\ x(0) = x_0, \tag{4.56}$$

where r is the growth rate of the population, K its carrying capacity, and x_0 is the initial population size at time $t = 0$. The exact solution of this model is easily computed by separation of variables and reads

$$x(t) = \frac{K}{1 + \frac{K-x_0}{x_0} e^{-rt}},\ t \geq 0. \tag{4.57}$$

Equation (4.56) has two fixed points at $x = 0$ and $x = K$, with the origin being unstable and the state $x = K$ being stable. For biologically meaningful positive initial values, the solution $x(t)$ will converge to the carrying capacity as the time t growing to infinity. A completely different situation occurs if the initial condition x_0 is negative. In this case, a blow-up occurs in finite time as the denominator in (4.57) becomes zero at time

$$t_\star = \frac{1}{r} \ln \frac{K - x_0}{x_0}. \tag{4.58}$$

Fix now the parameters r and K. Assume that the initial condition is modeled as a random variable Θ and one wishes to examine how the uncertainty in this parameter propagates through the logistic equation. Now, the positivity becomes an important issue: if at some time point realizations of the solution become negative, then the expectation of the solution will be in finite time infinite and the expansion of the solution will thus provide with no useful information about the population.

For the numerical simulations, it was assumed that the fixed parameters have the values $r = 3$ and $K = 1$. The initial value Θ was modeled as a log-normal distributed random variable with parameters (v, σ), with $v = -1.2$ and $\sigma = 1.26$ as in [106]. Note that a non-negative distribution is assigned to Θ. The basis random variable Ξ is chosen to be a

uniform random variable on $[-1, 1]$ and the Legendre polynomials are the corresponding orthogonal polynomials. The random variable Θ has finite variance and therefore it admits an expansion

$$\Theta = g_\Theta(\Xi) = \sum_{n=0}^{\infty} \lambda_n P_n(\Xi) h_n. \tag{4.59}$$

The PC coefficients are estimated by using the isoprobabilistic transformation and their definition

$$\lambda_n = \int_{-1}^{1} g_\Theta(\xi) P_n(\xi) d\mu(\xi) = \int_{-1}^{1} F_\Theta^{-1}(F_\Xi(\xi)) P_n(\xi) d\mu(\xi), \ n \in \mathbb{N}_0. \tag{4.60}$$

The known closed forms for the distribution function and its inverse for log-normal and uniform random variables were substituted in (4.60) and the resulting integrals were estimated by using Gauss-Legendre quadrature with 10 nodes in the interval $[-1, 1]$. In Figure 4.1, the density estimators from a sample from the log-normal distribution and from samples generated by the estimated PC expansions for approximation orders $N = 4, 8, 10$ are plotted.

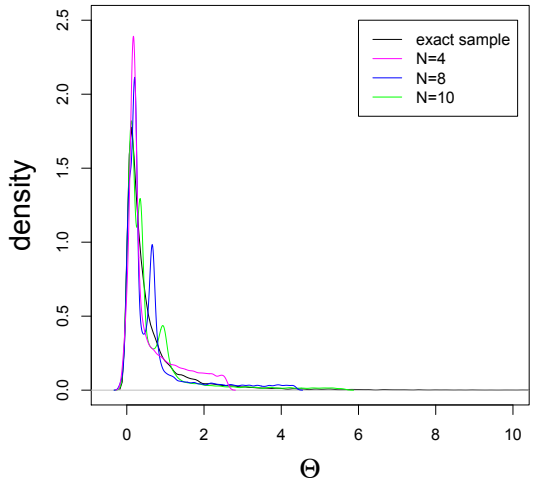

Figure 4.1: Density estimates from samples generated from a log-normal distribution and from the unweighted PCE for approximation orders $N = 4, 8, 10$.

Table 4.1 summarizes the minimum, maximum, mean, standard deviation and probability of negative values estimated from the same samples.

Tables 4.2 and 4.3 summarizes the minimum, maximum, mean, standard deviation and probability of negative values estimated from the weighted expansions. Comparing to table 4.1, it can be seen that although the mean value is estimated well, the standard deviation, minimum and maximum are underestimated. The probability of negative values is zero, as the weights introduced result in positive approximations. As seen in Figures 4.1 and 4.2, the tail of the log-normal distribution is not well captured by the finite (unweighted and weighted) expansions in Legendre polynomials. Therefore, an estimate for

Table 4.1: Sample estimates for the unweighted expansions

	min	max	mean	st. deviation	$P(\Theta < 0)$	$P(\Theta > \max(\Theta_N))$
$N = 4$	-0.0808866	2.60528	0.5730354	0.6564249	0.0414	0.0466
$N = 8$	-0.1688893	4.388988	0.6169931	0.8563759	0.0173	0.0174
$N = 10$	-0.117983	5.67271	0.6173979	0.9049591	0.0097	0.0101
exact	0.002915366	30.97578	0.6647008	1.209795	0	

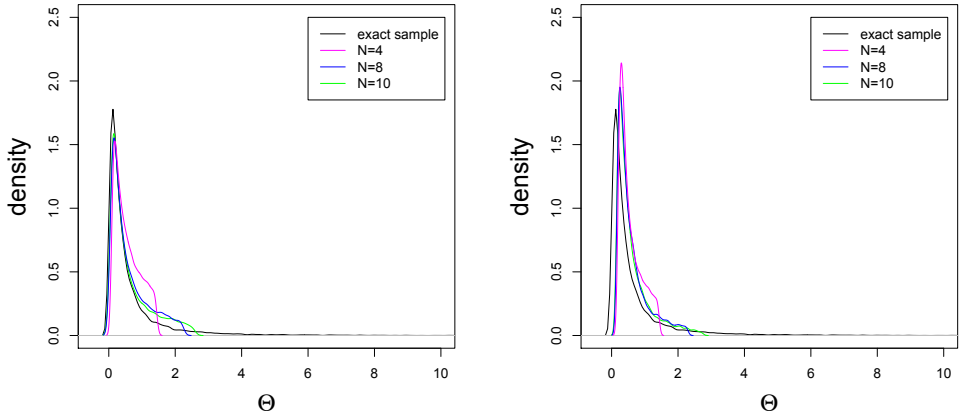

Figure 4.2: Density estimates from samples generated from a log-normal distribution and from the weighted PCE for approximation orders $N = 4, 8, 10$. For the left figure, the De la Vallée-Poussin weights were used and for the right figure the modified Fejér weights. In both cases, the basis polynomials used were the Legendre polynomials.

the probability of the log-normal distribution being greater than the maximum value of the truncated expansion is also reported in the last columns of all tables.

Assume now that the solution of the logistic equation with unknown initial condition represented by the random variable Θ, has the following PC expansion

$$x(t, \Xi) = \sum_{n=0}^{\infty} q_n(t) P_n(\Xi) h_n. \tag{4.61}$$

By substituting the truncated N-th order expansion of $x(t, \Xi)$ in the dynamics (4.56) and projecting the residual on the corresponding polynomial subspace, it follows that the coefficients $\{q_n(t)\}_{n=0,\dots,N}$ satisfy the following coupled differential equations

$$\dot{q}_n(t) = r q_n(t) - \frac{r}{K} \sum_{l=0}^{N} \sum_{m=0}^{N} q_l(t) q_m(t) \langle P_l P_m, P_n \rangle h_l h_m, \ n = 0, \dots, N, \tag{4.62}$$

with initial conditions

$$q_n(t_0) = \lambda_n, \ n = 0, \dots, N. \tag{4.63}$$

Table 4.2: Sample estimates for the weighted expansions (De la Vallée-Poussin)

	min	max	mean	st. deviation	$P(\Theta < 0)$	$P(\Theta > \max(\Theta_N))$
$N = 4$	0.09746752	1.443358	0.5712307	0.3795152	0	0.1091
$N = 8$	0.06042169	2.2542	0.6147099	0.5572159	0	0.0585
$N = 10$	0.05217915	2.611289	0.6317278	0.6202718	0	0.0462
exact	0.002915366	30.97578	0.6647008	1.209795	0	

Table 4.3: Sample estimates for the weighted expansions (modified Fejér)

	min	max	mean	st. deviation	$P(\Theta < 0)$	$P(\Theta > \max(\Theta_N))$
$N = 4$	0.1914131	1.42841	0.5709131	0.3297167	0	0.1108
$N = 8$	0.1667508	2.300286	0.6151546	0.4692217	0	0.0563
$N = 10$	0.1605935	2.74657	0.6305314	0.5162878	0	0.0425
exact	0.002915366	30.97578	0.6647008	1.209795	0	

The estimated coefficients for the log-normal random variable Θ correspond to random variables Θ_N which have small but non-zero probability of negative values, as seen in Table 4.1. Propagating these distributions through the dynamics results in a blow-up as seen in Figure 4.3. The mean value tends to $-\infty$ and the variance to $+\infty$ after approximately one-two units of time for the different approximation orders $N = 4, 8, 10$.

Assume now that the initial conditions of the Galerkin system are set to

$$q_n(t_0) = \omega_{N,n}\lambda_n,\ n = 0, \dots, N, \tag{4.64}$$

meaning that the distribution of the initial condition is strictly positive. In this case, the blow-up of the solution is avoided since a non-negative distribution is propagated through the dynamics. In Figures 4.4 and 4.5, the mean and the standard deviation of the solution over time are plotted. These are estimated through Monte Carlo integration and through weighted PC expansions for two different weight sequences.

The same behavior is demonstrated when employing a NISP approach. To this end, the PC coefficients of the solution were computed by numerically approximating their definition integrals

$$q_n(t) = \int_{-1}^{1} x(t, g_\Theta(\xi))P_n(\xi)d\mu(\xi) \approx \frac{1}{M}\sum_{m=1}^{M} x(t, g_\Theta(\xi^m))P_n(\xi^m),\ n = 0, \dots, N, \tag{4.65}$$

where $\{\Xi^m\}_{m=1}^M$ is a random sample from the uniform distribution on $[-1, 1]$. Again, if $g_\Theta(\xi^m) = \sum_{n=0}^{N} \lambda_n P_n(\xi^m)h_n$, corresponds to a negative realization of Θ, a blow-up of the solution of the logistic equation will occur in finite time and the PC coefficients will be undetermined after this time point. In Figure 4.6, the mean and the standard deviation of the solution are plotted when using an unweighted PC expansion for Θ for different approximation orders. In Figures 4.7 and 4.8, the initial conditions are sampled from a weighted PCE for Θ, where the De la Vallée-Poussin and the modified Fejér weights were used respectively.

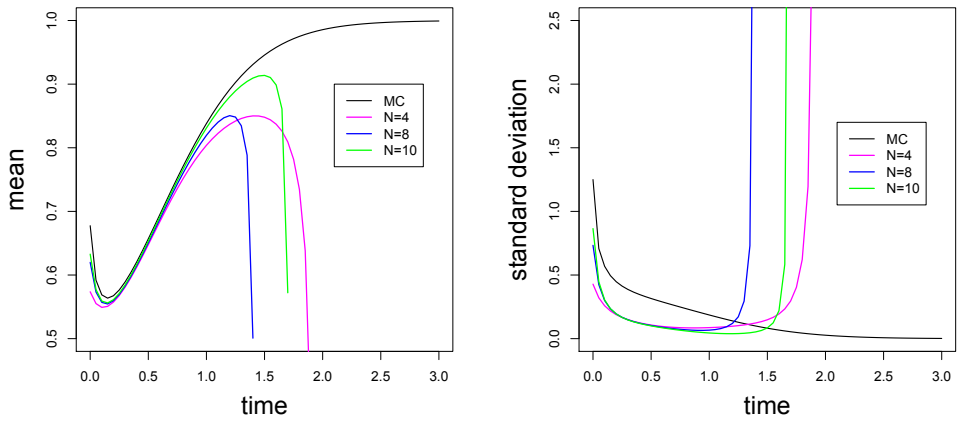

Figure 4.3: Mean and standard deviation over time estimated by Monte Carlo integration (black line) and via the PC expansion of the solution for approximation orders $N = 4, 8, 10$. The PC coefficients of the solution were computed via a Galerkin approach.

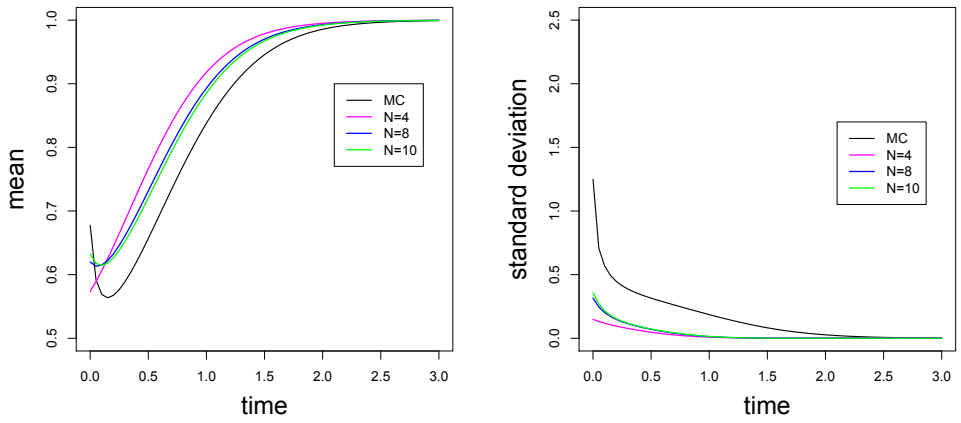

Figure 4.4: Mean and standard deviation over time estimated by Monte Carlo integration and via the Galerkin method for approximation orders $N = 4, 8, 10$. The De la Vallée-Poussin weights were used for the PC expansion of the initial condition.

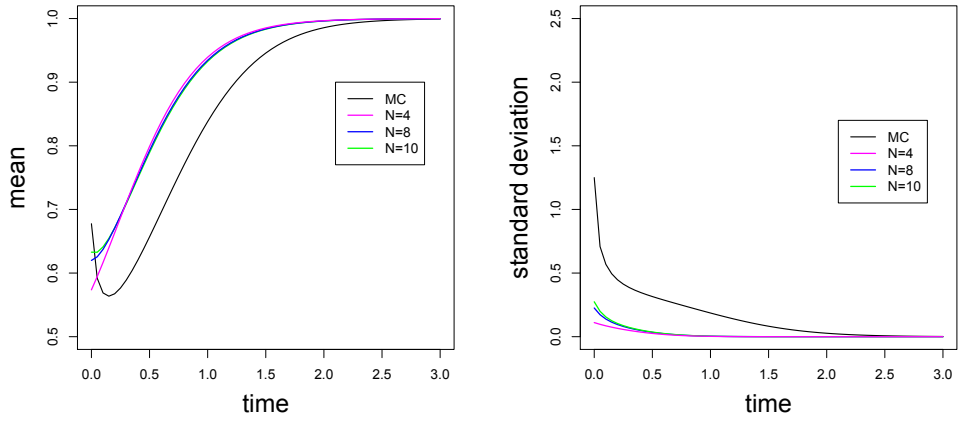

Figure 4.5: Mean and standard deviation over time estimated by Monte Carlo integration (black line) and via the Galerkin method for approximation orders $N = 4, 8, 10$. The modified Fejér weights were used for the PC expansion of the initial condition.

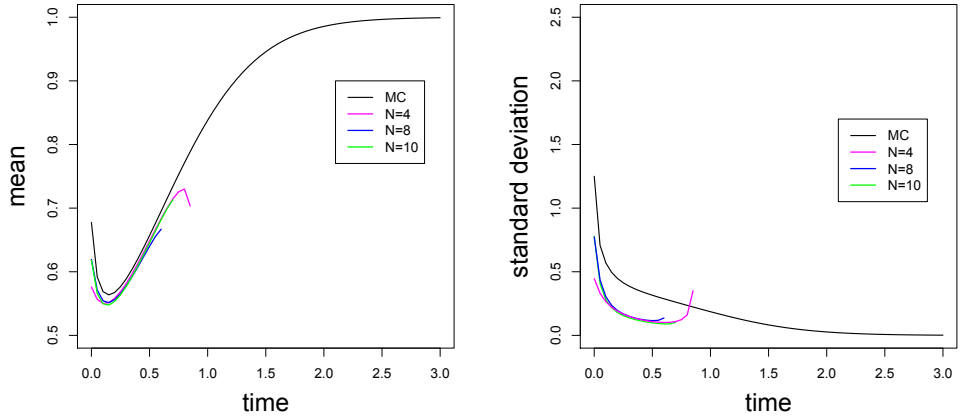

Figure 4.6: Mean and standard deviation over time estimated by Monte Carlo integration and via a NISP method for approximation orders $N = 4, 8, 10$.

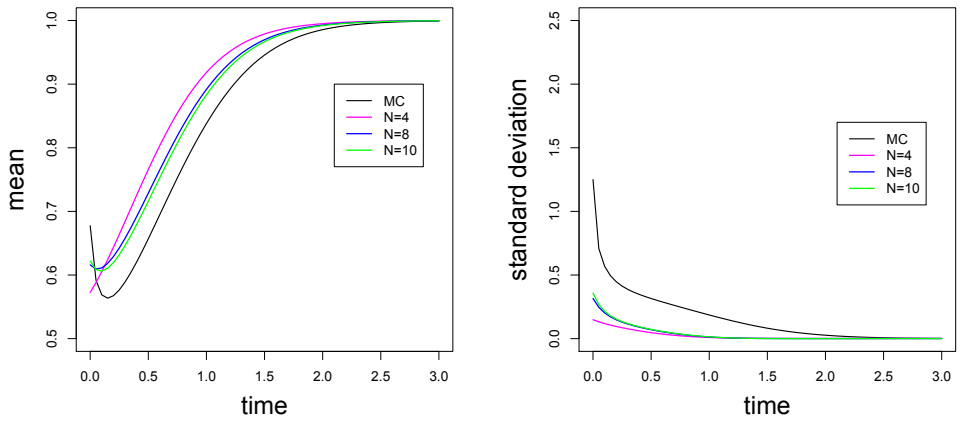

Figure 4.7: Mean and standard deviation over time estimated by Monte Carlo integration (black line) and via a NISP method for approximation orders $N = 4, 8, 10$. The De la Vallée-Poussin weights were used in the PC expansion of the initial condition.

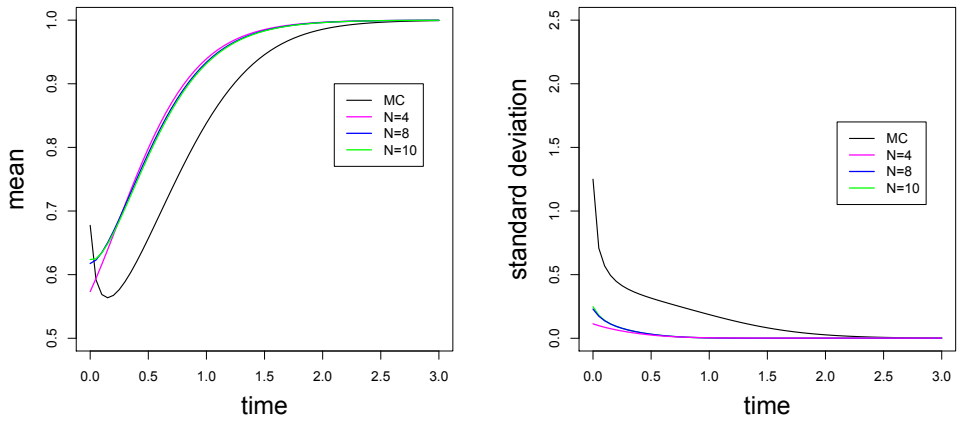

Figure 4.8: Mean and standard deviation over time estimated by Monte Carlo integration and via a NISP method for approximation orders $N = 4, 8, 10$. The modified Féjer weights were used in the PC expansion of the initial condition.

