University of Southampton
Faculty of Social, Human and Mathematical Sciences
Department of Mathematical Sciences

Global Optimisation of Noisy Grey-Box Functions with Financial Applications

by Dirk Banholzer

A dissertation submitted for the degree of
Doctor of Philosophy

October 2018

Abstract

Financial derivatives of both plain vanilla and exotic type are at the core of today's financial industry. For the valuation of these derivatives, mathematical pricing models are used that rely on different approaches such as (semi-)analytical transform methods, PDE approximations or Monte Carlo simulations. The calibration of the models to market prices, i.e. the estimation of appropriate model parameters, is a crucial procedure for making them applicable to real markets. Due to inherent complexity of the models, this typically results in a nonconvex optimisation problem that is hard to solve, thus requiring advanced techniques.

In this thesis, we study the general case of financial model calibration where model prices are approximated by standard Monte Carlo methods. We distinguish between two possibilities on how to employ Monte Carlo estimators along a calibration procedure: the simpler sample average approximation (SAA) strategy, which uses the same random sample for all function evaluations, and the more sophisticated variable sample average approximation (VSAA) strategy, which for each evaluation uses a different random sample with variable size. Both strategies have in common that they lead to (not prohibitively) expensive optimisation procedures providing approximating solutions to the original but unknown optimisation problem. Yet, whereas the former strategy results in a self-contained deterministic problem instance that may be fully solved by a suitable algorithm, the latter has to be considered together with a sequential sampling method that incorporates the strategy to select new evaluation points, which amounts to minimising a noisy objective function.

For both strategies, we initially establish essential convergence properties for the (optimal) estimators in the almost sure sense. Specifically, in the case of the SAA strategy, we complement the well-established strong consistency of optimal estimators with their almost sure rates of convergence. This, in turn, allows to draw several useful conclusions on their asymptotic bias and other notions of convergence. In the case of the VSAA strategy, we give conditions for the strong uniform consistency of the objective function estimators and provide corresponding uniform sample path bounds. Both results may be used to show convergence of a sequential sampling method adopting the VSAA scheme.

We then address the global optimisation within both considered procedures, and first present a novel modification to Gutmann's radial basis function (RBF) method for expensive and deterministic objective functions that is more suited for deterministic calibration problems. This modification exploits the particular data-fitting structure of these problems and additionally enhances the inherent search mechanism of the original method by an extended local search. We show convergence of the modified method and demonstrate its effectiveness on relevant test problems and by calibrating the Hull-White interest rate model under the SAA strategy in a real-world setting. Moreover, as the method may be applied equally well to similar data-fitting problems that are not necessarily expensive, we also demonstrate its practicability by fitting the Nelson-Siegel and Svensson models to market zero rates.

Based on the RBF method, we further present a novel method for the global optimisation of expensive and noisy objective functions, where the level of noise is controlled by means of error bounds. The method uses a regularised least-squares criterion to construct suitable radial basis function approximants, which are then also used to determine new sample points in a similar manner as the original RBF method. We provide convergence of the method, albeit under some simplifying assumption on the error bounds, and evaluate its applicability on relevant test problems and by calibrating the Hull-White interest rate model under the VSAA strategy.

Declaration of Authorship

I, ... , declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

Title of thesis:

....

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University.
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
3. Where I have consulted the published work of others, this is always clearly attributed.
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
5. I have acknowledged all main sources of help.
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.
7. None of this work has been published before submission.

Signed: ....

Date: ....

Acknowledgements

First of all, I would like to express my deep gratitude to my supervisors Professor Dr Jörg Fliege and Professor Dr Ralf Werner for having given me the opportunity to work on such an interesting topic eventually leading to this thesis. I am very thankful to both for having introduced me to the field of (global) optimisation, for having taught me how do research, and for their great commitment and utmost patience throughout. Their professional guidance and valuable advice have always been very encouraging, which has made things a lot more enjoyable during my research.

I am particularly grateful to the Operational Research Group within Mathematical Sciences at the University of Southampton for providing a pleasant research environment. A special mentioning here goes to Tri-Dung Nguyen and Edwin Tye for numerous fruitful discussions on topics of any kind, which has certainly contributed to my understanding over the past years. I also would like to thank Max Hughes, Christian Drescher and Jonas Schwinn for their accommodating support and insightful conversations during my research stays in Munich and Augsburg, and all those that have made these stays possible, after all.

I further would like to acknowledge the financial support of the DEVnet GmbH and the Engineering and Physical Sciences Research Council of the UK during the main part of my research. This has definitely helped in having a more pleasing PhD life and allowed me to visit several interesting workshops and conferences, among others.

Finally, I would like to thank my family and friends for their continuous support and encouragement throughout this adventurous journey. Without their help, this thesis would not have been written.

Contents

Chapter 1

Introduction

Over the past few decades, financial derivatives of both plain vanilla and exotic type have gained increasing importance and become one of the most essential products in the financial industry. They are heavily used by financial institutions and other entities for a variety of different purposes, mainly including the hedging and mitigation of risks arising in their financial positions, but also for speculation and arbitrage. In plain terms, a financial derivative can be defined as a security whose value depends on, or is derived from, the values of other more basic securities (often referred to as underlying assets or risk factors), such as stocks, commodities, interest/exchange rates or indices. To the most common financial derivatives belong forwards, futures and options.

For the fair valuation of financial derivatives, sophisticated mathematical pricing models are used throughout the industry, which describe the evolution of the underlying risk factors over time by means of stochastic processes and in turn lead to rigorous pricing formulas by the basic principles of derivatives pricing. Common to all pricing models is that they depend on a set of model-specific parameters to fully define the dynamics of each model and to produce a wide range of different model prices. Yet, to make the pricing models relevant to real financial markets and applicable for pricing, risk management and trading purposes, their parameters need to be estimated such that the computed model prices best match given market prices for a set of liquidly traded benchmark instruments[1]. This estimation procedure, which is commonly referred to as *calibration*, is typically set up as a numerical optimisation problem and as such has become an essential step in the pricing of financial derivatives. For a succinct illustration of a standard calibration procedure, see Figure 1.1.

Depending on the use of the calibrated model, there are a number of different ways in which a calibration procedure can be mounted. The main difference between these variants lies in the formulation of the objective function, the definition of a corresponding parameter space and the choice of an adequate algorithm to numerically solve the resulting optimisa-

[1] The set of instruments may refer to the same type of derivative with different contractual features, such as maturity, strike price and the like, but also to distinct types with different contractual features.

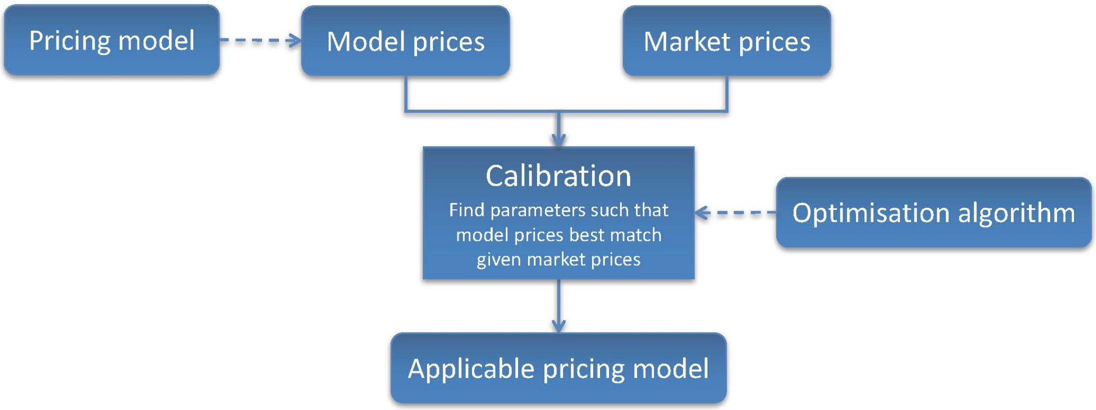


Figure 1.1: Calibration of a pricing model through a given set of benchmark instruments.

tion problem. Since most pricing models are inherently complex, the associated calibration problem typically results in a highly *nonlinear nonconvex data-fitting problem*, which is one of the harder optimisation problems to solve. There are often only a few parameter combinations to which the underlying model reacts quite strongly, whereas all other combinations (usually involving lower objective function values) have a weak impact. Further on, multiple local minima may exist, which in theory requires the use of global optimisation techniques to ensure convergence to an optimal solution.

A crucial role in solving calibration problems is thus played by the algorithm applied to determine appropriate model parameters. Despite the fact that most objective functions are nonconvex, they are frequently tackled by optimisation techniques designed to find a locally optimal solution only. Since these techniques are sensitive to the chosen initial value, 'globalising' them by starting from multiple initial values, or strategies alike, is a common remedy to increase the chances of finding a global optimum. Yet, no guarantee of convergence exists in these cases. By contrast, global techniques are obviously more suited to finding global solutions, but usually demand considerably more function evaluations.

Along with the decision of whether to aim at finding a global optimum or not, there are further aspects that should be taken into consideration when selecting a suitable algorithm for calibration. Firstly, the computational intensity of the pricing routine and thus of the objective function is of relevant importance and should be dealt with by the applied method in an adequate manner. Since the models rely on different pricing methodologies with distinct characteristics, the objective functions may turn out to be expensive to evaluate and therefore demand different optimisation techniques than others. Irrespective of the computational intensity of the pricing model, however, it is expected that a calibration is carried out fast and efficiently, within at most a few minutes of computation, due to their subsequent usage.

Moreover, depending on the type of pricing model, the evaluated objective function values can be exact or inaccurate as resulting, for instance, from numerical approximations or stochastic simulation. While the former are clearly easier to handle and consequently dealt with by most optimisation algorithms, efficient methods focusing on the latter are rare.