5 Real-time optimal control of the euglycemic hyperinsulinemic clamp (EHC) on mice

The euglucemic hyperinsulinemic clamp (EHC) is the gold standard test for quantifying insulin resistance in diabetes research. In this chapter, a method is proposed for the optimization of this experiment on mice. It is based on a Bayesian framework that combines sequential Monte Carlo methods for parameter inference with model predictive control for the underlying optimization problem. Two different ways are considered for the optimization: a sample-average and a polynomial chaos based approach.

In the next section, an introduction to the biological problem is given. A model for the glucose-insulin regulatory system in the special situation of the clamp test is presented and its performance is evaluated with comparison to real data. After a short introduction to Bayesian methods for parameter inference, the two algorithms for the real-time control of the linear non-autonomous system with parametric uncertainty are stated. The control and inference steps are separated and studied in more detail. The performance of the methods is evaluated through numerical simulations.

5.1 Biological background

Glucose is used as an energy source by living organisms and normal blood glucose concentrations are vital for the proper functioning of the body. The glucose levels can be affected by various factors such as food intake or exercise. The term *hyperglycemia* is used to describe the condition of elevated blood glucose concentration values. It can lead to destruction of the blood vessels, retinopathy, nephropathy and even more severe problems, as glucose is highly toxic. When glucose levels drop to levels lower than the physiological ones, the body cannot keep the organs functioning. This situation is referred to as *hypoglycemia* and can even lead to a coma. The endocrine hormones glucagon and insulin, which are produced by the α- and β-cells of the pancreas respectively, are responsible for keeping the glucose concentration at physiological levels. In healthy individuals, the glucose-insulin system is an extremely robust control system. In Figure 5.1 a schematic representation of this system is given.

Diabetes mellitus is a metabolic disease characterized by a disorder of this regulatory system and is divided into two categories. *Type-I* diabetes or insulin-dependent diabetes occurs when the β-cells produce little or no amounts of insulin and is the cause of about 5-10% of the diabetic cases. On the other hand, *type-II* diabetes or insulin-independent diabetes characterizes the rest 90-95% of the cases and is related to *insulin resistance*, the situation occurring when the body cells are not responding adequately to the action of insulin and are unable to absorb the hormone and metabolize the glucose in the blood. Under this

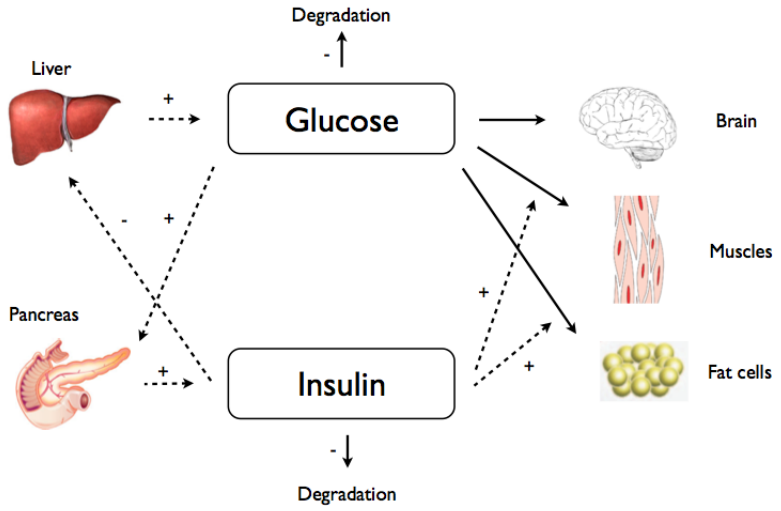

Figure 5.1: The glucose-insulin regulatory system. Glucose is produced by the liver and utilized by the brain, the muscles and fat cells. It stimulates the production of insulin by the pancreas, which in turn suppresses the liver glucose production.

condition, the liver fails to detect elevated glucose concentrations, glycogen is converted to glucose and the result is a combination of high glucose and insulin blood concentrations. According to data from the World Health Organization (WHO), diabetes is expected to rise from the eleventh to the sixth place of most common diseases in the world by the year 2030 [127]. These numbers emphasize the need and importance of the development of new effective diagnostic, preventive, and therapeutic methods for diabetes.

A variety of tests exist for the detection of hyperglycemia with most common the *intravenous glucose tolerance* test (IVGTT) and the *oral glucose tolerance* test (OGTT). These are easy to conduct and are routinely used to classify subjects as insulin-resistant or diabetic. In [42] two further tests were developed: the *hyperglycemic* and the *euglucemic hyperinsulinemic clamp* test. The idea behind them and their big advantage is to break the glucose-insulin feedback loop and place the system under the control of the experimentalist. In the first one, glucose is raised to high non-physiological values and kept constant (*clamped*) at those values by a variable glucose infusion. The test quantifies the amount of insulin secretion and is thus related to type-I diabetes. In the latter one, the goal is to keep the glucose level constant at physiological levels with a variable glucose infusion while insulin concentration is kept fixed in high levels by a continuous infusion of insulin. It is related to type-II diabetes and it is used to determine the *insulin-sensitivity index* (SI) of an individual. This corresponds to the amount of glucose infused in the last twenty minutes of the test, under *steady-state* conditions, i.e. under conditions where the glucose concentration remains constant. A low SI value characterizes the individual as insulin-resistant (or pre-diabetic). The clamp tests are expensive and difficult to execute and are mainly employed in studies targeted at drug development and at the reveal of the causes of diabetes.

In this thesis, only the EHC test on mice will be studied. Figure 5.2 describes how the EHC is conducted. Mice are fasted over night or for shorter periods to avoid the appear-

ance of meal-related glucose to the plasma. A preparation period of two hours (from -120 [min] to 0 [min]) precedes the actual clamp experiment. During this time, a solution of [3-^{3}H]-glucose is given to the mice and its concentration equilibrates in this period in all muscles and tissues. This radioactive glucose is used to quantify the endogenous glucose production (EGP) and the rate of disappearance of glucose (Rd). At the end of the preparatory period, blood samples are taken to estimate physiological values of glucose and insulin concentration. The actual clamp begins at 0 [min] with the injection of an insulin bolus, i.e. a high dose of insulin administered in a very short time. Its role is to shut down the EGP and stimulate the utilization of glucose by the muscles and tissues. After that and until the end of the experiment, insulin concentration is held constant at high non-physiological levels by a continuous constant insulin infusion. At the same time, a glucose solution is infused. Until the end of the experiment, blood samples are taken at regular times and the glucose infusion rate is adjusted accordingly to maintain a reference glucose concentration level. A bolus of 2 [^{14}C]-deoxy-glucose, i.e. another type of radioactive glucose, can be also administered during the second hour of the test in order to quantify the glucose uptake by specific tissues. More information on this test run on mice can be found in [6,7].

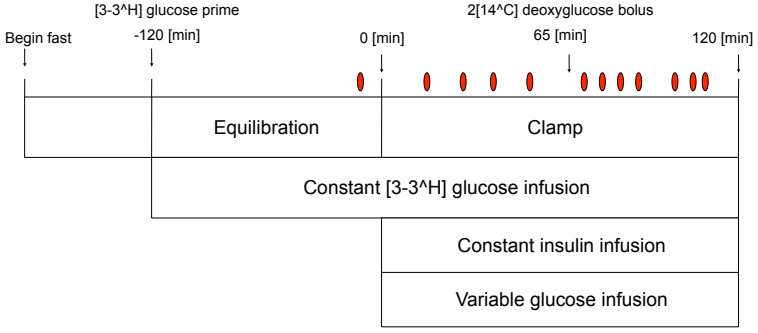

Figure 5.2: The euglycemic hyperinsulinemic clamp.

The basic problem experimentalists are facing is that there is no model-based strategy to determine the glucose infusion rate in order to maintain euglycemia throughout the test. Until now this was accomplished based on experience and intuition. The achievement of the steady-state conditions is crucial for the interpretation of the experiment: only after these conditions are achieved, one can extract useful information from the resulting data and make comparisons between different individuals. Clamps running on humans or even on rats are relative easier to run comparing to clamps on mice because the underlying system is more stable and one can rely on a large number of blood samples to monitor close the changes in glucose concentration and thus react quickly and efficiently to deviations from the desired reference level. Mice in contrary are very sensitive and large fluctuations occur around the reference concentration. These are caused mainly by the impact of stress on the glucose-insulin system, see for example in [7]. In addition, as the volume of blood in mice is rather limited, the number of measurements is accordingly small. These problems have to be taken under consideration in the development of the optimization algorithm. Despite the difficulties mentioned, mice are preferred for example when the relation of

gene mutations to diabetes are examined.

Next, the data and the developed model are presented.

5.2 Modeling the glucose dynamics during the EHC in mice

There exists a vast amount of literature on the modeling of the glucose-insulin system. See in [89–91] for reviews on the topic. One of the first models is the one by Bolie [24]. The most widely accepted model is the *minimal model* first proposed in [13], named like that because of its simple form and its low number of parameters compared to previous existing models. This model was developed to explain data obtained by the IVGTT. Although suitable for some situations, it fails to explain other tests and long term dynamics of the glucose-insulin system, as it was shown in [43]. Still, it is mainly used in clinical tools to extract information about the underlying physiological system [14, 15].

Particularly for the EHC test, a deterministic model was proposed in [103] based on data from long-duration clamps run on humans. This model takes into account the underlying physiology and has a complicated form including delay and integral terms and ten parameters which have to be estimated from data. The deterministic version was followed by a stochastic extension in [104], in which one of the parameters was assumed to depend linearly on a Wiener process.

In contrast with the aforementioned approach, the main goal here is not to explain the underlying physiology in a detailed way but to control the system and optimize the experiment. For this purpose, a simple model in the form of an ordinary differential equation was developed. As the number of available data was also relative small, the number of parameters had to be also reduced in comparison with the model in [103].

5.2.1 Data description

Two groups of mice were considered. The control group included 13 individuals and the case group included 12 individuals. The weight of the mice before and after fasting along with fasting hours were reported. From $0-3$ [min], an insulin bolus was injected whose concentration depends on the weight of each mouse. Its concentration in insuslin was given. From $3-120$ [min], insulin was infused at a given constant rate. From -120 [min] to 120 [min], the [3-^{3}H]-glucose was administered in two different solution concentrations before and after 0 [min]. At 65 [min], a bolus of 2[^{14}C]-deoxy-glucose was injected. This bolus is given from the same catheter that is used for the injection of insulin, glucose and [3-^{3}H]-glucose. This means that at this time an unspecified amount of these quantities is given to the subject and that after the labeled glucose is given, the subject does not receive any insulin, glucose and [3-^{3}H]-glucose until their flow in the catheter is restored. This procedure causes stress to the mouse and great fluctuations in the glucose concentrations around this time point. Furthermore, the concentrations of the injected glucose solution along with reference and physiological glucose levels were reported. Data were given on measured plasma glucose concentrations at times $-10, 10, 30, 45, 60, 67.5, 70, 75, 80, 90, 100, 110$ [min] and on the actual glucose infusion rates and their times of change. Note that no data are given on insulin concentrations throughout the experiment. This makes for example impossible to fit the minimal model since this is fitted in a decoupled manner on glucose and

insulin data [101].

5.2.2 The model

Denote the time by t and by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_6)$ the unknown vector of parameters of the model. $u(t)$ stands for the glucose infusion rate and $x(t, \boldsymbol{\theta}, u)$ for the plasma glucose concentration. The initial time point is $t_0 = -10$ [min] and the end point of the experiment is $T = 110$ [min]. These have been rescaled to $t_0 = 0$ [min] and $T = 120$ [min]. Assume that the glucose dynamics evolve during the clamp as follows

$$\begin{aligned} \dot{x}(t, \boldsymbol{\theta}, u) &= -(\theta_1 + \theta_2 t^2 e^{-\theta_3 t} + \theta_4 e^{\theta_5 t}) x(t, \boldsymbol{\theta}, u) + \frac{u(t)}{V_G}, \\ x(t_0, \boldsymbol{\theta}, u) &= \theta_6. \end{aligned} \tag{5.1}$$

The initial value is considered unknown and has to be also estimated. The parameter V_G represents the plasma volume of the mice per kilogram of body weight and its value was taken by the literature to be 0.49 [ml/kgBW] [1]. Note that no equation is considered for the dynamics of the insulin concentration. The action of insulin on the change of the glucose dynamics is included in the degradation rate

$$\gamma(t, \theta_1, \ldots, \theta_5) = \theta_1 + \theta_2 t^2 e^{-\theta_3 t} + \theta_4 e^{\theta_5 t}. \tag{5.2}$$

This function summarizes all possible reasons for appearance or disappearance of glucose

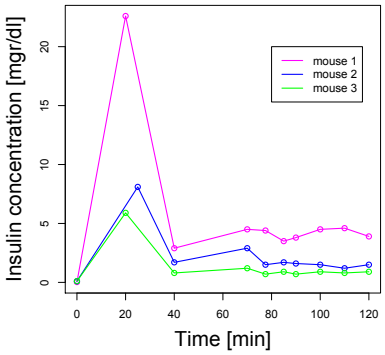

Figure 5.3: Data on insulin concentration

in the plasma. It thus includes the natural degradation rate of glucose, the disappearance of glucose from the blood due to the action of insulin, the insulin-independent absorption of glucose (for example by the brain, see Figure 5.2) as well as the endogenous glucose production. In Figure 5.3, data on insulin concentrations for three different individuals during a clamp test are plotted. As it can be seen, insulin values raise in the beginning of the experiment due to the action of the insulin bolus and then remain approximately constant. The elevated insulin levels in this phase, i.e. the pick of the insulin seen in Figure 5.3, are reflected to the bell shaped form of the degradation rate function during this period curve.

In Figures 5.4, 5.5 and 5.6, the glucose concentration data of three individuals and the fitted model are plotted. The value for the reference glucemia for all the subjects was $x_{\text{ref}} = 150$ [mg/dl]. Furthermore, the given glucose infusion rates and the estimated degradation rates are given for comparison. As it can be seen in these figures, the experimentalists fail to clamp the glucose concentrations around the reference value, thus explaining the need for an algorithmic approach to the problem. In each figure, the first subfigure includes the given glucose concentration data (circles) as well as the fitted model (black line). The dashed line corresponds to the reference glucose level. The second subfigure shows the estimated degradation rate and the third one, the given glucose infusion rate applied during the test.

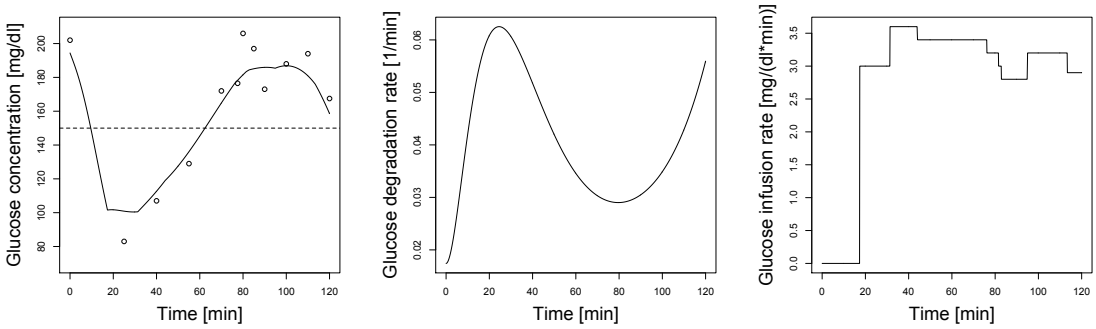

Figure 5.4: Glucose concentration, degradation rate and infusion rate for Mouse Di20.

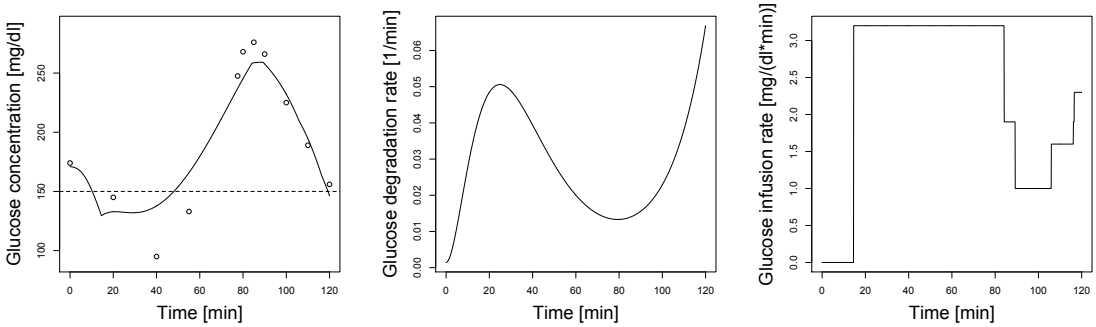

Figure 5.5: Glucose concentration, degradation rate and infusion rate for Mouse Di22.

Table 5.1 summarizes the median along with minimum and maximum values of the fitted parameters for the two groups. The parameters were estimated by a least squares fitting.

In order for the developed model to be adequate for control, it has to be able to predict the data. This is depicted in Figure 5.7. Here, the parameters are estimated from the three first data points and 95% estimation and prediction intervals are plotted around the data. It can be seen, that the data in general lie in these prediction intervals. This suggest that the model is suitable for the control.

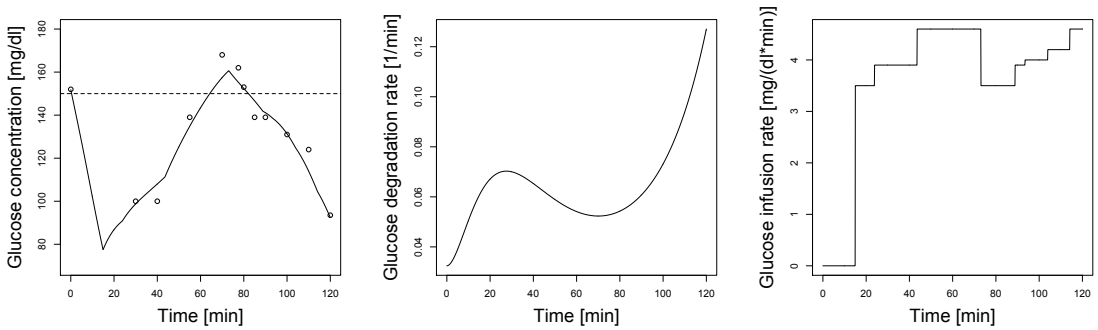

Figure 5.6: Glucose concentration, degradation rate and infusion rate for Mouse SD14.

Table 5.1: Estimated parameters for control and case subjects.

	Controls (13 subjects)	Cases (12 subjects)
θ_1	0.015 [0.0014, 0.038]	0.012 [0.0004, 0.026]
θ_2	0.0003 [0.00014, 0.00044]	0.0003 [0.00015, 0.0006]
θ_3	0.055 [0.054, 0.068]	0.065 [0.055, 0.085]
θ_4	0.0004 [0.00005, 0.00067]	0.00031 [0, 0.00075]
θ_5	0.033 [0.008, 0.05]	0.033 [0, 0.066]
θ_6	159 [140, 223]	173 [121, 215]

5.3 Bayesian methods for parameter inference

A short introduction of Bayesian methods is given here. More details can be found for example in [16, 57, 75].

Let $\boldsymbol{x} \in \mathbb{R}^n$ be a quantity of interest which depends on an unknown parameter vector $\boldsymbol{\theta} \in \mathbb{R}^d$. Furthermore, assume that $\boldsymbol{x}$ can be observed and that k independent noisy measurements $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k$ are available. Note that in some cases only some components of the vector $\boldsymbol{x}$ can be observed or even functions of its components. The presentation is here restricted to the full observable case for simplicity.

In contrast with the frequentist approach, in a Bayesian setting unknown parameters and observed data are considered as random variables. Therefore, according to the notation used in this manuscript, they will be denoted from now on with capital letters and the small letters will be reserved for their realizations. It is also assumed here for simplicity that these random variables are continuous, and thus their measures admit densities with respect to the Lebesgue measure.