Closely related to both previous issues is the availability of gradient information and its computational cost. For most pricing routines, gradients and Hessians cannot be calculated analytically and must therefore be approximated numerically, e.g. by finite differences or automatic differentiation. This, however, may require an excessive amount of computational resources or may become unreliable in the presence of noise. Similarly, Lipschitz constants are usually not known a priori and must be estimated during the optimisation if made use of. Yet, the highly nonlinear structure of the objective function, expensive or noisy objective function evaluations may acutely complicate their estimation, conceivably leading to an under- or overestimation of the Lipschitz constants.

Eventually, a suitable solver ideally also takes into account that pricing models are typically recalibrated on a frequent basis in order to keep the model prices consistent with their market observed counterparts. This thus amounts to solving a series of calibration problems, where a stable and robust behaviour of the calibrated parameters as a function of the input data is expected.

Although often not stated explicitly, the calibration of derivative pricing models is typically only carried out for those models that render analytical valuation formulas for liquidly traded instruments, the calibrated parameters of these models then being also used to evaluate more involved derivatives that do not allow for a valuation in closed-form. As opposed to this rather ad-hoc procedure, however, we pursue a more direct approach in this thesis and aim at considering also those models for calibration that do not yield explicit pricing formulas. Specifically, as a general approach to derivative pricing relies on standard Monte Carlo methods, we shall investigate the use of this approach within model calibration.

In view of this objective, the remainder of this introductory chapter is devoted to set the main stage for this thesis, providing the reader with a more detailed description on the calibration of pricing models and the related optimisation. To this end, we begin in Section 1.1 by formulating the general calibration setup which is of fundamental interest throughout the main part of this exposition. Considering the optimisation problems resulting from this setup, we then revise in Section 1.2 existing approaches for the global optimisation of a nonconvex objective function and discuss their applicability within our calibration setup. Eventually, in Section 1.3, we present the outline of this thesis, including our main contributions.

1.1 Financial Model Calibration

To derive the underlying calibration problems, we first present the basic setup in case a pricing model admits a closed-form expression for its implied derivative prices. We then motivate the use of Monte Carlo pricing methods within this setup, which leads us to two strategies for approximately solving the original calibration problem. These will be at the core of this exposition.

1.1.1 General Setup of Calibration Procedures

Given a parameterised pricing model and market prices for a set of different calibration instruments, the calibration of this model can be formulated in general terms as

$$\min_{x \in \mathcal{X}} \Big\{ f(x) := g\big(\Pi(x) - C^{\text{mkt}}\big) \Big\}, \tag{1.1}$$

where $\mathcal{X} \subset \mathbb{R}^d$ is a nonempty finite-dimensional set of feasible parameters, and $g : \mathbb{R}^l \to \mathbb{R}_{\geq 0}$ denotes a nonnegative continuous function that measures the discrepancy between the vector-valued pricing function of calibration instruments $\Pi : \mathcal{X} \to \mathbb{R}^l$ at any $x \in \mathcal{X}$ and the corresponding market prices[2] $C^{\text{mkt}} \in \mathbb{R}^l$. Throughout this thesis, we assume that $\mathcal{X}$ is compact such that the problem (1.1) is well-posed and also grants a solution if the model prices $\Pi(x)$ depend continuously on the parameter x. Since the parameters of most financial models have an intuitive meaning, the constraints for $\mathcal{X}$ may frequently be inferred from the interpretations the parameters have, provided that the applied algorithm is capable of handling them. Usually, the set $\mathcal{X}$ is then described by (deterministic) linear equality and inequality constraints.

Regarding problem (1.1), we can tacitly assume that it constitutes a low-dimensional optimisation problem, as pricing models typically depend only on a few parameters whose number rarely exceeds a single figure (otherwise, they are likely to be over-specified). However, to guarantee that it also defines an overdetermined fitting problem with more observations that unknowns, we let $l > d$ always holds, i.e. the number of calibration instruments is greater than the dimension of the problem.

If not further strengthened where required, the multivariate function g is supposed to be nonnegative and continuous, thus representing minimal requirements for a function to measure the discrepancy between a model and some observed data. Writing the objective function f as a composition of the residual vector $\Pi(x) - C^{\text{mkt}}$ and some function g allows to consider the pricing level and the subsequent transformation into an objective function separately, for a range of different specifications of g. Most generally, one may consider for g any vector p-norm $\|y\|_p := (\sum_{i=1}^{l} |y_i|^p)^{1/p}$, for $1 \leq p \leq \infty$, or any continuous monotone transformation thereof. By far the most common specification is obtained by the squared 2-norm, i.e. $g(y) = \|y\|_2^2 = \sum_{i=1}^{l} y_i^2$, $y \in \mathbb{R}^l$, which leads to a (nonlinear) *least-squares optimisation problem.* This formulation will be applied in all practical sections of this thesis. Other popular formulations include the 1-norm $g(y) = \|y\|_1 = \sum_{i=1}^{l} |y_i|$ or the ∞-norm $g(y) = \|y\|_\infty = \max\{|y_1|, \ldots, |y_l|\}$. In any case, as it is more beneficial to work with, we also consider in Chapter 4 the transformation g as a univariate nonnegative continuous function from $\mathbb{R}$ into $\mathbb{R}_{\geq 0}$ and then rewrite the corresponding objective function in (1.1) as the sum

$$f(x) = \sum_{i=1}^{l} g\big(\Pi_i(x) - C_i^{\text{mkt}}\big),$$

[2] For the sake of generality, we do not restrict the pricing function Π and the corresponding market prices C^{mkt} to be positive. Hence, by some abuse of notation, this formulation also allows for other setups not necessarily referring to the calibration of pricing models.

for residual components $\Pi_i(x) - C_i^{\mathrm{mkt}}$. Except the ∞-norm, this covers all relevant specifications of the multivariate case (possibly by use of some monotone transformation).

Quite naturally, positive weights may additionally be included in the objective function f if market prices are not assumed to be equally valid. In such a case, the residual vector in (1.1) may then be rewritten as $W_f(\Pi(x) - C^{\mathrm{mkt}})$ for some diagonal matrix W_f of positive weights. However, since the weights can always be absorbed into the model and the corresponding market prices, their inclusion does not present any further difficulties in the calibration procedure. For reasons of brevity and simplicity, we therefore assume that all weights are set equal to one in all theoretical and practical analysis. Further material on the specification of the objective functions and their usage can be found, for instance, in Hirsa [2012], Section 7.1, or Kienitz and Wetterau [2012], Section 9.1.

Even though we are mainly concerned in this thesis with optimisation problems that arise from the calibration of derivative pricing models, the above setup also applies to other data-fitting problems with similar characteristics. In particular, this holds for the fitting of yield curve models from the Nelson-Siegel family, cf. Section A.3 in the appendix. Yet, whereas the pricing function of financial derivatives may be put as an expected value which then leads to a stochastic programming problem, as described hereinafter, this representation is in general not viable in other data-fitting problems.

1.1.2 Derivative Pricing

By the theory of no-arbitrage pricing, the function Π can be further expressed as the expected value of the discounted payoffs of all involved calibration instruments, see, e.g., the textbooks of Björk [2009] or Musiela and Rutkowski [2005] for further background. This renders the calibration problem (1.1) a (transformed) *stochastic programming problem*, in which Π is then most conveniently written in the compact form

$$\Pi(x) := \mathbb{E}^{\mathbb{Q}}\big[h(x, Z)\big], \tag{1.2}$$

where $\mathbb{Q}$ denotes an equivalent martingale measure[3] and $h : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}^l$ describes the discounted payoff vector of the calibration instruments as a function of the model parameters $x \in \mathcal{X}$ and a random vector Z. In particular, the distribution of Z, denoted by $\mathbb{Q}^Z$, is assumed to be known and supported on a set $\mathcal{Z} \subset \mathbb{R}^{d_Z}$. Further, for Π to be well-defined, it shall also be assumed for every $x \in \mathcal{X}$ that the random vector $h(x, \cdot)$ is measurable with respect to the Borel σ-algebras $\mathcal{B}(\mathcal{Z})$ and $\mathcal{B}(\mathbb{R}^l)$, i.e. that h is a random function, and that it is $\mathbb{Q}^Z$-integrable.

The discounted payoff function h in (1.2) may typically be viewed as the result of a series of transformations, into which the randomness enters at the lowest level in form of Z.

[3] That is a probability measure which is equivalent to the objective measure $\mathbb{P}$ of the underlying probability space and under which the discounted price processes associated to the respective numeraire are martingales.

Most commonly, the latter may be thought of as a vector of independent and uniformly or standard normally distributed random variables, which is then transformed together with the parameter x of the employed pricing model to a random vector with a more complicated distribution. This transformation usually corresponds to constructing a path for the evolution of the underlying risk factors, where each time point affecting the payoff of any calibration instrument is simulated. Eventually, from the constructed sample path, the vector of discounted payoffs $h(x, Z)$ of all calibration instruments is calculated.

As an expected value of the discounted payoffs, the computability of the pricing function Π via (1.2) depends on the distribution of the derivative payoff $h(\,\cdot\,, Z)$ under the applied pricing model. If the structure of the payoff is reasonably simple, the expected value of $h(x, Z)$ can typically be calculated explicitly for any $x \in \mathcal{X}$, either directly or by some transform method, leading to (semi-)analytical pricing formulas. Such a case is given, for example, if swaptions or caps/floors are priced by the Hull-White one-factor interest rate model, cf. Section A.2. Unfortunately, however, the structure of the payoff vector turns out to be fairly complex for a variety of different pricing models, such that $h(x, Z)$ can indeed be evaluated at any $x \in \mathcal{X}$ but its exact distribution under any measure $\mathbb{Q}$ remains unknown. Some numerical method to approximate the unknown function Π is then required.