Let $(\Omega, \mathcal{A}, P)$ be an abstract probability space and define on this space the random variables $\boldsymbol{\Theta}$ and $\boldsymbol{Y}$ for the parameters and the data under consideration respectively. The available information on the parameters prior to the experiment based for example on previous experiments or on physical constraints, is summarized into a probability distribution, which is called the *prior* distribution of $\boldsymbol{\Theta}$. Its density will be denoted as $p_{\boldsymbol{\Theta}}$. Its construction is an important step for the method and different approaches exist. The existence of a prior distribution has been the main point of criticism of Bayesian methods,

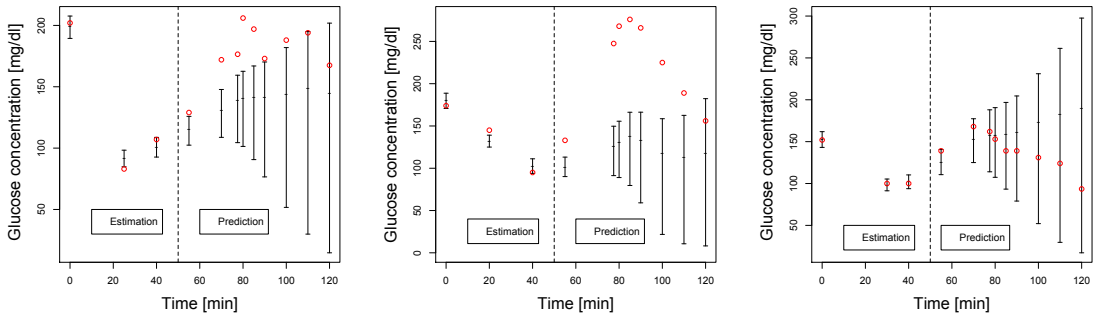

Figure 5.7: Prediction for the mice Di20, Di22 and SD14. The parameters are estimated from the first three data points and 95% estimation and prediction intervals are plotted around the given data (red circles).

as it introduces subjectivity to the method. Jeffreys defined *non-informative* priors to overcome this obstacle [73, 74]. Details are here omitted and the reader is referred to [76, 108]. In the context described here, the information available from previous experiment is useful as it restricts the domain where the parameters of the model are meaningful and thus helps also the control procedure. The next step is to construct the *likelihood* function $\mathcal{L}(\boldsymbol{\theta})$. This is defined as the density of the conditional probability of the observed data given the parameters

$$\mathcal{L}(\boldsymbol{\theta}) = \ell(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k \mid \boldsymbol{\Theta} = \boldsymbol{\theta}). \tag{5.3}$$

Note that the likelihood is considered as a function of the parameter vector although it depends also on the observations. The above are combined through Bayes' theorem which is next stated in terms of probabilities of events.

Theorem 5.1. *(Bayes) Let A_1, A_2 be subsets of the σ-algebra $\mathcal{A}$ with $P(A_2) > 0$. Then,*

$$P(A_1 \mid A_2) = \frac{P(A_2 \mid A_1)P(A_1)}{P(A_2)}. \tag{5.4}$$

Its continuous version is used to update the prior information on the parameter vector in the presence of data. The *posterior* distribution of the parameters is defined as the conditional probability of the parameters given the data and its density is given according to Bayes' theorem by

$$\pi_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) = \pi_{\boldsymbol{\Theta}}(\boldsymbol{\theta} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_k) = \frac{\ell(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k \mid \boldsymbol{\theta}) p_{\boldsymbol{\Theta}}(\boldsymbol{\theta})}{\int_{\mathbb{R}^d} \ell(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k \mid \boldsymbol{\theta}) p_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) d\boldsymbol{\theta}}. \tag{5.5}$$

The denominator in this expression is a normalizing constant (that does not depend on $\boldsymbol{\theta}$) so that $\pi_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$ is indeed a density. It is called the *evidence* or *marginal likelihood* and will be denoted by β. Thus, (5.5) can be written as

$$\pi_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) = \frac{\ell(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k \mid \boldsymbol{\theta}) p_{\boldsymbol{\Theta}}(\boldsymbol{\theta})}{\beta} = \frac{\tilde{\pi}_{\boldsymbol{\Theta}}(\boldsymbol{\theta})}{\beta} \propto \tilde{\pi}_{\boldsymbol{\Theta}}(\boldsymbol{\theta}), \tag{5.6}$$

where $\tilde{\pi}_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) = \ell(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k \mid \boldsymbol{\theta}) p_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$. In contrast with the classical statistical methods, in Bayesian theory the parameter estimation is based on a whole distribution and not only on a point estimate.

5.3.1 Monte Carlo methods

The information contained in the posterior distribution can be summarized in many ways. One can compute for example the posterior mean or the *maximum a posteriori (MAP) estimate*, i.e. the value(s) of $\boldsymbol{\theta}$ which maximize the posterior density. Measures of dispersion such as the posterior covariance as well as the construction of α*-credible regions* may be also used, i.e. subsets C of $\mathbb{R}^d$ such that $\int_C \pi_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) d\boldsymbol{\theta} = 1 - \alpha$. The computations of these quantities requires usually the evaluation of integrals with respect to the posterior density of the following form

$$E_{\boldsymbol{\Theta}}[h] = \int_{\mathbb{R}^d} h(\boldsymbol{\theta}) \pi_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{5.7}$$

where $h \in L^1(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \pi_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) d\boldsymbol{\theta})$. In high dimensional problems, this integration can be intractable with usual deterministic methods like numerical quadrature. Monte Carlo (MC) methods are a class of methods developed for the efficient numerical evaluation of these integrals. The idea is to approximate the measure $\boldsymbol{\mu}_{\boldsymbol{\Theta}}$ with density the posterior density by a discrete empirical measure $\boldsymbol{\mu}_{\boldsymbol{\Theta},N}$ based on a sample from the distribution of $\boldsymbol{\Theta}$. By the law of large numbers, if $\{\boldsymbol{\theta}^m\}_{m=1}^M$ is an i.i.d sample from the posterior of $\boldsymbol{\Theta}$, then the integrals in (5.7) can be approximated by

$$E_{\boldsymbol{\Theta}}[h] \approx \frac{1}{M} \sum_{m=1}^{M} h(\boldsymbol{\theta}^m). \tag{5.8}$$

The problem now arises of how to sample from a posterior distribution of complicated form. A class of methods in this direction are the Markov chain Monte Carlo methods which are next briefly reviewed.

5.3.2 Markov chain Monte Carlo (MCMC) methods

These sampling techniques are based on the construction of a *Markov* stochastic process $\{\boldsymbol{\Theta}_n\}_{n \in \mathbb{N}_0}$, i.e. a sequence of random vectors on $\mathbb{R}^d$ with the property that its value in the future depends only on the values of the current random vector and not on the past values. Mathematically, this is formulated as

$$P(\boldsymbol{\Theta}_{n+1} \in A \mid \boldsymbol{\Theta}_n = \boldsymbol{\theta}, \boldsymbol{\Theta}_{n-1} \in A_{n-1}, \ldots, \boldsymbol{\Theta}_0 \in A_0) = P(\boldsymbol{\Theta}_{n+1} \in A \mid \boldsymbol{\Theta}_n = \boldsymbol{\theta}), \tag{5.9}$$

for all $n \in \mathbb{N}_0$, all events $A, A_{n-1}, \ldots, A_0 \in \mathcal{B}(\mathbb{R}^d)$ and all $\boldsymbol{\theta} \in \mathbb{R}^d$. When the above probabilities do not depend on the index n, the Markov process is called *homogenous*. In this case, one can define the transition distribution $P(\boldsymbol{\theta}, A)$ of the process as follows.

Definition 5.2. *A function* $P \colon \mathbb{R}^n \times \mathcal{B}(\mathbb{R}^d) \in [0, 1]$ *is called a transition distribution or function if it has the following properties*

(i) for all $\boldsymbol{\theta} \in \mathbb{R}^d$*, the function* $P(\boldsymbol{\theta}, \cdot)$ *is a probability distribution, and*

(ii) for all $A \in \mathcal{B}(\mathbb{R}^d)$, the function $P(\cdot, A)$ is measurable.

One can then define the *transition kernel* $p(\boldsymbol{\theta}, \boldsymbol{\phi})$ as the density of the transition distribution.

A probability measure $\boldsymbol{\mu}$ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is called an *invariant measure* of the function $P(\boldsymbol{\theta}, A)$ if

$$\boldsymbol{\mu}(A) = \int_{\mathbb{R}^d} P(\boldsymbol{\theta}, A) d\boldsymbol{\mu}(\boldsymbol{\theta}). \tag{5.10}$$

If the measure $\boldsymbol{\mu}$ and the transition function admit densities π and p respectively, then (5.10) can be written as

$$\pi(\boldsymbol{\theta}) = \int_{\mathbb{R}^d} p(\boldsymbol{\theta}, \boldsymbol{\phi}) \pi(\boldsymbol{\phi}) d\boldsymbol{\phi}, \tag{5.11}$$

and then π is the invariant density for the kernel p. In Bayesian theory, MCMC methods provide with a sample from the posterior distribution $\pi_{\boldsymbol{\Theta}}$ by constructing a Markov process which admits this posterior distribution as its invariant measure. The mathematical justification of these methods is based on the Ergodic theorem. More details can be found for example in [62, 66].

The Metropolis-Hastings (MH) algorithm

A special class of methods for constructing a transition kernel with the desired properties are the Metropolis-Hastings algorithms. A nice introduction in this type of algorithms is given in [33]. Here, a basic version is considered. Let $q(\boldsymbol{\theta}, \boldsymbol{\phi})\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a function such that $\int_{\mathbb{R}^d} q(\boldsymbol{\theta}, \boldsymbol{\phi}) d\boldsymbol{\phi} = 1$. It can be shown that a density π is invariant for the transition function given by

$$P(\boldsymbol{\theta}, A) = \int_A q(\boldsymbol{\theta}, \boldsymbol{\phi}) \alpha(\boldsymbol{\theta}, \boldsymbol{\phi}) d\boldsymbol{\phi} + \mathbb{1}\{\boldsymbol{\theta} \in A\} \left(1 - \int_A q(\boldsymbol{\theta}, \boldsymbol{\phi}) \alpha(\boldsymbol{\theta}, \boldsymbol{\phi}) d\boldsymbol{\phi}\right), \tag{5.12}$$

where

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\phi}) = \begin{cases} \min\left(1, \frac{\pi(\boldsymbol{\phi}) q(\boldsymbol{\phi}, \boldsymbol{\theta})}{\pi(\boldsymbol{\theta}) q(\boldsymbol{\theta}, \boldsymbol{\phi})}\right), & \text{if } \pi(\boldsymbol{\theta}) q(\boldsymbol{\theta}, \boldsymbol{\phi}) > 0, \\ 1, & \text{else .} \end{cases} \tag{5.13}$$

In practice, the algorithm is implemented as follows

(i) Choose an arbitrary starting value $\boldsymbol{\theta}^{(0)}$ and set the index $m = 1$.

(ii) Generate a new value $\boldsymbol{\phi}$ from the density $q(\boldsymbol{\theta}^{(m-1)}, \cdot)$.

(iii) Evaluate the acceptance probability $\alpha(\boldsymbol{\theta}^{(m-1)}, \boldsymbol{\phi})$ and generate $u \sim U[0, 1]$. If $u < \alpha(\boldsymbol{\theta}^{(m-1)}, \boldsymbol{\phi})$, set $\boldsymbol{\theta}^{(m)} = \boldsymbol{\phi}$, else $\boldsymbol{\theta}^{(m)} = \boldsymbol{\theta}^{(m-1)}$.

(iv) Update the index $m \to m + 1$ and return to step (ii) until convergence is reached.

5.3.3 Sequential Monte Carlo (SMC) methods

The material in this section is based on [37, 44, 49]. When data are coming sequentially, MCMC methods are inadequate to solve the inference problem as they require to run a long chain for each new data point and they do not take into account previous computations. Sequential Monte Carlo methods or particle filters provide a faster alternative to MCMC and can be used for real-time parameter estimation. They are based on the following observation. Assume that after k measurements are given, the posterior distribution of $\boldsymbol{\Theta}$ is $\pi^k_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$. When the next $k+1$-measurement arrives, one has

$$\begin{aligned}\pi^{k+1}_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) &\propto \ell(\boldsymbol{y}_1,\dots,\boldsymbol{y}_k,\boldsymbol{y}_{k+1} \mid \boldsymbol{\theta}) p_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \\ &\propto \ell(\boldsymbol{y}_{k+1} \mid \boldsymbol{y}_1,\dots,\boldsymbol{y}_k,\boldsymbol{\theta}) \ell(\boldsymbol{y}_1,\dots,\boldsymbol{y}_k \mid \boldsymbol{\theta}) p_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \\ &\propto \ell(\boldsymbol{y}_{k+1} \mid \boldsymbol{\theta}) \pi^k_{\boldsymbol{\Theta}}(\boldsymbol{\theta}),\end{aligned} \tag{5.14}$$

which means that the posterior density at step k can be used as the prior for the next iteration.

SMC techniques are based on updating the weights of a discrete empirical measure $\{(\boldsymbol{\theta}^m, \frac{1}{M})\}_{m=1}^M$ from the prior distribution $p_{\boldsymbol{\Theta}}$ by a re-weighting procedure to produce a discrete measure converging for $M \to \infty$ to $\pi^k_{\boldsymbol{\Theta}}, \forall\, k = 1,\dots,K$. In other words, a weighted sample (or a set of *weighted particles*) $\{(\boldsymbol{\theta}^m, w^k(\boldsymbol{\theta}^m))\}_{m=1}^M$ is carried over time and the weights $\{w^k(\boldsymbol{\theta}^m)\}_{m=1}^M$ are updated at each iteration. This procedure relies on the *importance sampling* method, which is next presented.

Assume that integrals with respect to the posterior as in (5.7) are to be computed. When the posterior distribution is difficult to sample form, one can use an auxiliary measure to transform the integrals. Let $\boldsymbol{\nu}$ be a probability measure with density η, which is called the *importance density* such that $\eta(\boldsymbol{\theta}) > 0$, for all $\boldsymbol{\theta}$ in the support of the posterior measure. Then, by using equation (5.6), equation (5.7) can be written as

$$\begin{aligned}E_{\boldsymbol{\Theta}}[h] &= \int_{\mathbb{R}^d} h(\boldsymbol{\theta}) \pi_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int_{\mathbb{R}^d} h(\boldsymbol{\theta}) \frac{\tilde{\pi}_{\boldsymbol{\Theta}}(\boldsymbol{\theta})}{\beta} d\boldsymbol{\theta} \\ &= \frac{1}{\beta} \int_{\mathbb{R}^d} h(\boldsymbol{\theta}) \frac{\tilde{\pi}_{\boldsymbol{\Theta}}(\boldsymbol{\theta})}{\eta(\boldsymbol{\theta})} \eta(\boldsymbol{\theta}) d\boldsymbol{\theta} = \frac{1}{\beta} \int_{\mathbb{R}^d} h(\boldsymbol{\theta}) w(\boldsymbol{\theta}) \eta(\boldsymbol{\theta}) d\boldsymbol{\theta},\end{aligned} \tag{5.15}$$

where the weight function

$$w(\boldsymbol{\theta}) = \frac{\tilde{\pi}_{\boldsymbol{\Theta}}(\boldsymbol{\theta})}{\eta(\boldsymbol{\theta})} \tag{5.16}$$

is introduced. In the same way, one can rewrite the integral defining the evidence β as

$$\beta = \int_{\mathbb{R}^d} \tilde{\pi}_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int_{\mathbb{R}^d} \frac{\tilde{\pi}_{\boldsymbol{\Theta}}(\boldsymbol{\theta})}{\eta(\boldsymbol{\theta})} \eta(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int_{\mathbb{R}^d} w(\boldsymbol{\theta}) \eta(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{5.17}$$

Let now $\{\boldsymbol{\theta}^m\}_{m=1}^M$ be a random sample from $\boldsymbol{\nu}$ and denote by $\boldsymbol{\nu}_M$ the corresponding empirical measure. Then, one has

$$E_{\boldsymbol{\Theta}}[h] \approx \left(\frac{1}{M} \sum_{m=1}^M w(\boldsymbol{\theta}^m) \right)^{-1} \frac{1}{M} \sum_{m=1}^M h(\boldsymbol{\theta}^m) w(\boldsymbol{\theta}^m) = \sum_{m=1}^M \frac{w(\boldsymbol{\theta}^m)}{\sum_{m=1}^M w(\boldsymbol{\theta}^m)} h(\boldsymbol{\theta}^m). \tag{5.18}$$

Back to the sequential Bayesian context described above, assume one has a sample $\{\boldsymbol{\theta}^m\}_{m=1}^M$ to approximate $\pi_{\boldsymbol{\Theta}}^k$. To generate a set of particles from $\pi_{\boldsymbol{\Theta}}^{k+1}$, one uses as weight function

$$w^k(\boldsymbol{\theta}) = \frac{\pi_{\boldsymbol{\Theta}}^{k+1}(\boldsymbol{\theta})}{\pi_{\boldsymbol{\Theta}}^{k}(\boldsymbol{\theta})} \propto \frac{\tilde{\pi}_{\boldsymbol{\Theta}}^{k+1}(\boldsymbol{\theta})}{\tilde{\pi}_{\boldsymbol{\Theta}}^{k}(\boldsymbol{\theta})} \propto \ell(\boldsymbol{y}_{k+1} \mid \boldsymbol{\theta}). \tag{5.19}$$

This means, that at each iteration one has to multiply each $w^k(\boldsymbol{\theta})$ with the weight $w^{k+1}(\boldsymbol{\theta})$. This has as a consequence the reduction of the number of particles with significant weight at each step. This phenomenon is called *particle degeneracy*.

A resampling step may be used to eliminate samples with insignificant weight and save computational time. The most common algorithm for resampling is multinomial selection, i.e. at stage k, one chooses M particles from the set $\{\boldsymbol{\theta}^m\}_{m=1}^M$ with probability the corresponding normalized weights $\{\frac{w^k(\boldsymbol{\theta}^m)}{\sum_{m=1}^M w^k(\boldsymbol{\theta}^m)}\}_{m=1}^M$, see [64]. For other approaches, see for example in [87]. The improved method is called *sampling importance resampling (SIR)*. Note that resampling does not save from degeneracy.

The combination of SIR with a MH-step was proposed first in [63] to reduce degeneracy. One chooses a transition kernel and applies one MH-step on each particle and accepts or rejects the proposed value through the acceptance probability as in the classical MH-algorithm. In this way, new sample values will enter the set $\{\boldsymbol{\theta}^m\}_{m=1}^M$ and will have a signifiant weight so that particle degeneracy is reduced. The acceptance rate can be used as a measure of the rejuvenation of the sample. The choice of the kernel influences the efficiency of this step and different approaches exist.

All together, an SMC algorithm in the Bayesian context is as follows

(i) Generate a sample $\{\boldsymbol{\theta}^m\}_{m=1}^M$ from the prior, assign the weight $\frac{1}{M}$ to each particle and set the index $k = 1$.

(ii) Compute the weight $w^k(\boldsymbol{\theta}^m)$ given by (5.19) and update the weight of each particle by multiplying with the corresponding w^k. Normalize the new weights by dividing with the total weight.

(iii) Resample and assign the weight $\frac{1}{M}$ again to each particle.

(ii) Draw $\boldsymbol{\phi}^m \sim K^k(\boldsymbol{\theta}^m, \cdot)$, for $m = 1, \ldots, M$, where K^k is a kernel with stationary distribution $\boldsymbol{\pi}_{\boldsymbol{\Theta}}^k$ and accept or reject $\boldsymbol{\phi}^m$ by computing the acceptance probability $\alpha(\boldsymbol{\theta}^m, \boldsymbol{\phi}^m)$ as in (5.13).