Two possible options to estimate the price of a single derivative instrument are *multinomial lattice (tree) methods* (e.g., Hull [2017], Chapters 13 and 21) and *finite-difference methods* (e.g., Hirsa [2012], Chapters 4 and 5), which work by approximating the stochastic processes of the underlying risk factors and by numerically solving the associated pricing PDEs/PIDEs, respectively. They both are able to handle similar kinds of derivative pricing problems, which include many popular pricing models and common European-style derivatives, as well as American-style derivatives with a path-dependent payoff (which in the case of PDEs/PIDEs, though, become more difficult to solve as free boundary problems). Yet, both have the drawback that their practicalness for pricing derivative instruments is restricted to models in which the stochastic processes driving the evolution of the underlying risk factors are Markov and of low dimension. In case the processes are Markov with a high-dimensional scheme or non-Markov, or the payoff structure is too complex, then these methods soon become unworkable. In particular, this holds for a calibration where an entire portfolio of instruments has to be evaluated to obtain an approximation to the pricing vector $\Pi(x)$ for any $x \in \mathcal{X}$, and thus typically requires a feasible high-dimensional approximation scheme.

In any case, though, *standard Monte Carlo methods* provide the simplest general technique to approximate the pricing function Π, see, e.g., Glasserman [2003] or Hirsa [2012], Chapter 6. They basically allow for pricing under any model that can be described by a set of random variables with a known distribution and for any evaluable derivative payoff, thus also applying in most situations in which any of the above methods is unrealistic. Above all, they do not suffer from any dimensionality issues[4], which makes them a well-suited approx-

[4] Note that since the effectiveness of quasi-Monte Carlo methods deteriorates with increasing dimension

imation technique for our calibration purposes. Nevertheless, one has to bear in mind that they typically also represent the most expensive pricing methods.

To approximate the unknown expected value in (1.2) at any $x \in \mathcal{X}$, Monte Carlo methods proceed by drawing a sample of i.i.d. random vectors $Z_1, \ldots, Z_N$, $N \in \mathbb{N}$, from the same distribution as Z, and taking the sample average

$$\widehat{\Pi}_N(x; Z_1, \ldots, Z_N) := \frac{1}{N} \sum_{i=1}^{N} h(x, Z_i), \tag{1.3}$$

as Monte Carlo estimator for $\Pi(x)$. Hence, by analogy with the above construction of h, the estimator (1.3) thus requires the simulation of N sample paths for the evolution of the underlying risk factors, where from each a discounted payoff $h(x, Z_i)$ is calculated to ultimately build the sample average $\widehat{\Pi}_N(x)$.

The use of the sample average (1.3) as a pointwise estimator of $\Pi(x)$ for any $x \in \mathcal{X}$ is backed up by the multivariate strong law of large numbers (e.g., Serfling [1980], Theorem 1.8(B)), which states that for $\mathbb{Q}^Z$-integrable $h(x, \cdot)$,

$$\widehat{\Pi}_N(x; Z_1, \ldots, Z_N) \to \Pi(x), \quad \text{as } N \to \infty,$$

$\mathbb{Q}$-almost surely. Under the same assumption, the multivariate weak law of large numbers (e.g., Serfling [1980], Theorem 1.8(A)) says that, for any $\epsilon > 0$,

$$\mathbb{Q}\big(\|\widehat{\Pi}_N(x) - \Pi(x)\| > \epsilon\big) \to 0, \quad \text{as } N \to \infty,$$

where the theory of large deviations (e.g., Dembo and Zeitouni [1998]) then asserts that, under a finite moment generating function of $h(x, Z)$ in a neighbourhood of zero, this probability converges exponentially fast. Eventually, assuming that $h(x, \cdot)$ is even square-integrable, the multivariate central limit theorem (e.g., Serfling [1980], Theorem 1.9.1(B)) yields

$$\sqrt{N}\big(\widehat{\Pi}_N(x; Z_1, \ldots, Z_N) - \Pi(x)\big) \xrightarrow{d} \mathcal{N}\big(0, \Sigma(x)\big), \quad \text{as } N \to \infty,$$

where $\Sigma(x)$ denotes the covariance matrix of the random vector $h(x, Z)$, and which thus allows to (approximately) quantify the accuracy of the estimator in distribution. In particular, the error made by $\widehat{\Pi}_N(x)$ is asymptotically normally distributed with mean zero and variance $\frac{1}{N}\Sigma(x)$, the rate at which it converges to zero in distribution being of order $\mathcal{O}(1/\sqrt{N})$.

1.1.3 Monte Carlo Sampling Strategies

If Monte Carlo methods are to be used in an optimisation procedure to approximatively solve a stochastic programming problem, different strategies for generating underlying random samples and employing the estimators along the procedure may be adopted, see, e.g.,

of the random vector Z, their applicability in the current setup where d_Z also becomes fairly large is quite impractical, see, e.g., Glasserman [2003].

Homem-de-Mello and Bayraksan [2014], for a comprehensive survey. In the case of the above calibration setup, where we seek to replace the pricing function Π by estimators of the form (1.3) and subsequently build an objective function through the transformation g, we will mainly investigate the following two strategies within this thesis.