(v) Renew the set $\{\boldsymbol{\theta}^m\}_{m=1}^M$ by including the accepted particles. All particles have again weight $\frac{1}{M}$. Set $k \to k+1$ and return to step (ii) until $k = K$.

5.4 Real-time parameter estimation and optimal control for the EHC

In this section, the first algorithm for the optimization of the test is presented. It is based on Monte Carlo approximation of the underlying expectations. First, the problem is formulated and then the algorithm is stated. Its individual steps are analyzed in sections 5.4.1 and 5.4.2.

Problem formulation Estimate the parameter vector $\boldsymbol{\Theta}$ and drive the glucose concentration to the reference state x_{ref} by minimizing the cost functional

$$J_0(u) = E_{\boldsymbol{\Theta}}[\int_{t_0}^{T} (x(t, \boldsymbol{\Theta}, u) - x_{\text{ref}})^2 dt] \tag{5.20}$$

over all piecewise constant control functions of the form

$$u(t) = \sum_{i=1}^{K} u_i \mathbb{1}\{(t_{i-1}, t_i]\}(t),\ t \in [t_0, T], \tag{5.21}$$

with $u(t) \in [0, u_{\text{max}}]$ and under the dynamic constraint

$$\begin{aligned} \dot{x}(t, \boldsymbol{\Theta}, u) &= -\gamma(t, \boldsymbol{\Theta}) x(t, \boldsymbol{\Theta}, u) + \frac{u(t)}{V_G},\ t \in [t_0, T], \\ x(t_0, \boldsymbol{\Theta}, u) &= x_0(\boldsymbol{\Theta}). \end{aligned} \tag{5.22}$$

Noisy data y_k are given sequentially at times t_k

$$y_k = x(t_k) + \varepsilon_k,\ \varepsilon_k \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma_M^2),\ k = 0, \ldots, K. \tag{5.23}$$

Here, σ_M^2 denotes the assumed known measurement error and u_{max} is an upper bound on the glucose infusion rate which is physiologically accepted and also satisfies the capability of the device used. The above formulation is based on the special experimental setup and the given protocol. According to it, the control (the glucose infusion rate) is changed after a new measurement is given and is a piecewise constant function, as seen in Figures 5.4, 5.5 and 5.6. The parameter vector is noted with a capital letter to emphasize the fact that it is now modeled as a random variable. Define for each $k = 0, \ldots, K-1$, the cost functionals

$$J_k(u) = E_{\boldsymbol{\Theta}}[\int_{t_k}^{T} (x(t, \boldsymbol{\Theta}, u) - x_{\text{ref}})^2 dt]. \tag{5.24}$$

The complete algorithm is now given.

Algorithm

- Set counter $k = 1$, initialize a prior distribution $p_{\boldsymbol{\Theta}}$ and generate an equally weighted sample $\{(\boldsymbol{\theta}^m, \frac{1}{M})\}_{m=1}^M$ from $p_{\boldsymbol{\Theta}}$. While $k < K$:

  A. Estimation

  - get at time t_{k-1} the measurement y_{k-1},
  - use Bayes theorem to obtain the posterior distribution $\pi_{\boldsymbol{\Theta}}^k$,
  - use an SMC step and update the sample to $\{(\boldsymbol{\theta}^m, w^k(\boldsymbol{\theta}^m))\}_{m=1}^M$,

  B. Control

  - compute $\bar{u} = (\bar{u}_k, \ldots, \bar{u}_K) = \text{argmin} J_{k-1}(u)$,
  - apply $\bar{u}_k$ to the system in the time interval $(t_{k-1}, t_k]$,

- set as new prior the posterior $\pi_{\boldsymbol{\Theta}}^k$, $k \to k+1$ and go to the estimation step.

5.4.1 Exact solution of the optimal control problem

It is here shown how the minimizer of $J_0(u)$ can be computed. The same procedure is valid for all $J_k(u), k = 1, \ldots, K-1$.

Assume a control function as in equation (5.21). The solution of the homogenous system corresponding to (5.22) for a fixed realization $\boldsymbol{\theta}$ of the parameter vector reads

$$x_h(t, \boldsymbol{\theta}) = \Phi(t, \boldsymbol{\theta})\Phi^{-1}(t_0, \boldsymbol{\theta})x_0(\boldsymbol{\theta}), \tag{5.25}$$

where

$$\Phi(t, \boldsymbol{\theta}) = exp\left(-\theta_1 t + \frac{\theta_2}{\theta_3^2}e^{-\theta_3 t}(t^2 + 2t + 2) - \frac{\theta_4}{\theta_5}e^{\theta_5 t}\right). \tag{5.26}$$

The solution of the non-homogenous differential equation is given by

$$x(t, \boldsymbol{\theta}, u) = x_h(t, \boldsymbol{\theta}) + \Phi(t, \boldsymbol{\theta}) \int_{t_0}^{t} \Phi^{-1}(s, \boldsymbol{\theta}) \frac{u(s)}{V_G} ds. \tag{5.27}$$

Substituting (5.21) in the last integral, one has

$$\int_{t_0}^{t} \Phi^{-1}(s, \boldsymbol{\theta}) \frac{u(s)}{V_G} ds = \frac{1}{V_G} \sum_{i < \max\{j\colon t_{j-1} < t\}} u_i \int_{t_{i-1}}^{\min(t, t_i)} \Phi^{-1}(s, \boldsymbol{\theta}) ds. \tag{5.28}$$

It is next shown that the minimization of the cost functional $J_0(u)$ in (5.20) over the assumed piecewise constant control functions leads to an equivalent quadratic programming problem of the following form

$$J_0(u) = u^T A u + Bu \to \min!_{\substack{u_i \in [0, u_{\max}], \\ i=1,\ldots,K}}, \; u = (u_1, \ldots, u_K)^T, \tag{5.29}$$

where $A \in \mathbb{R}^{K \times K}$, $B \in \mathbb{R}^{K \times 1}$. Firstly, by substituting the exact solution into the cost functional one obtains

$$\begin{aligned}
J_0(u) &= E_{\boldsymbol{\Theta}}\left[\int_{t_0}^{T} (x(t, \boldsymbol{\Theta}, u) - x_{\text{ref}})^2 dt\right] \qquad (5.30)\\
&= E_{\boldsymbol{\Theta}}\left[\int_{t_0}^{T} \left(x_h(t, \boldsymbol{\Theta}) + \Phi(t, \boldsymbol{\Theta}) \int_{t_0}^{t} \Phi^{-1}(s, \boldsymbol{\Theta}) \frac{u(s)}{V_G} ds - x_{\text{ref}}\right)^2 dt\right] \\
&= E_{\boldsymbol{\Theta}}\left[\int_{t_0}^{T} \left(x_h(t, \boldsymbol{\Theta}) + \frac{\Phi(t, \boldsymbol{\Theta})}{V_G} \sum_{i < \max\{j\colon t_{j-1} < t\}} u_i \int_{t_{i-1}}^{\min(t, t_i)} \Phi^{-1}(s, \boldsymbol{\Theta}) ds - x_{\text{ref}}\right)^2 dt\right] \\
&= E_{\boldsymbol{\Theta}}\left[\int_{t_0}^{T} \left(x_h(t, \boldsymbol{\Theta}) + \frac{\Phi(t, \boldsymbol{\Theta})}{V_G} \sum_{i=1}^{K} u_i \mathbb{1}\{t > t_{i-1}\} \int_{t_{i-1}}^{\min(t, t_i)} \Phi^{-1}(s, \boldsymbol{\Theta}) ds - x_{\text{ref}}\right)^2 dt\right].
\end{aligned}$$

Expand now the quadratic term in the integral and neglect additive terms which do not depend on the control u. Denoting again by $J_0(u)$ the equivalent quantity to be minimized,

one has

$$J_0(u) = E_{\boldsymbol{\Theta}}\left[\int_{t_0}^{T}\left(\frac{\Phi(t,\boldsymbol{\Theta})^2}{V_G^2}\left(\sum_{i=1}^{K}u_i\mathbb{1}\{t>t_{i-1}\}\int_{t_{i-1}}^{\min(t,t_i)}\Phi^{-1}(s,\boldsymbol{\Theta})ds\right)^2\right.\right.$$
$$+2x_h(t,\boldsymbol{\Theta})\frac{\Phi(t,\boldsymbol{\Theta})}{V_G}\sum_{i=1}^{K}u_i\mathbb{1}\{t>t_{i-1}\}\int_{t_{i-1}}^{\min(t,t_i)}\Phi^{-1}(s,\boldsymbol{\Theta})ds$$
$$\left.\left.-2x_{\text{ref}}\frac{\Phi(t,\boldsymbol{\Theta})}{V_G}\sum_{i=1}^{K}u_i\mathbb{1}\{t>t_{i-1}\}\int_{t_{i-1}}^{\min(t,t_i)}\Phi^{-1}(s,\boldsymbol{\Theta})ds\right)dt\right].$$

Using the linearity property of the expectation and the integration, expanding the quadratic appearing in the first term in the cost integral and collecting together similar terms, it follows that

$$J_0(u) = \sum_{i=1}^{K}\frac{u_i^2}{V_G^2}E_{\boldsymbol{\Theta}}\left[\int_{t_0}^{T}\left(\Phi(t,\boldsymbol{\Theta})^2\mathbb{1}\{t>t_{i-1}\}\left(\int_{t_{i-1}}^{\min(t,t_i)}\Phi^{-1}(s,\boldsymbol{\Theta})ds\right)^2\right)dt\right] \tag{5.31}$$
$$+\sum_{i=1}^{K}\sum_{j=i+1}^{K}\frac{2u_iu_j}{V_G^2}E_{\boldsymbol{\Theta}}\left[\int_{t_0}^{T}\left(\Phi(t,\boldsymbol{\Theta})^2\mathbb{1}\{t>t_{i-1}\}\mathbb{1}\{t>t_{j-1}\}\right.\right.$$
$$\left.\left.\int_{t_{i-1}}^{\min(t,t_i)}\Phi^{-1}(s,\boldsymbol{\Theta})ds\int_{t_{j-1}}^{\min(t,t_j)}\Phi^{-1}(s,\boldsymbol{\Theta})ds\right)dt\right] \tag{5.32}$$
$$+\sum_{i=1}^{K}\frac{2u_i}{V_G}E_{\boldsymbol{\Theta}}\left[\int_{t_0}^{T}\left((x_h(t,\boldsymbol{\Theta})-x_{\text{ref}})\Phi(t,\boldsymbol{\Theta})\mathbb{1}\{t>t_{i-1}\}\int_{t_{i-1}}^{\min(t,t_i)}\Phi^{-1}(s,\boldsymbol{\Theta})ds\right)dt\right].$$

Returning now to the quadratic programming formulation in (5.29), the exact expressions for the elements of the matrices A and B are now given. The diagonal entries of A for $i=1,\ldots,K$ are

$$a_{ii} = \frac{1}{V_G^2}E_{\boldsymbol{\Theta}}\left[\int_{t_0}^{T}\left(\Phi(t,\boldsymbol{\Theta})^2\mathbb{1}\{t>t_{i-1}\}\left(\int_{t_{i-1}}^{\min(t,t_i)}\Phi^{-1}(s,\boldsymbol{\Theta})ds\right)^2\right)dt\right] \tag{5.33}$$
$$=\frac{1}{V_G^2}E_{\boldsymbol{\Theta}}\left[\int_{t_{i-1}}^{T}\left(\Phi(t,\boldsymbol{\Theta})^2\left(\int_{t_{i-1}}^{\min(t,t_i)}\Phi^{-1}(s,\boldsymbol{\Theta})ds\right)^2\right)dt\right],$$

and the off-diagonal elements for $i,j=1,\ldots,K$ and $j>i$ are

$$a_{ij} = \frac{2}{V_G^2}E_{\boldsymbol{\Theta}}\left[\int_{t_0}^{T}\left(\Phi(t,\boldsymbol{\Theta})^2\mathbb{1}\{t>t_{i-1}\}\mathbb{1}\{t>t_{j-1}\}\right.\right.$$
$$\left.\left.\int_{t_{i-1}}^{\min(t,t_i)}\Phi^{-1}(s,\boldsymbol{\Theta})ds\int_{t_{j-1}}^{\min(t,t_j)}\Phi^{-1}(s,\boldsymbol{\Theta})ds\right)dt\right] \tag{5.34}$$
$$=\frac{2}{V_G^2}E_{\boldsymbol{\Theta}}\left[\int_{t_{j-1}}^{T}\left(\Phi(t,\boldsymbol{\Theta})^2\int_{t_{i-1}}^{t_i}\Phi^{-1}(s,\boldsymbol{\Theta})ds\int_{t_{j-1}}^{\min(t,t_j)}\Phi^{-1}(s,\boldsymbol{\Theta})ds\right)dt\right].$$

The elements of the vector B for $i = 1, \ldots, K$ are given by

$$b_i = \frac{2}{V_G} E_{\boldsymbol{\Theta}} \left[\int_{t_0}^{T} \left((x_h(t, \boldsymbol{\Theta}) - x_{\text{ref}}) \Phi(t, \boldsymbol{\Theta}) \mathbb{1}\{t > t_{i-1}\} \int_{t_{i-1}}^{\min(t,t_i)} \Phi^{-1}(s, \boldsymbol{\Theta}) ds \right) dt \right] \quad (5.35)$$

$$= \frac{2}{V_G} E_{\boldsymbol{\Theta}} \left[\int_{t_{i-1}}^{T} \left((x_h(t, \boldsymbol{\Theta}) - x_{\text{ref}}) \Phi(t, \boldsymbol{\Theta}) \int_{t_{i-1}}^{\min(t,t_i)} \Phi^{-1}(s, \boldsymbol{\Theta}) ds \right) dt \right].$$

Once the matrices A and B are known, standard numerical procedures for the solution of box constrained quadratic optimization problems can be used for the numerical solution of problem (5.29), see for example in [19, 25].

The control policy is to be applied in the following way. At each measurement time t_{k-1}, the vector $\bar{u} = (\bar{u}_k, \ldots, \bar{u}_K)^T$ is computed as the solution to the optimization problem in equation (5.29). Next, the control $\bar{u}_k$ is applied to the system in the time period $(t_{k-1}, t_k]$. After the measurement at t_k is given, the new $\bar{u} = (\bar{u}_{k+1}, \ldots, \bar{u}_K)^T$ is computed as the minimizer of the cost functional J_k and only $\bar{u}_{k+1}$ is applied. This procedure continues until all measurements are given. The expectations appearing in the control optimization are going to be approximated by sample averages based on the current distribution $\pi^k_{\boldsymbol{\Theta}}$ of the random vector $\boldsymbol{\Theta}$, as in (5.8).

5.4.2 Parameter inference based on SMC methods

A particle filter algorithm will be used to update the distribution of the model parameters after each new glucose measurement becomes available.

To build up an informative prior for the vector $\boldsymbol{\Theta}$, the fitted parameters to the available mice data were used. For simplicity, a uniform prior was assumed for each parameter in the range of the estimated values and the parameters were assumed to be independent before the experiment. All together, the prior on $\boldsymbol{\Theta}$ is of the following form

$$p_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) = \prod_{i=1}^{6} (\theta_i - a_i)(b_i - \theta_i), \ \boldsymbol{\theta} = (\theta_1, \ldots, \theta_6), \quad (5.36)$$

where a_i, b_i, $i = 1, \ldots, 6$ are the minimum and maximum bounds of the estimated parameters. The measurement errors ε_k, $k = 0, \ldots, K$ were assumed to be independent and normally distributed with variance $\sigma^2_M = 15$, so that the likelihood $\mathcal{L}$ after K given measurements has the following form

$$\mathcal{L}(\boldsymbol{\theta}) \propto \prod_{k=0}^{K} \exp\left(-\frac{(y_k - x(t_k, \boldsymbol{\theta}, u))^2}{2\sigma^2_M} \right). \quad (5.37)$$

At each iteration, a multinomial sampling scheme was applied to reduce the number of particles with negligible weight. After this step, all the particles are equally weighted. To avoid degeneracy, an independent transition kernel was used based on the normal approximation of the current set of particles, i.e. for $m = 1, \ldots, M$ a particle $\boldsymbol{\phi}^m$ was proposed from the normal distribution $\mathcal{N}(\hat{\boldsymbol{r}}, \hat{\Sigma})$, with mean

$$\hat{\boldsymbol{r}} = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{\theta}^m, \quad (5.38)$$

and covariance matrix

$$\hat{\Sigma} = \frac{1}{M} \sum_{m=1}^{M} (\boldsymbol{\theta}^m - \hat{\boldsymbol{r}})(\boldsymbol{\theta}^m - \hat{\boldsymbol{r}})^T. \tag{5.39}$$

5.4.3 Numerical simulations

The performance of the above method is demonstrated on the example of one in silico individual. A parameter vector is drawn from the prior distribution and it is fixed. It is denoted by $\boldsymbol{\theta}_{\text{true}}$. Measurements are taken at the times indicated by the protocol and stated in section 5.2.1. The SMC algorithm was based on a sample of size $M = 5000$ particles. In Figure 5.8, the time evolution of the "true" underlying glucose concentration and the sampled data are shown, along with the degradation rate obtained with the true parameter vector and the rate obtained with the mean of the posterior distribution at the end of the experiment. The computed optimal control policy is also given. Figures 5.9 and 5.10 show the density estimator for the distribution of each model parameter after a new measurement became available.

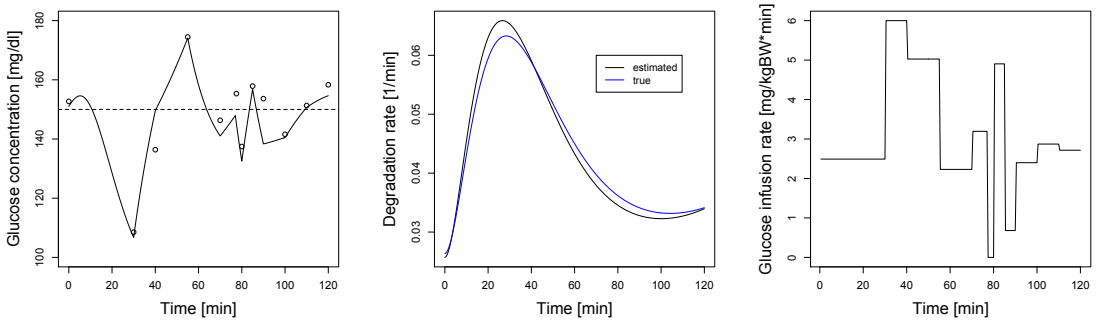

Figure 5.8: First subfigure: Sampled data (circles), model solution for the fixed value $\boldsymbol{\theta}_{\text{true}}$ (solid line) and reference glucemia (dashed line). Second subfigure: true and estimated degradation rate. Third subfigure: computed glucose infusion rate. The parameter vector was fixed at $\boldsymbol{\theta}_{\text{true}} = (2.593369e-02, 3.395887e-04, 7.093037e-02, 4.171830e-04, 2.376361e-02, 1.510564e+02)$. The solution is based on the Monte Carlo version of the algorithm.

It can be seen that the glucose concentration stays around the reference value $G_{\text{ref}} = 150$ [mg/dl] in a more consistent way than the intuitive approach used until now by the biologists. The posterior distribution is centered for all parameters around the nominal value.