Sample Average Approximation Strategy

The first strategy consists of taking for fixed N a sample of i.i.d. random vectors $Z_1, \dots, Z_N$, with the same distribution as Z, and using it throughout the entire optimisation procedure for each evaluation of the estimator (1.3). Consequently, this amounts to considering the problem

$$\min_{x \in \mathcal{X}} \left\{ \hat{f}_N(x) := g\Big(\widehat{\Pi}_N\big(x; Z_1, \dots, Z_N\big) - C^{\mathrm{mkt}} \Big) \right\}, \tag{1.4}$$

as an approximation to the original problem (1.1). Since the problem (1.4) depends on the set of random vectors $Z_1, \dots, Z_N$, it essentially constitutes a random problem. Its optimal value $\hat{f}_N^* = \min_{x \in \mathcal{X}} \hat{f}_N(x)$ is thus an estimator of the optimal value f^* of the original problem (1.1), and an optimal solution $\hat{x}_N^* \in \operatorname{argmin}_{x \in \mathcal{X}} \hat{f}_N(x)$ is an estimator of an optimal solution x^* of the original problem (1.1). However, for a particular realisation of the random sample $Z_1, \dots, Z_N$, the approximating problem (1.4) represents a deterministic problem instance, which can then be solved by adequate optimisation algorithms.

The strategy of using a single sample represents the most frequently used way to approximately solve a stochastic programming problem and is commonly known as the *sample average approximation* (SAA) approach. Its related literature, addressing mainly the original setup where an untransformed scalar-valued expectation is approximated by sample averages, is vast and has been dealt with by many authors. Even though the approach is closely related to maximum likelihood estimation and M-estimation in statistics, its main ideas have most likely emerged in the stochastic counterpart method (Rubinstein and Shapiro [1990, 1993]) and the sample path optimisation (Plambeck et al. [1996], Robinson [1996]). The term 'sample average approximation' seems to have eventually appeared in Kleywegt et al. [2001]. Given the way in which samples are generated, the approach is sometimes also referred to as the *external sampling approach*. For a more recent treatment of the topic, we refer, for instance, to the book of Shapiro et al. [2014] and the paper of Kim et al. [2015].

The SAA approach is simple, intuitive and based on a strong theoretical foundation, which allows to establish sound convergence properties for the sequences of optimal values and solutions. Notwithstanding, it has several drawbacks when used in its original form, as reported in various papers, e.g., Homem-de-Mello [2003] and Royset [2013], but also when applied to calibration problems. Firstly, it is often argued that the approach lacks efficiency and flexibility. In order to obtain a good estimate of a true optimal solution, the number of simulations N typically needs to be set to a considerable high value throughout the entire optimisation procedure; computational resources may thus be wasted in stages of the optimisation procedure where a rough estimate of the objective function is in fact entirely sufficient, such as in the beginning of a procedure. Secondly, due to using a single set of random vectors for all function evaluations, there is no variability and robustness in the estimation of an optimal solution. In unfavourable cases, this may lead to a bad sample path in which convergence to a global minimum only occurs for a considerable sample size N, such that the optimisation with a lower N is likely to get trapped in a spurious local minimum. Eventually, when applied to financial model calibration, the SAA approach does not reflect the way in which derivative instruments are actually priced by Monte Carlo methods. In fact, since Monte Carlo prices are computed at different parameter sets by using different sets of random vectors, this feature should also be taken into account in the calibration procedure of the underlying models.

In view of these potential shortcomings, it is thus reasonable to consider a second, more elaborate strategy to approximately solve the original problem (1.1) by means of Monte Carlo methods, which gives the user more flexibility and makes the calibration consistent with the way in which Monte Carlo model prices are computed.

Variable Sample Average Approximation Strategy

Specifically, instead of using a single sample with a fixed size for all evaluations, the strategy is characterised in that it uses a different set of i.i.d. random vectors with an adaptively chosen sample size every time the estimator (1.3) is evaluated. Accordingly, letting $Z_1^k, \dots, Z_{N_k}^k$ denote the sample used for the k-th evaluation with size N_k, where the Z_i^k's follow the same distribution as Z and are also drawn independently of the previous random samples, the objective function approximating f at the k-th stage is given by

$$\hat{f}_{N_k}(x) := g\Big(\widehat{\Pi}_{N_k}\big(x; Z_1^k, \dots, Z_{N_k}^k\big) - C^{\text{mkt}}\Big), \qquad x \in \mathcal{X}. \tag{1.5}$$

Hence, this procedure comes with the benefit of allowing the user to control the accuracy of the approximation to f and the time spent on each evaluation adaptively, via the number of Monte Carlo simulations N_k. As each objective function defined by (1.5) depends on a set of random vectors $Z_1^k, \dots, Z_{N_k}^k$, it is a random function and thus an estimator of the objective function f of the original problem (1.1). However, unlike the SAA strategy which leads to a self-contained random problem that may be fully minimised for a particular realisation of the underlying random sample by an algorithm, the VSAA strategy is tied to the sequential sampling algorithm that incorporates the VSAA scheme and selects new sample points (thus, the formulation of (1.5) as an objective function only). In particular, since $\hat{f}_{N_k}$ changes at every function evaluation k and (typically) returns a different value if called again at the same x, the observed objective function values turn out to be perturbed by noise. This requires the applied algorithm to account for the presence of noise.