5.5 Optimal control based on PC expansions and optimal maps

In this section it is shown how polynomial chaos methods can be used to solve the optimal control and parameter estimation problem discussed in the previous section. Here, the optimal control policy will be computed by expressing the cost functional in terms of

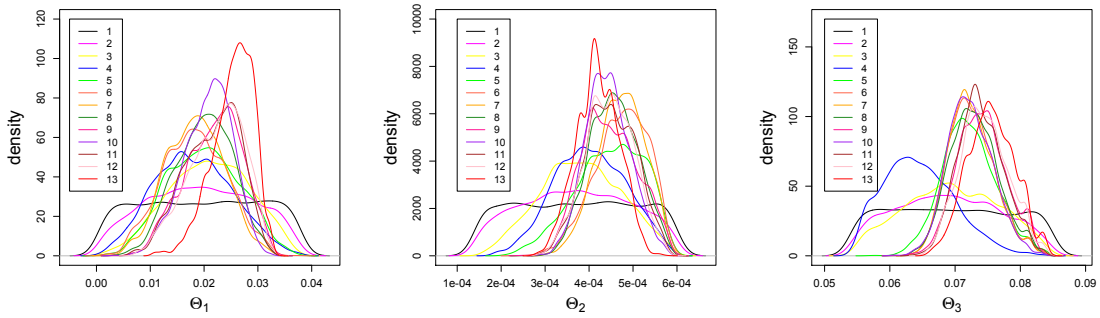

Figure 5.9: Kernel density estimators for parameters Θ_1, Θ_2 and Θ_3. The corresponding components of the true parameter vector were $\theta_{\text{true},1} = 2.593369e-02, \theta_{\text{true},2} = 3.395887e-04, \theta_{\text{true},3} = 7.093037e-02$.

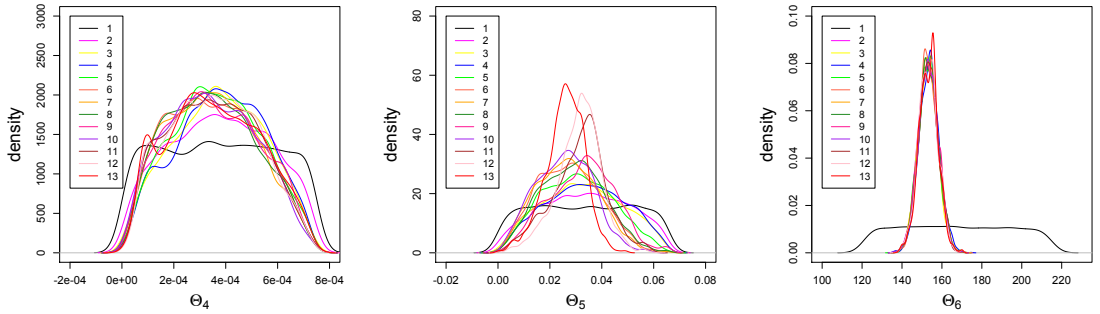

Figure 5.10: Kernel density estimators for parameters Θ_4, Θ_5 and Θ_6. The corresponding components of the true parameter vector were $\theta_{\text{true},4} = 4.171830e-04, \theta_{\text{true},5} = 2.376361e-02, \theta_{\text{true},6} = 1.510564e+02$.

the PC coefficients of the solution and solving the related quadratic optimization problem. The parameter inference problem will be solved by combining SMC methods and the construction of optimal maps discussed in chapter 3. Recall the problem formulation and the notation introduced in equations (5.20) - (5.24). The algorithm flow is stated here and then the individual steps will be closer examined.

- Set counter $k = 1$, initialize a prior distribution $p_{\boldsymbol{\Theta}}$, build a PCE for $\boldsymbol{\Theta}$ based on this prior, and generate an equally weighted sample $\{(\boldsymbol{\theta}^m, \frac{1}{M})\}_{m=1}^M$ from the prior. While $k < K$,

  A. Estimation

  - get at time t_{k-1} the measurement y_{k-1},
  - apply Bayes theorem to obtain the posterior distribution $\pi_{\boldsymbol{\Theta}}^k$,
  - use an SMC step and update the sample to $\{(\boldsymbol{\theta}^m, w^k(\boldsymbol{\theta}^m))\}_{m=1}^M$,
  - use this sample to build a PCE for $\boldsymbol{\Theta}$ via the optimal maps,

B. Prediction

- build the Galerkin system for the PC coefficients of the glucose dynamics,

C. Control

- compute $\bar{u} = (\bar{u}_k, \ldots, \bar{u}_K) = \text{argmin} J_{k-1}(u)$,
- apply $\bar{u}_k$ to the system in the time interval $(t_{k-1}, t_k]$ and

- set as new prior the posterior $\pi^k_{\boldsymbol{\Theta}}$, $k \to k+1$ and go to the estimation step.

5.5.1 Polynomial chaos approximation for the glucose dynamics

The polynomial chaos approximation of the homogenous glucose dynamics

$$\begin{aligned} \dot{x}_h(t, \boldsymbol{\theta}) &= -\gamma(t, \boldsymbol{\theta}) x_h(t, \boldsymbol{\theta}), \\ x_h(t_0, \boldsymbol{\theta}) &= x_0(\boldsymbol{\theta}) \end{aligned} \tag{5.40}$$

is presented here. This forms the base for the PC approximation of the non-homogenous solution, since the non-homogeneity is an additive deterministic function. By using firstly a NISP approach the polynomial order sufficient for the application is determined.

Let $\boldsymbol{\Xi}$ be a vector of basis random variables and let $\{P_{\boldsymbol{n}} : \boldsymbol{n} \in \mathbb{N}_0^6\}$ be the corresponding orthogonal polynomial sequence. Assume that each parameter has a polynomial expansion of the form

$$\Theta_i = g_{\Theta_i}(\boldsymbol{\Xi}) = \sum_{|\boldsymbol{n}|=0}^{\infty} \lambda_{i;\boldsymbol{n}} P_{\boldsymbol{n}}(\boldsymbol{\Xi}) h_{\boldsymbol{n}}, \ i = 1, \ldots, 6, \ \boldsymbol{n} \in \mathbb{N}_0^6, \tag{5.41}$$

and so does the solution of equation (5.40)

$$x_h(t, \boldsymbol{\Theta}) = x_h(t, g_{\boldsymbol{\Theta}}(\boldsymbol{\Xi})) \equiv x_h(t, \boldsymbol{\Xi}) = \sum_{|\boldsymbol{n}|=0}^{\infty} q_{h,\boldsymbol{n}}(t) P_{\boldsymbol{n}}(\boldsymbol{\Xi}) h_{\boldsymbol{n}}, \ \boldsymbol{n} \in \mathbb{N}_0^6, \tag{5.42}$$

where $g_{\boldsymbol{\Theta}}(\boldsymbol{\Xi}) = (g_{\Theta_1}(\boldsymbol{\Xi}), \ldots, g_{\Theta_6}(\boldsymbol{\Xi}))$. Let $N \in \mathbb{N}_0$ and denote by $x_{h,N}$ the PC approximation of order N of the homogenous solution x_h in (5.40). $\mathcal{P}^N$ stands for the corresponding polynomial subspace as defined in equation (2.42). For the numerical simulations, the basis random vector $\boldsymbol{\Xi}$ was chosen to consist of independent uniformly distributed random variables on $[-1, 1]$, so that the corresponding orthogonal polynomial sequence are the 6-dimensional Legendre polynomials. The parameter vector $\boldsymbol{\Theta}$ is assumed to be distributed according to the uniform prior given in equation (5.36), so that before any data are given only the 0-th and 1-st terms in the functions g_{Θ_i}, $i = 1, \ldots, 6$ have non-zero PC coefficients.

NISP approach

As explained in chapter 2, one way to estimate the PC coefficients in equation (5.42) is by numerically computing the integrals

$$q_{h,\boldsymbol{n}}(t) = \int_{\boldsymbol{S}} x_h(t, g_{\boldsymbol{\Theta}}(\boldsymbol{\xi})) P_{\boldsymbol{n}}(\boldsymbol{\xi}) d\boldsymbol{\mu}(\boldsymbol{\xi}), \ |\boldsymbol{n}| = 0, \ldots, N, \ \boldsymbol{n} \in \mathbb{N}_0^6, \tag{5.43}$$

where $\boldsymbol{S} = [-1, 1]^6$ is the support of the uniform measure $\boldsymbol{\mu}$. Two orders of approximations are considered, $N = 3, 4$ and the integrals were approximated by Monte Carlo integration based on a sample of size $M = 60000$ generated from the distribution of $\boldsymbol{\Xi}$. The relative error over time

$$\text{err}_{\text{rel}}(t) = \frac{\| x_h(t, \boldsymbol{\Xi}) - x_{h,N}(t, \boldsymbol{\Xi}) \|_{L^2}}{\| x_h(t, \boldsymbol{\Xi}) \|_{L^2}} \tag{5.44}$$

is estimated also via Monte Carlo integration for both approximation orders and shown in Figure 5.11. As it can be seen, an approximation order of $N = 4$ is enough to keep the relative error smaller than 5% in the time interval $[t_0, T]$.

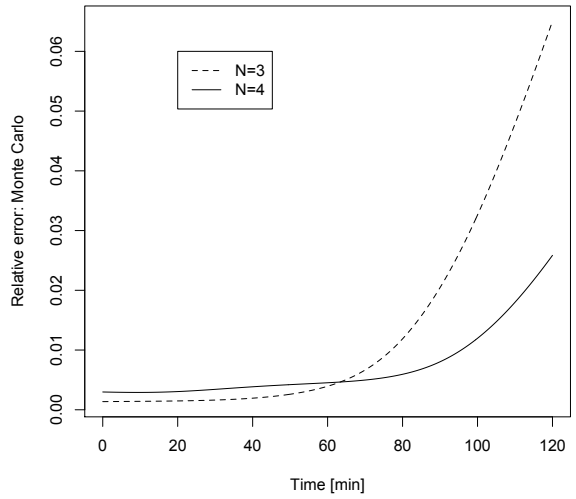

Figure 5.11: Relative error of the PC approximation of the homogenous solution computed via a Monte Carlo NISP approach.

Galerkin approach

In order to build the Galerkin system for the homogenous solution, one needs to compute the PC coefficients in the expansion of the degradation rate $\gamma(t, \boldsymbol{\Theta})$,

$$\gamma(t, \boldsymbol{\Theta}) = \gamma(t, g_{\boldsymbol{\Theta}}(\boldsymbol{\Xi})) \equiv \gamma(t, \boldsymbol{\Xi}) = \sum_{|\boldsymbol{n}|=0}^{\infty} \gamma_{\boldsymbol{n}}(t) P_{\boldsymbol{n}}(\boldsymbol{\Xi}) h_{\boldsymbol{n}}, \boldsymbol{n} \in \mathbb{N}_0^6. \tag{5.45}$$

The PC coefficients of the N-th order approximation of the degradation rate

$$\gamma_N(t, \boldsymbol{\Xi}) = \sum_{|\boldsymbol{n}|=0}^{N} \gamma_{\boldsymbol{n}}(t) P_{\boldsymbol{n}}(\boldsymbol{\Xi}) h_{\boldsymbol{n}} \tag{5.46}$$

will be approximated by a NISP approach based on Monte Carlo and sparse grids integration. Once this PC expansion is build, then one can solve the Galerkin system for the

coefficients of the model solution. In Figure 5.12 the relative error for the degradation rate is estimated by Monte Carlo integration with a sample of size $M = 40000$ and for orders $N = 3, 4$, and by using sparse grids based on the Kronrod-Patterson rule for level $l = 5$, [69,79]. This error is defined as

$$\text{err}_{\text{rel}}(t) = \frac{\| \gamma(t, \boldsymbol{\Xi}) - \gamma_N(t, \boldsymbol{\Xi}) \|_{L^2}}{\| \gamma(t, \boldsymbol{\Xi}) \|_{L^2}}. \tag{5.47}$$

As the relative errors are comparable, sparse grids will be in what follows preferred as they require less computational effort and time.

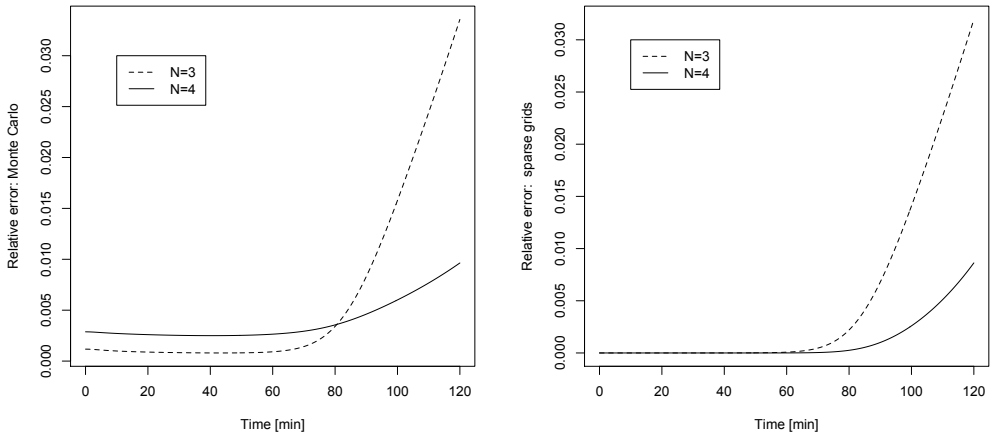

Figure 5.12: Relative error of the PC approximation of the degradation rate estimated via NISP approaches: Monte Carlo (first subfigure) and sparse grid (second subfigure) integration.

The Galerkin system for the homogenous solution is now formulated. By substituting the truncated PC expansions of total order N of the stochastic processes $\gamma(t, \boldsymbol{\Xi})$ and $x_h(t, \boldsymbol{\Xi})$ in the system dynamics, one has

$$\sum_{|\boldsymbol{n}|=0}^{N} \dot{q}_{h,\boldsymbol{n}}(t) P_{\boldsymbol{n}}(\boldsymbol{\Xi}) h_{\boldsymbol{n}} = - \left(\sum_{|\boldsymbol{m}|=0}^{N} \gamma_{\boldsymbol{m}}(t) P_{\boldsymbol{m}}(\boldsymbol{\Xi}) h_{\boldsymbol{m}} \right) \left(\sum_{|\boldsymbol{n}|=0}^{N} q_{h,\boldsymbol{n}}(t) P_{\boldsymbol{n}}(\boldsymbol{\Xi}) h_{\boldsymbol{n}} \right). \tag{5.48}$$

By projecting the residual on the subspace $\mathcal{P}^N$ and taking expectations, one obtains

$$\sum_{|\boldsymbol{n}|=0}^{N} \dot{q}_{h,\boldsymbol{n}}(t) \langle P_{\boldsymbol{n}}, P_{\boldsymbol{l}} \rangle h_{\boldsymbol{n}} = - \sum_{|\boldsymbol{m}|=0}^{N} \gamma_{\boldsymbol{m}}(t) \sum_{|\boldsymbol{n}|=0}^{N} q_{h,\boldsymbol{n}}(t) \langle P_{\boldsymbol{m}} P_{\boldsymbol{n}}, P_{\boldsymbol{l}} \rangle h_{\boldsymbol{m}} h_{\boldsymbol{n}}, \tag{5.49}$$

and by orthogonality one has for all $\boldsymbol{l} \in \mathbb{N}_0^6$ with $|\boldsymbol{l}| = 0, \ldots, N$

$$\dot{q}_{h,\boldsymbol{l}}(t) = - \sum_{|\boldsymbol{m}|=0}^{N} \gamma_{\boldsymbol{m}}(t) \sum_{|\boldsymbol{n}|=0}^{N} q_{h,\boldsymbol{n}}(t) \langle P_{\boldsymbol{m}} P_{\boldsymbol{n}}, P_{\boldsymbol{l}} \rangle h_{\boldsymbol{m}} h_{\boldsymbol{n}}, \tag{5.50}$$

or in matrix form

$$\dot{q}_h(t) = -A(t)q_h(t), \tag{5.51}$$

where $q_h(t) = [q_{h,0}, q_{h,1}, \ldots, q_{h,d_P}]^T$ is the vector of PC coefficients in single index notation, $d_P + 1 = dim\mathcal{P}^N$ and

$$A(t) = \begin{bmatrix} \sum_{|\boldsymbol{m}|=0}^{N} \gamma_{\boldsymbol{m}}(t)\langle P_{\boldsymbol{m}}P_0, P_0\rangle h_{\boldsymbol{m}}h_0 & \cdots & \sum_{|\boldsymbol{m}|=0}^{N} \gamma_{\boldsymbol{m}}(t)\langle P_{\boldsymbol{m}}P_{d_P}, P_0\rangle h_{\boldsymbol{m}}h_{d_P} \\ \sum_{|\boldsymbol{m}|=0}^{N} \gamma_{\boldsymbol{m}}(t)\langle P_{\boldsymbol{m}}P_0, P_1\rangle h_{\boldsymbol{m}}h_1 & \cdots & \sum_{|\boldsymbol{m}|=0}^{N} \gamma_{\boldsymbol{m}}(t)\langle P_{\boldsymbol{m}}P_{d_P}, P_1\rangle h_{\boldsymbol{m}}h_{d_P} \\ \vdots & & \\ \sum_{|\boldsymbol{m}|=0}^{N} \gamma_{\boldsymbol{m}}(t)\langle P_{\boldsymbol{m}}P_0, P_{d_P}\rangle h_{\boldsymbol{m}}h_1 & \cdots & \sum_{|\boldsymbol{m}|=0}^{N} \gamma_{\boldsymbol{m}}(t)\langle P_{\boldsymbol{m}}P_{d_P}, P_{d_P}\rangle h_{\boldsymbol{m}}h_{d_P} \end{bmatrix}. \tag{5.52}$$

This system has to be solved with initial value the vector of PC coefficients of the parameter Θ_6, which represents the unknown initial condition for the glucose dynamics. In single index notation, this means that

$$q_{h,j}(t_0) = \lambda_{6,j}, \ \forall\, j = 0, \ldots, d_P. \tag{5.53}$$

In Figure 5.13, the relative error of using a Galerkin approximation of orders $N = 3, 4$ for the homogenous solution based on a sparse grid approximation of the degradation rate of depth $l = 6$ is plotted. As in the NISP approach, the 4-th order approximation provides with the desired accuracy.

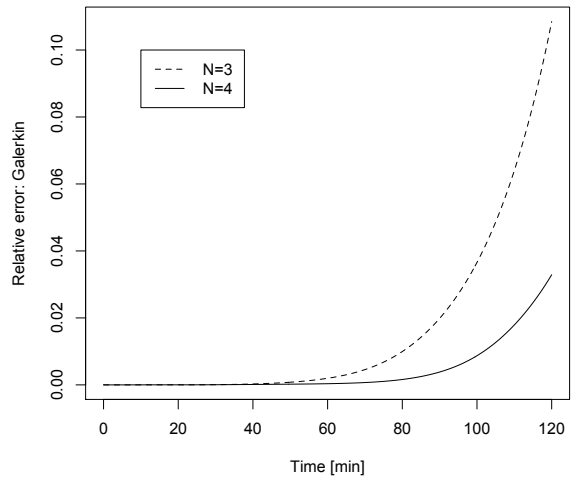

Figure 5.13: Relative error of the PC approximation of the homogenous solution via Galerkin projection.

5.5.2 Quadratic optimization

Let now

$$x(t,\mathbf{\Theta},u)=x(t,g_{\mathbf{\Theta}}(\mathbf{\Xi}),u)\equiv x(t,\mathbf{\Xi},u)=\sum_{|\boldsymbol{n}|=0}^{\infty}q_{\boldsymbol{n}}(t,u)P_{\boldsymbol{n}}(\mathbf{\Xi})h_{\boldsymbol{n}},\ \boldsymbol{n}\in\mathbb{N}_0^6 \tag{5.54}$$

be the PC expansion of the non-homogenous system and denote by $x_N(t,\mathbf{\Xi})$ its approximation in the subspace $\mathcal{P}^N$.