Due to the way in which the samples are generated, the strategy can be regarded as a variant of the so-called *variable sample average approximation* (VSAA) approach, as advocated by Homem-de-Mello [2003] and Royset [2013] to overcome the drawbacks of the

SAA approach. The approach describes a general framework, where in each iteration the unknown objective function in form of a single-valued expectation is approximated by a new sample average estimator with an adaptively chosen sample size. On each single approximation, an unspecified number of function evaluations may then be performed depending on the user's preference, before a new approximation is generated with a different and usually larger sample size.

Characterisation of Approximation Procedures

It is important to note that both described strategies are not optimisation algorithms themselves but merely describe ways of replacing the original problem (1.1) with an approximative procedure. To eventually obtain approximations to the true optimal value and a corresponding optimal solution, adequate solvers must be applied, either on the problem instance as specified by the SAA strategy or in combination with the VSAA strategy, whose suitability depends on the characteristics of the underlying objective functions.

Due to the inherent complexity of the pricing models and the involved payoff structure, both procedures typically lead to highly nonlinear nonconvex objective functions and therefore demand global optimisation techniques to be adequately approached. Moreover, since Monte Carlo methods are employed to approximate the underlying pricing function Π, the objective functions $\hat{f}_N$ and $\hat{f}_{N_k}$ are computationally expensive to evaluate for fairly large numbers of Monte Carlo simulations and considering that only a restricted time budget of a few minutes is available to carry out the optimisation. However, unlike other expensive optimisation problems where an evaluation of the objective functions may take even hours, both considered approaches are not deemed to be prohibitively expensive. The use of Monte Carlo estimators for the unknown Π also introduces an additive, asymptotically normally distributed pointwise random error into the approximation procedures, which can be transferred under weak differentiability assumptions on the function g to the objective functions by the multivariate delta method (e.g., Serfling [1980], Theorem 3.3(A)). Yet, whereas the SAA strategy uses the same source of randomness at each evaluated $x \in \mathcal{X}$ such that observed objective function values $\hat{f}_N(x)$ are exact, the VSAA strategy employs different sets of random vectors at each evaluation which produces noisy objective function observations. Specifically, in the latter case, we then have at the k-th stage the approximate relation

$$\hat{f}_{N_k}(x) \overset{d}{\approx} f(x) + \epsilon_{N_k}(x), \qquad x \in \mathcal{X}, \tag{1.6}$$

where $\epsilon_{N_k}(x)$ denotes a normally distributed random variable with mean zero and variance $\frac{1}{N_k}\nabla g(\Pi(x))^\top \Sigma(x) \nabla g(\Pi(x))$, thus allowing to control the level of noise entering into an evaluation by N_k. Eventually, it shall be noted that gradient or Hessian information may only be of limited use in both approaches, due to the computational expensiveness of its numerical approximation. In particular, in the case of the VSAA strategy, the involved random noise is likely to produce inexact gradient or Hessian information if the level is not adjusted carefully.

Concerning the terminology used in latter context, it has to be clarified that, throughout this thesis, we speak of minimising a noisy objective function $\hat{f}$ on the parameter space $\mathcal{X}$ if we intend to minimise the underlying original (but unknown) objective function f and to this end use observed noisy function values $\hat{f}(x_k)$ with $x_k \in \mathcal{X}$, $k \in \mathbb{N}$. This way of expressing has become well-established in the optimisation literature, see, for instance, Kelley [1999] or Nocedal and Wright [2006]. In particular, in the present setup where we use Monte Carlo methods to approximately solve the problem (1.1) via the VSAA strategy, we thus have the noisy objective function values $\hat{f}_{N_k}(x_k)$, indexed by N_k, whose relation to f is characterised by the approximation (1.6).

1.2 Global Optimisation

Since the above calibration problems are essentially nonconvex, our principal objective is to tackle them by proper global optimisation techniques. This is in contrast to most approaches in the financial industry where local search techniques are employed, see, e.g., Hirsa [2012], Section 7.6, or Kienitz and Wetterau [2012], Section 9. Yet, as described above, the possible existence of several local minima is not the only difficulty arising in these problems since the highly nonlinear (data-fitting) structure of the objective functions poses a major obstacle for any global solver. This is further amplified by the fact that function evaluations are expensive and potentially contaminated by noise.

Given these additional challenges, we will now briefly review existing approaches from the literature on global optimisation regarding their suitability for solving problems of the above kind. To this end, recall that a point $x^* \in \mathcal{X}$ is a *global minimiser* of a general nonconvex objective function $f : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R}$ if $f(x^*) \leq f(x)$ for all $x \in \mathcal{X}$. The corresponding minimum function value $f(x^*)$ is then referred to as the *global minimum* of f on $\mathcal{X}$, which, however, may be attained at more than a single point x^*. In comparison, a point $x^* \in \mathcal{X}$ is a *local minimiser* of f if $f(x^*) \leq f(x)$ for all x in some neighbourhood of x^*, the related function value $f(x^*)$ then being said to be a *local minimum* of f. In particular, it should be noted that finding a global minimum of f is much more difficult than finding a local minimum, as every part of the domain $\mathcal{X}$ has to be explored in order to guarantee that a global minimum is found. As most of the approaches for global optimisation fall into one of the three categories exact methods, heuristic methods and response surface methods, we shall consider each of these in turn. For a more comprehensive treatment of topics on global optimisation, we refer to the handbooks of Horst and Pardalos [1995], Pardalos and Romeijn [2002], Locatelli and Schoen [2013], as well as the survey by Neumaier [2004] and the references therein.