The Galerkin system for the coefficients $\{q_{\boldsymbol{n}}(t,u)\colon |\boldsymbol{n}|=0,\dots,N,\boldsymbol{n}\in\mathbb{N}_0^6\}$ has the following form

$$\begin{aligned}\dot{q}(t,u)&=-A(t)q(t,u)+u(t)\\ q(t_0,u)&=q^0,\end{aligned} \tag{5.55}$$

where $q(t,u)=(q_0(t,u),\dots,q_{d_P}(t,u))$ is the vector of the PC coefficients in single index notation, $A(t)$ is the matrix in equation (5.52), q^0 is the initial condition defined by equations (5.53) and now $u(t)$ is the vector

$$u(t)=[\frac{1}{V_g}\sum_{i=1}^{K}u_i\mathbb{1}\{(t_{i-1},t_i]\}(t),0,\dots,0]^T. \tag{5.56}$$

Recall now the optimization problem formulation defined through equations (5.20)-(5.23). By exchanging the integral over time and the expectation in equation (5.20) and expanding the square, one obtains

$$J_0(u)=\int_{t_0}^{T}E_{\mathbf{\Theta}}[(x(t,\mathbf{\Theta},u)-x_{\text{ref}})^2]dt=\int_{t_0}^{T}\left(E_{\mathbf{\Theta}}[x(t,\mathbf{\Theta},u)^2]-2x_{\text{ref}}E_{\mathbf{\Theta}}[x(t,\mathbf{\Theta},u)]+x_{\text{ref}}^2\right)dt. \tag{5.57}$$

As seen in Chapter 2, the moments of a random variable can be expressed by its PC coefficients. Substituting expressions of the form (2.45) and (2.46) in (5.57) and neglecting additive terms that do not depend on the control u, one has

$$J_0(u)=\int_{t_0}^{T}\left(\sum_{|\boldsymbol{n}|=0}^{\infty}q_{\boldsymbol{n}}^2(t,u)h_{\boldsymbol{n}}-2x_{\text{ref}}q_{\mathbf{0}}(t,u)\right)dt. \tag{5.58}$$

The infinite summation will be substituted in what follows by the order N-th approximation for computational purposes.

Next, it will be shown how the cost functional in equation (5.58) can be brought to the form

$$J_0(u)=u^TCu+Du,\ u=(u_1,\dots,u_K),\ C\in\mathbb{R}^{K\times K},\ D\in\mathbb{R}^{K\times 1}, \tag{5.59}$$

which is suitable for quadratic optimization.

Let $\Psi(t,s)=\Phi(t)\Phi(s)^{-1}$, where $\Phi(t)$ is the fundamental matrix of the linear system (5.51). Then, the solution of the system (5.55) can be computed by the method of variation of constants and has the form

$$q(t,u)=\Psi(t,t_0)q^0+\int_{t_0}^{t}\Psi(t,s)u(s)ds. \tag{5.60}$$

Substituting in the above equation $u(s)$ by equation (5.56), the integral term can be rewritten as

$$\int_{t_0}^{t} \Psi(t,s)u(s)ds = \int_{t_0}^{t} \Psi(t,s)\frac{1}{V_g}\begin{bmatrix} \sum_{i=1}^{K} u_i \mathbb{1}\{(t_{i-1},t_i]\}(s) \\ 0 \\ \vdots \\ 0 \end{bmatrix} ds \tag{5.61}$$

$$= \frac{1}{V_g}\int_{t_0}^{t} \Psi(t,s) \sum_{i=1}^{K} \begin{bmatrix} u_i \\ 0 \\ \vdots \\ 0 \end{bmatrix} \mathbb{1}\{(t_{i-1},t_i]\} ds$$

$$= \frac{1}{V_g}\sum_{i=1}^{K} u_i \mathbb{1}\{t > t_{i-1}\} \int_{t_{i-1}}^{\min(t,t_i)} \Psi(t,s)\, b\, ds,$$

where $b = [1, 0, \ldots, 0]^T$. Thus, returning to (5.60) this can be written as

$$q(t,u) = \Psi(t,t_0)q^0 + \frac{1}{V_g}\sum_{i=1}^{K} u_i \mathbb{1}\{t > t_{i-1}\} \int_{t_{i-1}}^{\min(t,t_i)} \Psi(t,s)\, b\, ds \tag{5.62}$$

$$= \Psi(t,t_0)q^0 + \frac{1}{V_g}\sum_{i=1}^{K} u_i \mathbb{1}\{t > t_{i-1}\}\Psi(t,\min(t,t_i)) \int_{t_{i-1}}^{\min(t,t_i)} \Psi(\min(t,t_i),s)\, b\, ds,$$

where the translation property of the matrix $\Psi(t,s)$ has been used. Define the functions

$$f_{i-1}(t) = \left(\int_{t_{i-1}}^{t} \Psi(t,s)\, b\, ds\right),\ i = 1, \ldots, K, \tag{5.63}$$

and

$$v_{i-1}(t) = \left(\Psi(t,t_i)\int_{t_{i-1}}^{t_i} \Psi(t_i,s)\, b\, ds\right),\ i = 1, \ldots, K-1. \tag{5.64}$$

Then, it holds

$$q(t,u) = q_h(t) + \frac{1}{V_g}\sum_{i=1}^{K} u_i \mathbb{1}\{t > t_{i-1}\}\left(\mathbb{1}\{t > t_i\}v_{i-1}(t) + \mathbb{1}\{t < t_i\}f_{i-1}(t)\right), \tag{5.65}$$

where $q_h(t)$ can be computed by solving the homogenous problem of equation (5.51)

$$\dot{q}_h(t) = -A(t)q_h(t),\ q_h(t_0) = q^0. \tag{5.66}$$

The functions f_{i-1}, $i = 1, \ldots K$ and v_{i-1}, $i = 1, \ldots, K$ can be computed by solving appropriate initial value problems as shown next. By differentiating the functions f_{i-1} one has

$$\dot{f}_{i-1}(t) = \frac{d}{dt}\left(\int_{t_{i-1}}^{t} \Psi(t,s)\, b\, ds\right) = \left(\int_{t_{i-1}}^{t} \frac{d}{dt}\Psi(t,s)\, b\, ds\right) + \Psi(t,t)\, b - \Psi(t,t_i)\, 0 \tag{5.67}$$

$$= b + A(t)f_{i-1}(t),\ i = 1, \ldots, K.$$

This system has to be solved with initial condition

$$f_{i-1}(t_{i-1}) = 0, \tag{5.68}$$

as one can easily see by substituting t with t_{i-1} in equation (5.63). For the functions v_{i-1}, one has

$$\begin{aligned}\dot{v}_{i-1}(t) &= \frac{d}{dt}\left(\Psi(t,t_i)\int_{t_{i-1}}^{t_i}\Psi(t_i,s)\,b\,ds\right) = \left(\frac{d}{dt}\Psi(t,t_i)\right)\int_{t_{i-1}}^{t_i}\Psi(t_i,s)\,b\,ds + \Psi(t,t_i)\,0 \\ &= A(t)\Psi(t,t_i)\int_{t_{i-1}}^{t_i}\Psi(t_i,s)\,b\,ds = A(t)v_{i-1}(t),\ i=1,\dots,K-1,\end{aligned} \tag{5.69}$$

with initial condition computed by equations (5.63) and (5.64)

$$v_{i-1}(t_i) = f_{i-1}(t_i). \tag{5.70}$$

Here, the Leibniz rule for differentiating an integral has been used, namely

$$\frac{d}{dt}\int_{a(t)}^{b(t)} f(t,s)ds = \int_{a(t)}^{b(t)} f_t(t,s)ds + f(t,b(t))\,b'(t) - f(t,a(t))\,a'(t), \tag{5.71}$$

for all continuous functions $f\colon \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ such that the derivative $f_t(t,s)$ exists and is continuous.

Denote the n-th component of the functions $q_h(t), f_{i-1}(t), v_{i-1}(t)$ as $q_h^n(t), f_{i-1}^n(t)$ and $v_{i-1}^n(t)$ respectively. Then, the cost functional $J_0(u)$ can be written as

$$\begin{aligned}J_0(u) = \int_{t_0}^{T}\Bigg[\sum_{n=0}^{N} h_n\Bigg(q_h^{n+1}(t) + \frac{1}{V_g}\sum_{i=1}^{K} u_i \mathbb{1}\{t>t_{i-1}\}(\mathbb{1}\{t>t_i\}v_{i-1}(t) + \mathbb{1}\{t<t_i\}f_{i-1}(t))\Bigg)^2 \\ - 2x_{\text{ref}}\Bigg(q_h^1(t) + \frac{1}{V_g}\sum_{i=1}^{K} u_i \mathbb{1}\{t>t_{i-1}\}(\mathbb{1}\{t>t_i\}v_{i-1}^1(t) + \mathbb{1}\{t<t_i\}f_{i-1}^1(t))\Bigg)\Bigg]\,dt.\end{aligned}$$

Expanding the quadratic term yields

$$\begin{aligned}J_0(u) = \int_{t_0}^{T}\Bigg[\sum_{n=0}^{N} h_n\Bigg((q_h^{n+1}(t))^2 + 2q_h^{n+1}(t)\frac{1}{V_g}\sum_{i=1}^{K} u_i \mathbb{1}\{t>t_{i-1}\} \\ (\mathbb{1}\{t>t_i\}v_{i-1}(t) + \mathbb{1}\{t<t_i\}f_{i-1}(t)) \\ + \frac{1}{V_g^2}\Bigg(\sum_{i=1}^{K} u_i \mathbb{1}\{t>t_{i-1}\}(\mathbb{1}\{t>t_i\}v_{i-1}(t) + \mathbb{1}\{t<t_i\}f_{i-1}(t))\Bigg)^2\Bigg) \\ - 2x_{\text{ref}}q_h^1(t) - 2x_{\text{ref}}\frac{1}{V_g}\sum_{i=1}^{K} u_i \mathbb{1}\{t>t_{i-1}\}(\mathbb{1}\{t>t_i\}v_{i-1}^1(t) + \mathbb{1}\{t<t_i\}f_{i-1}^1(t))\Bigg]\,dt.\end{aligned}$$

Neglecting additive terms that do not depend on the control u, it holds

$$J_0(u) = \int_{t_0}^{T} \Bigg[\sum_{n=0}^{N} h_n \Bigg(2q_h^{n+1}(t) \frac{1}{V_g} \sum_{i=1}^{K} u_i \mathbb{1}\{t > t_{i-1}\} (\mathbb{1}\{t > t_i\} v_{i-1}(t) + \mathbb{1}\{t < t_i\} f_{i-1}(t)) + \frac{1}{V_g^2} \left(\sum_{i=1}^{K} u_i \mathbb{1}\{t > t_{i-1}\} (\mathbb{1}\{t > t_i\} v_{i-1}(t) + \mathbb{1}\{t < t_i\} f_{i-1}(t)) \right)^2 \Bigg) - 2x_{\text{ref}} \frac{1}{V_g} \sum_{i=1}^{K} u_i \mathbb{1}\{t > t_{i-1}\} (\mathbb{1}\{t > t_i\} v_{i-1} 1(t) + \mathbb{1}\{t < t_i\} f_{i-1}^1(t)) \Bigg] dt. \tag{5.72}$$

Consider now only the quadratic term

$$Q(t) = \left(\sum_{i=1}^{K} u_i \mathbb{1}\{t > t_{i-1}\} \left(\mathbb{1}\{t > t_i\} v_{i-1}(t) + \mathbb{1}\{t < t_i\} f_{i-1}(t) \right) \right)^2. \tag{5.73}$$

Expanding the square, one obtains

$$
\begin{aligned}
Q(t) &= \sum_{i=1}^{K} u_i^2 \mathbb{1}\{t > t_{i-1}\}(\mathbb{1}\{t > t_i\} v_{i-1}(t) + \mathbb{1}\{t < t_i\} f_{i-1}(t))^2 \\
&\quad + 2 \sum_{i=1}^{K} \sum_{j=i+1}^{K} u_i u_j \mathbb{1}\{t > t_{i-1}\}\mathbb{1}\{t > t_{j-1}\}(\mathbb{1}\{t > t_i\} v_{i-1}(t) \\
&\qquad + \mathbb{1}\{t < t_i\} f_{i-1}(t))(\mathbb{1}\{t > t_j\} v_{j-1}(t) + \mathbb{1}\{t < t_j\} f_{j-1}(t)) \\
&= \sum_{i=1}^{K} u_i^2 \mathbb{1}\{t > t_{i-1}\}(\mathbb{1}\{t > t_i\}(v_{i-1}(t))^2 + 2\mathbb{1}\{t > t_i\}\mathbb{1}\{t < t_i\} v_{i-1}(t) f_{i-1}(t) \\
&\qquad + \mathbb{1}\{t < t_i\}(f_{i-1}(t))^2) \\
&\quad + 2 \sum_{i=1}^{K} \sum_{j=i+1}^{K} u_i u_j \mathbb{1}\{t > t_{j-1}\}(\mathbb{1}\{t > t_i\}\mathbb{1}\{t > t_j\} v_{i-1}(t) v_{j-1}(t) \\
&\qquad + \mathbb{1}\{t > t_i\}\mathbb{1}\{t < t_j\} v_{i-1}(t) f_{j-1}(t) + \mathbb{1}\{t < t_i\}\mathbb{1}\{t > t_j\} f_{i-1}(t) v_{j-1}(t) \\
&\qquad + \mathbb{1}\{t < t_i\}\mathbb{1}\{t < t_j\} f_{i-1}(t) f_{j-1}(t)) \\
&= \sum_{i=1}^{K} u_i^2 \mathbb{1}\{t > t_{i-1}\}(\mathbb{1}\{t > t_i\}(v_{i-1}(t))^2 + \mathbb{1}\{t < t_i\}(f_{i-1}(t))^2) \\
&\quad + 2 \sum_{i=1}^{K} \sum_{j=i+1}^{K} u_i u_j \mathbb{1}\{t > t_{j-1}\}(\mathbb{1}\{t > t_j\} v_{i-1}(t) v_{j-1}(t) + \mathbb{1}\{t_i < t < t_j\} v_{i-1}(t) f_{j-1}(t) \\
&\qquad + \mathbb{1}\{t < t_i\} f_{i-1}(t) f_{j-1}(t)) \\
&= \sum_{i=1}^{K} u_i^2 \mathbb{1}\{t > t_{i-1}\}\mathbb{1}\{t > t_i\}(v_{i-1}(t))^2 + \sum_{i=1}^{K} u_i^2 \mathbb{1}\{t > t_{i-1}\}\mathbb{1}\{t < t_i\}(f_{i-1}(t))^2 \\
&\quad + 2 \sum_{i=1}^{K} \sum_{j=i+1}^{K} u_i u_j \mathbb{1}\{t > t_{j-1}\}\mathbb{1}\{t > t_j\} v_{i-1}(t) v_{j-1}(t) \\
&\quad + 2 \sum_{i=1}^{K} \sum_{j=i+1}^{K} u_i u_j \mathbb{1}\{t > t_{j-1}\}\mathbb{1}\{t_i < t < t_j\} v_{i-1}(t) f_{j-1}(t) \\
&\quad + 2 \sum_{i=1}^{K} \sum_{j=i+1}^{K} u_i u_j \mathbb{1}\{t > t_{j-1}\}\mathbb{1}\{t < t_i\} f_{i-1}(t) f_{j-1}(t).
\end{aligned}
$$

All together, one has

$$
\begin{aligned}
Q(t) = &\sum_{i=1}^{K} u_i^2 \mathbb{1}\{t > t_i\}(v_{i-1}(t))^2 + \sum_{i=1}^{K} u_i^2 \mathbb{1}\{t_{i-1} < t < t_i\}(f_{i-1}(t))^2 \\
&+ 2\sum_{i=1}^{K}\sum_{j=i+1}^{K} u_i u_j \mathbb{1}\{t > t_j\} v_{i-1}(t) v_{j-1}(t) \\
&+ 2\sum_{i=1}^{K}\sum_{j=i+1}^{K} u_i u_j \mathbb{1}\{t_{j-1} < t < t_j\} v_{i-1}(t) f_{j-1}(t).
\end{aligned} \tag{5.74}
$$

Substitute this expression back in the cost functional in (5.72) to obtain

$$
\begin{aligned}
J_0(u) = &\int_{t_0}^{T} \left[\sum_{n=0}^{N} h_n 2 q_h^{n+1}(t) \frac{1}{V_g} \sum_{i=1}^{K} u_i \mathbb{1}\{t > t_{i-1}\}(\mathbb{1}\{t > t_i\} v_{i-1}(t) + \mathbb{1}\{t < t_i\} f_{i-1}(t)) \right] dt \\
&+ \int_{t_0}^{T} \left[\sum_{n=0}^{N} h_n \frac{1}{V_g^2} Q(t) \right] dt \\
&- \int_{t_0}^{T} \left[2x_{\text{ref}} \frac{1}{V_g} \sum_{i=1}^{K} u_i \mathbb{1}\{t > t_{i-1}\}(\mathbb{1}\{t > t_i\} v_{i-1} 1(t) + \mathbb{1}\{t < t_i\} f_{i-1}^1(t)) \right] dt \\
&= \int_{t_0}^{T} \left[\frac{2}{V_g} \sum_{n=0}^{N} h_n q_h^{n+1}(t) \left(\sum_{i=1}^{K} u_i \mathbb{1}\{t > t_{i-1}\}\mathbb{1}\{t > t_i\} v_{i-1}(t) \right.\right. \\
&\qquad \left.\left. + \sum_{i=1}^{K} u_i \mathbb{1}\{t > t_{i-1}\}\mathbb{1}\{t < t_i\} f_{i-1}(t) \right) \right] dt \\
&+ \int_{t_0}^{T} \left[\frac{1}{V_g^2} \sum_{n=0}^{N} h_n \left(\sum_{i=1}^{K} u_i^2 \mathbb{1}\{t > t_i\}(v_{i-1}(t))^2 + \sum_{i=1}^{K} u_i^2 \mathbb{1}\{t_{i-1} < t < t_i\}(f_{i-1}(t))^2 \right.\right. \\
&\qquad + 2\sum_{i=1}^{K}\sum_{j=i+1}^{K} u_i u_j \mathbb{1}\{t > t_j\} v_{i-1}(t) v_{j-1}(t) \\
&\qquad \left.\left. + 2\sum_{i=1}^{K}\sum_{j=i+1}^{K} u_i u_j \mathbb{1}\{t_{j-1} < t < t_j\} v_{i-1}(t) f_{j-1}(t) \right) \right] dt \\
&+ \int_{t_0}^{T} \left[\frac{2x_{\text{ref}}}{V_g} \left(\sum_{i=1}^{K} u_i \mathbb{1}\{t > t_{i-1}\}\mathbb{1}\{t > t_i\} v_{i-1}^1(t) + \sum_{i=1}^{K} u_i \mathbb{1}\{t > t_{i-1}\}\mathbb{1}\{t < t_i\} f_{i-1}^1(t) \right) \right] dt.
\end{aligned}
$$

By exchanging the finite sums and the integration in the above expression, it holds

$$
\begin{aligned}
J_0(u) = & \frac{2}{V_g}\sum_{i=1}^{K} u_i \int_{t_0}^{T} \left[\mathbb{1}\{t > t_{i-1}\}\mathbb{1}\{t > t_i\} \sum_{n=0}^{N} h_n q_h^{n+1}(t) v_{i-1}(t) \right] dt \\
& + \frac{2}{V_g}\sum_{i=1}^{K} u_i \int_{t_0}^{T} \left[\mathbb{1}\{t > t_{i-1}\}\mathbb{1}\{t < t_i\} \sum_{n=0}^{N} h_n q_h^{n+1}(t) f_{i-1}(t) \right] dt \\
& + \frac{1}{V_g^2}\sum_{i=1}^{K} u_i^2 \int_{t_0}^{T} \left[\mathbb{1}\{t > t_i\} \sum_{n=0}^{N} h_n (v_{i-1}(t))^2 \right] dt \\
& + \frac{1}{V_g^2}\sum_{i=1}^{K} u_i^2 \int_{t_0}^{T} \left[\mathbb{1}\{t_{i-1} < t < t_i\} \sum_{n=0}^{N} h_n (f_{i-1}(t))^2 \right] dt \\
& + \frac{2}{V_g^2}\sum_{i=1}^{K}\sum_{j=i+1}^{K} u_i u_j \int_{t_0}^{T} \left[\mathbb{1}\{t > t_j\} \sum_{n=0}^{N} h_n v_{i-1}(t) v_{j-1}(t) \right] dt \\
& + \frac{2}{V_g^2}\sum_{i=1}^{K}\sum_{j=i+1}^{K} u_i u_j \int_{t_0}^{T} \left[\mathbb{1}\{t_{j-1} < t < t_j\} \sum_{n=0}^{N} h_n v_{i-1}(t) f_{j-1}(t) \right] dt \\
& - \frac{2x_{\text{ref}}}{V_g}\sum_{i=1}^{K} u_i \int_{t_0}^{T} \left[\mathbb{1}\{t > t_{i-1}\}\mathbb{1}\{t > t_i\} v_{i-1}^1(t) \right] dt \\
& - \frac{2x_{\text{ref}}}{V_g}\sum_{i=1}^{K} u_i \int_{t_0}^{T} \left[\mathbb{1}\{t > t_{i-1}\}\mathbb{1}\{t < t_i\} f_{i-1}^1(t) \right] dt.
\end{aligned}
$$

Taking under consideration the characteristic functions, the above expression becomes

$$
\begin{aligned}
J_0(u) = & \frac{2}{V_g}\sum_{i=1}^{K} u_i \int_{t_i}^{T} \left[\sum_{n=0}^{N} h_n q_h^{n+1}(t) v_{i-1}(t) \right] dt + \frac{2}{V_g}\sum_{i=1}^{K} u_i \int_{t_{i-1}}^{t_i} \left[\sum_{n=0}^{N} h_n q_h^{n+1}(t) f_{i-1}(t) \right] dt \\
& + \frac{1}{V_g^2}\sum_{i=1}^{K} u_i^2 \int_{t_i}^{T} \left[\sum_{n=0}^{N} h_n (v_{i-1}(t))^2 \right] dt + \frac{1}{V_g^2}\sum_{i=1}^{K} u_i^2 \int_{t_{i-1}}^{t_i} \left[\sum_{n=0}^{N} h_n (f_{i-1}(t))^2 \right] dt \\
& + \frac{2}{V_g^2}\sum_{i=1}^{K}\sum_{j=i+1}^{K} u_i u_j \int_{t_j}^{T} \left[\sum_{n=0}^{N} h_n v_{i-1}(t) v_{j-1}(t) \right] dt \\
& + \frac{2}{V_g^2}\sum_{i=1}^{K}\sum_{j=i+1}^{K} u_i u_j \int_{t_{j-1}}^{t_j} \left[\sum_{n=0}^{N} h_n v_{i-1}(t) f_{j-1}(t) \right] dt \\
& - \frac{2x_{\text{ref}}}{V_g}\sum_{i=1}^{K} u_i \int_{t_i}^{T} v_{i-1}^1(t) dt - \frac{2x_{\text{ref}}}{V_g}\sum_{i=1}^{K} u_i \int_{t_{i-1}}^{t_i} f_{i-1}^1(t) dt.
\end{aligned}
$$

Returning back to the formulation of the cost functional in the quadratic form

$$
J_0(u) = u^T C u + D u, \tag{5.75}
$$

the elements d_i, $i = 1, \ldots, K$ of the vector D are given by

$$d_i = \frac{2}{V_g} \int_{t_i}^{T} \left[\sum_{n=0}^{N} h_n q_h^{n+1}(t) v_{i-1}(t) \right] dt + \frac{2}{V_g} \int_{t_{i-1}}^{t_i} \left[\sum_{n=0}^{N} h_n q_h^{n+1}(t) f_{i-1}(t) \right] dt + \frac{2x_{\text{ref}}}{V_g} \int_{t_i}^{T} v_{i-1}^1(t) dt - \frac{2x_{\text{ref}}}{V_g} \int_{t_{i-1}}^{t_i} f_{i-1}^1(t) dt.$$

The diagonal elements c_{ii}, $i = 1, \ldots, K$ of the matrix C are given by

$$c_{ii} = \frac{1}{V_g^2} \int_{t_i}^{T} \left[\sum_{n=0}^{N} h_n (v_{i-1}(t))^2 \right] dt + \frac{1}{V_g^2} \int_{t_{i-1}}^{t_i} \left[\sum_{n=0}^{N} h_n (f_{i-1}(t))^2 \right] dt, \tag{5.76}$$

and the off diagonal elements c_{ij}, $i, j = 1, \ldots, K$ with $j > i$ are given by

$$c_{ij} = \frac{2}{V_g^2} \int_{t_j}^{T} \left[\sum_{n=0}^{N} h_n v_{i-1}(t) v_{j-1}(t) \right] dt + \frac{2}{V_g^2} \int_{t_{j-1}}^{t_j} \left[\sum_{n=0}^{N} h_n v_{i-1}(t) f_{j-1}(t) \right] dt. \tag{5.77}$$

The control policy is to be applied in the same way as explained at the end of section 5.4.1.

5.5.3 Parameter inference combined with polynomial chaos expansions

The parameter inference will be based again on SMC methods as in the Monte Carlo version of the algorithm. Assume that k measurements are available, where $k = 1, \ldots, K$ and that one has an equally weighted sample $\{(\boldsymbol{\theta}^m, \frac{1}{M})\}_{m=1}^{M}$ from the posterior $\pi_{\boldsymbol{\Theta}}^k$ after one run of the SMC algorithm. This sample will be used to build the PC expansion of the model parameters $\boldsymbol{\Theta}$ distributed now according to the posterior $\pi_{\boldsymbol{\Theta}}^k$. This means, that after each measurement is given, one builds the transformation

$$\boldsymbol{\Theta} = g_{\boldsymbol{\Theta}}^k(\boldsymbol{\Xi}), \ k = 1, \ldots, K, \tag{5.78}$$

where $\boldsymbol{\Theta} \sim \pi_{\boldsymbol{\Theta}}^k$ and $\boldsymbol{\Xi}$ is the basis random vector. The functions $g_{\boldsymbol{\Theta}}^k$ are expanded in the polynomial basis and are used to update the PC coefficients of the degradation rate $\gamma(t, \boldsymbol{\Theta}) = \gamma(t, g_{\boldsymbol{\Theta}}^k(\boldsymbol{\Theta}))$ and thus update the Galerkin system for the solution $x(t, \boldsymbol{\Xi})$.

5.5.4 Numerical simulations

The performance of the algorithm based on PC approximations is now demonstrated on the example of one in silico individual. The underlying model parameters are fixed to $\boldsymbol{\theta}_{\text{true}} = (2.593369e-02, 3.395887e-04, 7.093037e-02, 4.171830e-04, 2.376361e-02, 1.510564e+02)$, which is the same parameters used for the numerical simulation of the first MC version of the algorithm. The SMC algorithm was based again on the same sample size, namely $M = 5000$ particles for comparison.

The first subfigure in Figure 5.14 shows the time evolution of the glucose concentration obtained by solving the model with the parameters $\boldsymbol{\theta}_{\text{true}}$ along with the sampled data. The second subfigure contains the time evolution of the degradation rate obtained with

the true parameter vector and with the mean of the posterior distribution at the end of the experiment. The computed optimal control policy is also given in the third subfigure. Figures 5.15 and 5.16 show the density estimators of the posterior distributions for each model parameter.

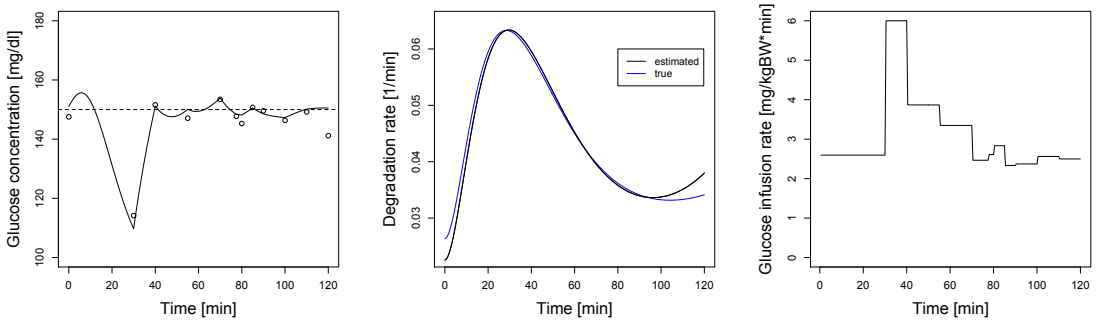

Figure 5.14: First subfigure: Sampled data (circles), solution of the glucose concentration model for $\boldsymbol{\theta}_{\text{true}}$ (solid line) and reference glycemia (dashed line). Second subfigure: true and estimated degradation rate. Third subfigure: computed glucose infusion rate. The parameter vector was fixed at $\boldsymbol{\theta}_{\text{true}} = (2.593369e-02, 3.395887e{-}04, 7.093037e{-}02, 4.171830e{-}04, 2.376361e{-}02, 1.510564e{+}02)$. The solution is based on the polynomial chaos version of the algorithm.

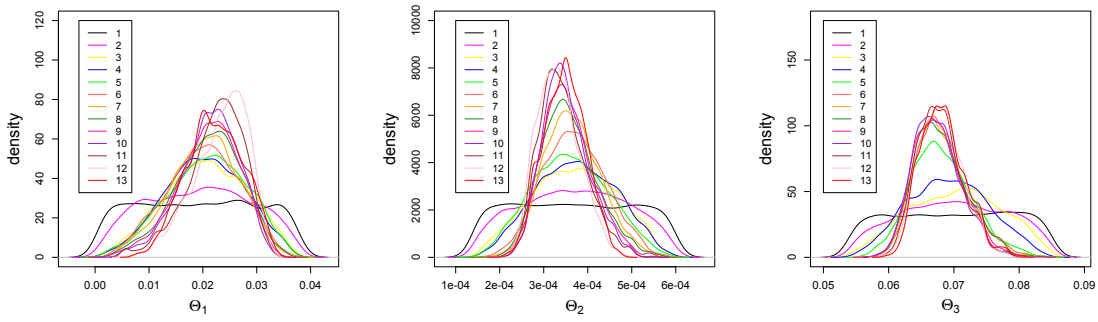

Figure 5.15: Kernel density estimators for parameters Θ_1, Θ_2 and Θ_3. The corresponding components of the true parameter vector were $\theta_{\text{true},1} = 2.593369e{-}02, \theta_{\text{true},2} = 3.395887e-04, \theta_{\text{true},3} = 7.093037e-02$.

The glucose concentration deviates only in the first 40 minutes from the reference value $G_{\text{ref}} = 150$ [mg/dl]. After this point, the glucose concentration stays very close to the target glucemia. This method obviously performs better than the Monte Carlo version of the algorithm, comparing Figures 5.8 and 5.14. This is due to the fact that the PC summarizes the entire distribution of the parameters, while the Monte Carlo algorithm is based on a sample approximation of the expectations. This is known to converge very slow, as it was pointed out in chapter 2. Another important observation is that the control policy based

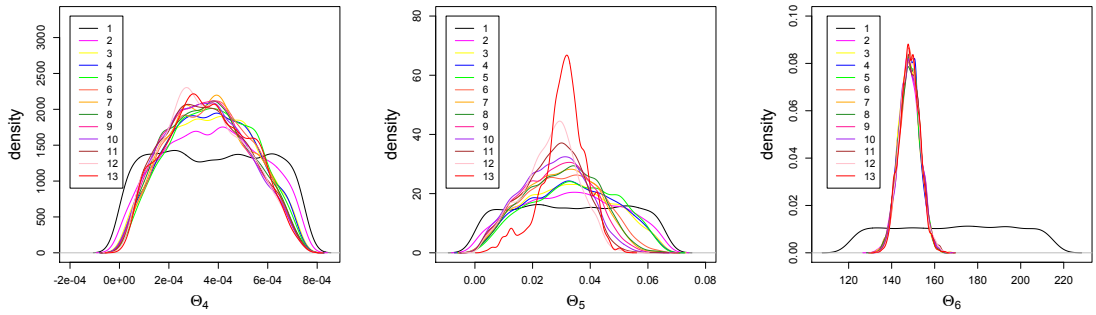

Figure 5.16: Kernel density estimators for parameters Θ_4, Θ_5 and Θ_6. The corresponding components of the true parameter vector were $\theta_{\text{true},4} = 4.171830e{-}04, \theta_{\text{true},5} = 2.376361e - 02, \theta_{\text{true},6} = 1.510564e + 02$.

on PC expansions is much more stable than its Monte Carlo counterpart. This is very important for the application as rapid changes in the glucose infusion rate adds additional stress to the mice, which in turn results in abrupt changes in the glucose dynamics and thus make the control of the system more difficult. Finally, the posterior distributions are centered for all parameters around the nominal value, as it can be seen in Figures 5.15 and 5.16.

REMARK

(i) The preservation of positivity of the PC approximation was not taken into account in the above algorithm, as negative realizations, although making no biological sense, do not lead to a finite time blow up in the linear model under consideration. As seen from the simulations, the existence of negative values does not have any important consequences for the optimization of the test. In addition, the tensor product structure of the multivariate kernels proposed in chapter 4 would result in a very high dimensional Galerkin system for the 6-dimensional model considered here, which in turn would lead to an additional computational effort.

(ii) The algorithm proposed here does not incorporate the PC expansion of the solution in the likelihood. This would require an adaption of the SMC algorithm to take into account the transformation of the random variables corresponding to the posterior and the proposal distribution. This step is not crucial for the running time and the performance of the algorithm, since the dynamics have here the simple form of a one-dimensional ordinary differential equation. The SMC sample is used only to build the PC approximation of the model parameters when they are distributed according to the posterior. As seen in chapter 3, in a 5-dimensional space and for a well-behaved distribution a random sample of size $M = 1000$ can already capture well the marginals densities and the dependence structure. If the method is to be applied to more complicated and/or higher-dimensional systems, the saving of computation time may be significant when using the PC expansion of the solution also in the inference part of the algorithm.

6 Conclusion

Two problems related with applications of polynomial chaos theory in the propagation of uncertainty in dynamical systems were examined in this thesis: the problem of the representation of an arbitrary random variable in terms of a basic random variable and the problem of preservation of positivity in truncated polynomial expansions. The basis of the solution for both cases was the Doob-Dynkin lemma from probability theory, which assures the existence of transformations of the basis random variable to the general variable.

A solution to the former problem was given by constructing a discrete map between random samples from the two variables and combining it with a regression approach in order to finally estimate the PC coefficients. Proofs for convergence were given for the one- and the multi-dimensional case. It was shown that the estimated PC coefficients via the discrete maps converge to the PC coefficients obtained when the transformation appearing in the Doob-Dynkin lemma is assumed to be an optimal map. The estimation of the coefficients directly by their definition performed bad even with an increased sample size. The convergence theorem revealed that only a subsequence convergences. Therefore, the approach via regression and optimal transportation is essential in practice. The theoretical results were verified by numerical simulations on the example of a one-dimensional uniform distribution and of a 5-dimensional Dirichlet distribution.

The second problem was solved by viewing the finite PC expansions as operators on the function space $L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, to which the transformations appearing in the Doob-Dynkin lemma belong. Weight sequences were then introduced in the PC expansions whose role was to enforce the positivity of these operators. The approximation error from the introduction of weights was proven to have possibly a slower convergence rate than the one resulting from the unweighted expansions. This is expected since for finite N the weighted expansions are a worse approximation than the best L^2 classical PC approximations. Examples for suitable weight sequences were stated and the performance of the method was demonstrated on the example of the logistic equation with uncertain initial condition.

Polynomial chaos methods may be used for the design of controllers in systems with parametric uncertainty. They enable to separate the deterministic and the stochastic part of the problem and thus transform the stochastic control problem to a deterministic one, for which a well established theory exists. Both problems considered above may occur in such considerations. When the control design is combined with a sequential Bayesian parameter estimation procedure, one has to find the PC representation of a random variable distributed according to a posterior distribution, a question falling into the first class of problems mentioned above. Even when the prior and posterior distributions have compact supports, the estimated finite PC representation will admit also realizations outside this support. When working with nonlinear systems, these realizations corresponding to the fact that the underlying estimated density function has non zero probability mass outside the support. This may lead to instabilities and failure of the control design.

Bibliography

[1] P. L. ALTMAN, S. D. DITTMER [ED.], *Biology data book*, Federation of american societies for experimental biology, Washington D.C., 1964.

[2] M. ARNST, R. GHANEM, C. SOIZE, Identification of Bayesian posteriors for coefficients of chaos expansions, *Journal of Computational Physics* 229 (2010), pp. 3134-3154.

[3] R. ASKEY, *Orthogonal polynomials and special functions*, SIAM, Philadelphia, 1975.

[4] K. ATKINSON, W. HAN, *Theoretical numerical analysis: a functional analysis framework*, Springer, New York, 2001.

[5] F. AUGUSTIN, A. GILG, M. PAFFRATH, R. RENTROP, U. WEVER, Polynomial chaos for the approximation of uncertainties: chances and limits, *European Journal of Applied Mathematics* 19 (2008), pp. 149-190.

[6] J. E. AYALA, D. P. BRACY, O. P. MCGUINNESS, D. H. WASSERMAN, Considerations in the design of hyperinsulinemic-euglycemic clamps in the conscious mouse, *Diabetes* 55 (2006), pp. 390-397.

[7] J. E. AYALA, V. T. SAMUEL, G. J. MORTON, S. OBICI, C. M. CRONIGER, G. I. SHULMAN, D. H. WASSERMAN, O. P. MCGUINNESS, Standard operating procedures for describing and performing metabolic tests of glucose homeostasis in mice, *Disease Models and Mechanisms* 3 (2010), pp. 525-534.

[8] I. BABUŠKA, F. NOBILE, R. TEMPONE, A stochastic collocation method for elliptic partial differential equations with random input data, *SIAM Journal on Numerical Analysis* 45 (2007), pp. 1005-1034.

[9] R. R. BAHADUR, A note on quantiles in large samples, *The Annals of Mathematical Statistics* 37 (1966), pp. 577-580.

[10] H. BAVINCK, A special class of Jacobi series and some applications, *Journal of Mathematical Analysis and Applications* 37 (1972), pp. 767-797.

[11] H. BAVINCK, On positive convolution operators for Jacobi series, *Tôhoku Mathematical Journal* 24 (1972), pp. 55-69.

[12] H. BERENS, Y. XU, On Bernstein-Durrmeyer polynomials with Jacobi weights, *Approximation Theory and Functional Analysis* 1991, C. K. Chui [ed.], Academic Press, New York, pp. 25-46.

[13] R. N. BERGMAN, Y. Z. IDER, C. R. BOWDEN, C. COBELLI, Quantitative estimation of insulin sensitivity, *American Journal of Physiology* 236 (1979), pp. 667-677.

[14] R. N. BERGMAN, The minimal model: yesterday, today and tomorrow, *The minimal model approach and determination of glucose tolerance*, R. N. Bergman, J.C. Lovejoy [ed.], Louisiana State University Press, Baton Rouge, 1997.

[15] R. N. BERGMAN, Orchestration of glucose homeostasis: from a small acorn to california oak, *Diabetes* 56 (2007), pp. 1489-1501.

[16] J. M. BERNARDO, A. F. M. SMITH, *Bayesian theory*, Wiley, Chichester, 1994.

[17] D. P. BERTSEKAS, A distributed algorithm for the assignment problem, *Laboratory for Information and Decision Systems Working Paper*, MIT, Cambridge, 1979.

[18] D. P. BERTSEKAS, *Linear network optimization: algorithms and codes*, MIT Press, Cambridge, 1991.

[19] D. P. BERTSEKAS, A. NEDIC, A. E. OZDAGLAR, *Convex analysis and optimization*, Athena Scientific, Nashua, 2003.

[20] M. BERVEILLER, B. SUDRET, M. LEMAIRE, Stochastic finite element: a non-intrusive approach by regression, *European Journal of Computational Mechanics* 15 (2006), pp. 81-92.

[21] G. BIRKHOFF, Tres observaciones sobre el algebra lineal, *Universidad Nacional de Tucumn Revista A* 5 (1946), pp. 147-151.

[22] G. BLATMAN, B. SUDRET, Sparse polynomial chaos expansions and adaptive stochastic finite elements using a regression approach, *Comptes Rendus Mécanique* 336 (2008), pp. 518-523.

[23] E. D. BLANCHARD, *Polynomial chaos approaches to parameter estimation and control design for mechanical systems with uncertain parameters*, PhD Thesis, Virginia Polytechnic Institute and State University, 2010.

[24] V. W. BOLIE, Coefficients of normal blood glucose regulation, *Journal of Applied Physiology* 16 (1961), pp. 783-788.

[25] S. BOYD, L. VANDENBERGHE, *Convex optimization*, Cambridge University Press, Cambridge, 2004.

[26] R. BURKHARD, M. D. AMICO, S. MARTELLO, *Assignment problems*, SIAM, Philadelphia, 2009.

[27] R. CAMERON, W. MARTIN, The orthogonal development of nonlinear functionals in series of Fourier-Hermite functionals, *Annals of Mathematics* 28 (1947), pp. 385-392.

[28] C. CANUTO, A. QUARTERONI, Approximation results for orthogonal polynomials in Sobolev spaces, *Mathematics of Computation* 38 (1982), pp. 67-86.

[29] G. CALAFIORE, F. DABBENE [ED.], *Probabilistic and randomized methods for design under uncertainty*, Springer, London, 2006.

[30] R. E. CALFISH, Monte Carlo and quasi Monte Carlo methods, *Acta Numerica* 13 (1998), pp. 1-49.

[31] G. CARLIER, A. GALICHON, F. SANTAMBROGIO, From Knothe's Transport to Brenier's Map and a Continuation Method for Optimal Transport, *SIAM Journal on Mathematical Analysis* 41 (2010), pp. 2554-2576.

[32] D. A. CASTANON, Reverse auction algorithms for assignment problems, *Network flows and matching*, J. S. Johnson, C. C. McGeoch [ed.], American Mathematical Society, Providence, 1992.

[33] S. CHIB, E. GREENBERG, Understanding the Metropolis-Hastings algorithm, *The American Statistician* 49 (1995), pp. 327-335.

[34] C. CHICONE, *Ordinary differential equations with applications*, Springer, New York, 1999.

[35] T. S. CHIHARA, *An introduction to orthogonal polynomials*, Gordon and Breach, New York, 1978.

[36] S. K. CHOI, R. V. GRANDHI, R. A. CANFIELD, Structural reliability under non-Gaussian stochastic behavior, *Computers and Structures* 82 (2004), pp. 1113-1121.

[37] N. CHOPIN, A sequential particle filter method for static models, *Biometrika* 89 (2002), pp. 539-551.

[38] J. A. CUESTA-ALBERTOS, L. RÜSCHENDORF, A. TUERO-DIAZ, Optimal coupling for multivariate distributions and stochastic processes, *Journal of Multivariate Analysis* 46 (1993), pp. 335-361.

[39] S. DAS, R. GHANEM, J. C. SPALL, Asymptotic sampling distribution for polynomial chaos representation from data: a maximum entropy and Fisher information approach, *SIAM Journal on Scientific Computing* 30 (2008), pp. 2207-2234.

[40] S. DAS, R. GHANEM, S. FINETTE, Polynomial chaos representation of spatio-temporal random fields from experimental measurements, *Journal of Computational Physics* 228 (2009), pp. 8726-8751.

[41] B. J. DEBUSSCHERE, H. N. NAJM, P. P. PÉBAY, O. M. KNIO, R. G. GHANEM, O. P. LE MAÎTRE, Numerical challenges in the use of polynomial chaos representations for stochastic processes, *SIAM Journal on Scientific Computing* 26 (2005), pp. 698-719.

[42] R. A. DE FRONZO, J. D. TOBIN, R. ANDRES, Glucose clamp technique: a method for quantifying insulin secretion and resistance, *American Journal of Physiology* 237 (1979), pp. 214-223.

[43] A. DE GAETANO, O. ARINO, Mathematical modeling of the intravenous glucose tolerance test, *Journal of Mathematical Biology* 40 (2000), pp. 136-168.

[44] P. DEL MORAL, A. DOUCET, A. JASRA, Sequential Monte Carlo samplers, *Journal of the Royal Statistical Society B* 68 (2006), pp. 441-436.

[45] C. DESCELIERS, R. GHANEM, C. SOIZE, Maximum likelihood estimation of stochastic chaos representations from experimental data, *International Journal for Numerical Methods in Engineering* 66 (2006), pp. 978-1001.

[46] C. DESCELIERS, C. SOIZE, R. GHANEM, Identification of chaos representations of elastic properties of random media using experimental vibration tests, *Computational Mechanics* 39 (2007), pp. 831-838.

[47] R. A. DE VORE, G. G. LORENTZ, *Constructive approximation*, Springer, Berlin, 1993.

[48] J. L. DOOB, *Stochastic processes*, Wiley, New York, 1953.

[49] A. DOUCET, N. DE FREITAS, N. GORDON [ED.], *Sequential Monte Carlo methods in practice*, Springer, New York, 2001.

[50] G. E. DULLERUD, F. PAGANINI, *A course in robust control theory: a convex approach*, Springer, New York, 2000.

[51] A. DVORETZKY, J. KIEFER, J. WOLFOWITZ, Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator, *The Annals of Mathematical Statistics* 27 (1956), pp. 642-669.

[52] O. G. ERNST, A. MUEGLER, H. J. STARKLOFF, E. ULLMANN, On the convergence of generalized polynomial chaos expansions, *Mathematical Modeling and Numerical Analysis* 46 (2012), pp. 317-339.

[53] J. FISHER, R. BHATTACHARYA, Linear quadratic regulation of systems with stochastic parametric uncertainties, *Automatica* 45 (2009), pp. 2831-2841.

[54] G. GASPER, Linearization of the product of Jacobi polynomials I, *Canadian Journal of Mathematics* 22 (1970), pp. 171-175.

[55] G. GASPER, Linearization of the product of Jacobi polynomials II, *Canadian Journal of Mathematics* 22 (1970), pp. 582-593.

[56] G. GASPER, Banach algebras for Jacobi series and positivity of a kernel, *Annals of Mathematics* 95 (1972), pp. 261-280.

[57] A. GELMAN, J. B. CARLIN, H. S. STERN, D. B. RUBIN, *Bayesian Data Analysis*, Chapman and Hall/CRC, Boca Raton, 2003.

[58] T. GERSTNER, M. GRIEBEL, Numerical integration using sparse grids, *Numerical Algorithms* 18 (1998), pp. 209-232.

[59] R. G. GHANEM, A. DOOSTAN, On the construction and analysis of stochastic models: characterization and propagation of the errors associated with limited data, *Journal of Computational Physics* 217 (2006), pp. 63-81.

[60] R. G. GHANEM, A. DOOSTAN, J. RED-HORSE, A probabilistic construction of model validation, *Computer Methods in Applied Mechanics and Engineering* 197 (2008), pp. 2585-2595.

[61] R. GHANEM, P. D. SPANOS, *Stochastic finite elements: a spectral approach*, Springer, New York, 1991.

[62] W. R. GILKS, S. RICHARDSON, D. SPIEGELHALTER [ED.], *Markov chain Monte Carlo in practice*, Chapman and Hall/CRC, Boca Raton, 1995.

[63] W. R. GILKS, C. BERZUINI, Following a moving target: Monte Carlo inference for dynamic Bayesian models, *Journal of the Royal Statistical Society B* 63 (2001), pp. 127-146.

[64] N. J. GORDON, D. J. SALMOND, A. F. M. SMITH, Novel approach to nonlinear/non-Gaussian Bayesian state estimation, *IEE Proceedings Radar and Signal Processing* 140 (1993), pp. 107-113.

[65] D. GOTTLIEB, S. ORSZAG, *Numerical analysis of spectral methods: theory and applications*, SIAM, Philadelphia, 1997.

[66] P. GUTTORP, *Stochastic modeling of scientific data*, Chapman and Hall/CRC, London, 1995.

[67] R. HARDIN, N. SLOANE, A new approach to the construction of optimal designs, *Journal of Statistical Planning and Inference* 37 (1993), pp. 339-369.

[68] E. HEWITT, K. STROMBERG, *Real and abstract analysis*, Springer, New York, 1969.

[69] F. HEISS, V. WINSCHEL, Likelihood approximation ny numerical integration on sparse grids, *Journal of Econometrics* 144 (2008), pp. 62-80.

[70] JAN S. HESTHAVEN, S. GOTTLIEB, D. GOTTLIEB, *Spectral methods for time-dependent problems*, Cambridge University Press, Cambridge, 2007.

[71] F. S. HOVER, M. S. TRIANTAFYLLOU, Application of polynomial chaos in stability and control, *Automatica* 42 (2006), pp. 789-795.

[72] S. JANSON, *Gaussian Hilbert spaces*, Cambridge University Press, Cambridge, 1997.

[73] H. JEFFREYS, An invariant form for the prior probability in estimation problems, *Proceedings of the Royal Society A* 186 (1946), pp. 453-461.

[74] H. JEFFREYS, *Theory of probability*, Oxford University Press, London, 1961.

[75] J. KAIPIO, E. SOMERSALO, *Statistical and computational inverse problems*, Springer, New York, 2005.

[76] R. E. KASS, L. WASSERMAN, Formal rules for selecting prior distributions: a review and annotated bibliography, *Journal of the American Statistical Association* 91 (1996), pp. 343-1370.

[77] A. KIRSCH, *An introduction to the mathematical theory of inverse problems*, Springer, New York, 1996.

[78] A. KLENKE, *Probability theory: a comprehensive course*, Springer, London, 2008.

[79] K. PETRAS, Smolyak cubature of given polynomial degree with few nodes for increasing dimension, *Numerische Mathematik* 93 (2003), pp. 729-753.

[80] T. P. LAINE, The product formula and convolution structure for the generalized Chebyshev polynomials, *SIAM Journal on Mathematical Analysis* 11 (1980), pp. 133-146.

[81] R. LASSER, J. OBERMAIER, On Fejér means with respect to orthogonal polynomials: a hypergroup theoretic approach, *Journal of Approximation Theory*, special Volume Progress in Approximation Theory (1991), pp. 551-565.

[82] R. LASSER, J. OBERMAIER, On the convergence of weighted Fourier expansions, *Acta Scientiarum Mathematicarum* 61 (1995), pp. 345-355.

[83] R. LASSER, D. H. MACHE, J. OBERMAIER, On approximation methods by using orthogonal polynomial expansions, *Advanced Problems in Constructive Approximation*, M. D. Buhmann, D. H. Mache [ed.], Birkhäuser, Basel, 2003.

[84] O. P. LE MAÎTRE, H. NAJM, R. GHANEM, O. KNIO, Uncertainty propagation using Wiener-Haar expansions, *Journal of Computational Physics* 197 (2004), pp. 28-57.

[85] O. P. LE MAÎTRE, H. N. NAJM, P. P. PÉBAY, R. G. GHANEM, O. M. KNIO, Multiresolution-analysis scheme for uncertainty quantification in chemical systems, *SIAM Journal on Scientific Computing* 29 (2007), pp. 864-889.

[86] O. P. LE MAÎTRE, O. M. KNIO, *Spectral methods for uncertainty quantification*, Springer, Dordrecht, 2010.

[87] J. S. LIU, R. CHEN, Sequential Monte Carlo methods for dynamic systems, *Journal of the American Statistical Association* 93 (1998), pp. 1032-1044.

[88] M. LOÈVE, *Probability theory*, Springer Verlag, New York, 1977.

[89] A. MAKROGLOU, J. LI, Y. KUANG, Mathematical models and software tools for the glucose-insulin regulatory system and diabetes: an overview, *Applied Numerical Mathematics* 56 (2006), pp. 559-573.

[90] A. MAKROGLOU, I. KARAOUSTAS, J. LI, Y. KUANG, Delay differential equation models in diabetes modeling: a review, *EOLSS encyclopedia*, Oxford, 2011.

[91] A. MARI, Mathematical modeling in glucose metabolism and insulin secretion, *Current Opinion in Clinical Nutrition and Metabolic Care* 5 (2002), pp. 495-501.

[92] Y. M. MARZOUK, H. N. NAJM, L. A. RAHN, Stochastic spectral methods for efficient Bayesian solution of inverse problems, *Journal of Computational Physics* 224 (2007), pp. 560-586.

[93] Y. M. MARZOUK, H. N. NAJM, Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems, *Journal of Computational Physics* 228 (2009), pp. 1862-1902.

[94] P. MASSART, The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality, *Annals of Probability* 18 (1990), pp. 1269-1283.

[95] M. MCKAY, W. CONOVER, R. BECKMAN, A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, *Technometrics* 21 (1979), pp. 239-245.

[96] A. MONTI, F. PONCI, T. LOVETT, A polynomial chaos theory approach to the control design of a power converter, *Proceedings of the 35th Annual IEEE Power Electronics Specialists Conference* 2004, pp. 4809-4813.

[97] T. A. MOSELHY, Y. M. MARZOUK, Bayesian inference with optimal maps, *Journal of Computational Physics* 231 (2012), pp. 7815-7850.

[98] F. NOBILE, R. TEMPONE, C. WEBSTER, A sparse grid stochastic collocation method for partial differential equations with random input data, *SIAM Journal on Numerical Analysis* 46 (2008), pp. 2411-2442.

[99] J. OBERMAIER, The de la Vallée Poussin kernel for orthogonal polynomial systems, *Analysis* 21 (2001), pp. 277-288.

[100] J. OBERMAIER, A modified Fejér and Jackson summability method with respect to orthogonal polynomials, *Journal of Approximation Theory* 163 (2011), pp. 554-567.

[101] G. PACINI, R. N. BERGMAN, Minmod: a computer program to calculate insulin sensitivity and pancreatic responsivity from the frequently sampled intravenous glucose tolerance test, *Computer Methods and Programs in Biomedicine* 23 (1986), pp. 113-122.

[102] Y.-B. PENG, R. GHANEM, J. LI, Polynomial chaos expansions for optimal control of nonlinear random oscillators, *Journal of Sound and Vibration* 329 (2010), pp. 3660-3678.

[103] U. PICCHINI, A. DE GAETANO, S. PANUNZI, S. DITLEVSEN, G. MINGRONE, A mathematical model of the euglycemic hyperinsulinemic clamp, *Theoretical Biology and Medical Modeling* 2 (2005).

[104] U. PICCHINI, S. DITLEVSEN, A. DE GAETANO, Modeling the euglycemic hyperinsulinemic clamp by stochastic differential equations, *Journal of Mathematical Biology* 53 (2006), pp. 771-796.

[105] F. PULKELSHEIM, *Optimal design of experiments*, SIAM, Philadelphia, 2006.

[106] M. T. REAGAN, H. N. NAJM, B. J. DEBUSSCHERE, O. P. LE MAÎTRE, O. M. KNIO, R. G. GHANEM, Spectral stochastic uncertainty quantification in chemical systems, *Combustion Theory and Modeling* 8 (2004), pp. 607-632.

[107] S. REICH, A non-parametric ensemble transform method for Bayesian inference, *SIAM Journal on Scientific Computing* 35 (2013), pp. 2013-2024.

[108] C. P. ROBERT, *The Bayesian choice: from decision-theoretic foundations to computational implementation*, Springer, New York, 2007.

[109] L. RUESCHENDORF, S. T. RACHEV, A characterization of random variables with minimum L^2 distance, *Journal of Multivariate Analysis* 32 (1990), pp. 48-54.

[110] M. ROSENBLATT, Remarks on multivariate transformation, *The Annals of Mathematical Statistics* 23 (1952), pp. 470-472.

[111] R. J. SERFLING, *Approximation theorems for mathematical statistics*, Wiley, New York, 1980.

[112] A. L. SHIRYAEV, *Probability*, Springer, New York, 1996.

[113] S. SMOLYAK, Quadrature and interpolation formulas for tensor products of certain classes of functions, *Doklady Akademii Nauk SSSR* 4 (1963), pp. 240-243.

[114] C. SOIZE, Construction of probability distributions in high dimension using the maximum entropy principle: applications to stochastic processes, random fields and random matrices, *International Journal for Numerical Methods in Engineering* 76 (2008), pp. 1583-1611.

[115] C. SOIZE, Identification of high-dimensional polynomial chaos expansions with random coefficients for non-Gaussian tensor-valued random fields using partial and limited experimental data, *Computer Methods in Applied Mechanics and Engineering* 199 (2010), pp. 2150-2164.

[116] J. STOER, R. BULIRSCH, *Introduction to numerical analysis*, Springer, Berlin, 1993.

[117] G. SZEGÖ, *Orthogonal polynomials*, American Mathematical Society, Providence, 1959.

[118] B. A. TEMPLETON, *A polynomial chaos approach to control design*, PhD Thesis, Virginia Polytechnic Institute and State University, 2009.

[119] R. TEMPO, G. CALAFIORE, F. DABBENE, *Randomized algorithms for analysis and control of uncertain systems*, Springer, London, 2005.

[120] D. BERTSIMAS, J. N. TSITSIKLIS, *Introduction to linear optimization*, Athena Scientific, Nashua, 1997.

[121] A. W. VAN DER VAART, *Asymptotic statistics*, Cambridge University Press, Cambridge, 1998.

[122] P. F. VERHULST, Recherches mathématiques sur la loi d' accroissement de la population, *Mémoires de l' académie royale des sciences et belles lettres de Bruxelles* 18 (1845), pp. 1-38.

[123] C. VILLANI, *Topics in optimal transportation*, American Mathematical Society, Providence, 2003.

[124] C. VILLANI, *Optimal transport: old and new*, Springer, Berlin, 2009.

[125] X. WAN, G. E. KARNIADAKIS, Multi-element generalized polynomial chaos for arbitrary probability measures, *SIAM Journal on Scientific Computing* 28 (2006), pp. 901-928.

[126] N. WIENER, The homogenous chaos, *The American Journal of Mathematics* 60 (1938), pp. 897-936.

[127] S. WILD, G. ROGLIC, A. GREEN, R. SICREE, H. KING, Global prevalence of diabetes: estimates for the year 2000 and projections for 2030, *Diabetes Care* 27 (2004), pp. 1047-1053.

[128] D. XIU, G. E. KARNIADAKIS, The Wiener-Askey polynomial chaos for stochastic differential equations, *SIAM Journal on Scientific Computing* 24 (2002), pp. 619-644.

[129] D. XIU, J. S. HESTHAVEN, High-order collocation methods for differential equations with random inputs, *SIAM Journal on Scientific Computing* 27 (2005), pp. 1118-1139.

[130] D. XIU, Efficient collocation approach for parametric uncertainty analysis, *Communications in Computational Physics* 2 (2007), pp. 293-309.

[131] D. XIU, *Numerical methods for stochastic computations: a spectral method approach*, Princeton University Press, Princeton, 2010.

[132] K. ZHOU, J. C. DOYLE, *Essentials of robust control*, Prentice Hall, Upper Saddle River, 1997.

[133] N. ZOHRABI, H. R. MOMENI, H. ZAKERI, Markov jump modeling and control of stress effects on the blood glucose regulation system, *Proceedings of the 20th Iranian Conference on Electric Engineering* 2012, pp. 985-989.