1.2.1 Exact Methods

The first category of *exact methods* provides the most rigorous approach to global optimisation of a nonconvex objective function and covers deterministic methods that have a guarantee of finding a global solution. The most prominent methods within this category are based on the general concept of *branch-and-bound*, which subsumes many specific cases but also allows for further generalisations. The basic idea behind is to recursively split a problem into disjoint subproblems (branching) and determine on each of them a lower and upper bound on the objective function (bounding), such that any of these subproblems may be discarded sooner or later if it cannot lead to a better point than the best point found thus far, e.g. if its lower bound is larger than the current minimum function value.

The concept of branch-and-bound proves to be very efficient if the underlying problem has certain structure which is known and can be exploited to obtain sufficiently tight bounds. Broadly structured global optimisation problems which possibly allow for such kind of exploitation are, for instance, *d.c. programming problems*, i.e. problems in which the objective and the constraints can be put as the difference between two convex functions, problems with explicit analytical expressions that enable the use of *interval methods*, and *Lipschitz optimisation problems* where the objective function is assumed to be Lipschitz continuous with known constant. Yet, if no or not enough structural information on the problem is available, such as in our general approach to model calibration where only the data-fitting structure of the objective function is assumed to be known, then the branch-and-bound concept is not directly applicable. In particular, unless a specific model and some fitting function with usable properties are considered, one is not able to compute powerful and reliable bounds, such that any considered method presumably has to explore every region in the domain and becomes very inefficient.

Besides the mentioned approaches, there are also (partly heuristic) methods that use ideas from the branch-and-bound technique but do not require any particular structure of the objective function. These involve, among others, the original DIRECT method developed by Jones et al. [1993], which is based on splitting the domain into boxes and a Pareto principle for suitable box selection, and the Multilevel Coordinate Search (MCS) algorithm from Huyer and Neumaier [1999], which is similar to DIRECT but allows for a more irregular splitting procedure based on multilevel coordinate search to balance between global and local search. Both methods are derivative-free techniques, relying on function values only, and could therefore even be used in the presence of noise subject to careful modifications. However, since the methods essentially perform a grid search and do not further exploit any inherent structure of the objective function like smoothness, a large number of function evaluations is generally needed, which makes them quite unsuitable for expensive problems. Moreover, they may also exhibit difficulties in coping with the highly nonlinear (data-fitting) structure of the objective functions. In any case, though, we will in Chapter 4 of this thesis also assess the performance of our modified RBF method against the DIRECT and the MCS methods, for the sake of comparison.

1.2.2 Heuristic Methods

As opposed to exact methods, the second category of *heuristic methods* comprises methods that do not come with a strict guarantee of finding a global minimum. In fact, these methods are not designed for exploiting any structure of the underlying problem and rather employ some kind of random factor in their methodology, which thus gives a good chance that the global minimum is found in practice, although no certainty. Theoretically, many heuristic methods can be shown to find a global minimum with probability one.

A very common class of methods falling into the category of heuristics are *multi-start methods*, which initialise suitable local search methods from several starting points. The points are often chosen at random but could just as well be set by some more elaborate strategy, in the hope that one of them lies in the basin of attraction of a global minimiser. Whereas multi-start methods employ local searches that may most effectively be used with gradient or Hessian information if feasible for the objective function, other classes of heuristic methods make only use of function values. *Random search methods*, for instance, such as pure random search or pure adaptive search, are characterised by randomly generating points over the domain from a given distribution, which may or may not take previously evaluated points into account. A newly sampled candidate point is then accepted by the method if its function value is less than the current minimum function value, and otherwise rejected. A similar heuristic method using analogies to physics is *simulated annealing*. It essentially consists in sampling the objective function in a neighbourhood of the current best point according to some distribution, where a new sample is then always accepted if it is better than the current minimum function value but also with a certain probability if it is worse. Latter probability depends on the difference in function values and the so-called temperature which decreases along the iterations, such that the method gradually switches from a global exploration phase to a local one and ends in locally random sampling in the neighbourhood of the current best solution. Eventually, there is the broad class of *evolutionary algorithms*, among them genetic algorithms and differential evolution, which is based on the principles of evolutionary biology such as mutation, crossover and selection. Here, at each stage, a whole population of candidate solutions is maintained with their fitness level (usually corresponding to their objective function values), and further evolved by these evolutionary modifications towards better solutions until some terminal condition is reached.

Most heuristic methods are, in their simpler forms, easy to understand and implement, which makes them very popular in practice. They also often turn out to work well, if only slowly, and may thus be regarded as useful tools for applications in which function values are not expensive (but potentially perturbed by noise) and the primary objective is to find near-optimal solutions. Yet, as these methods typically require a very large number of function evaluations to yield a reasonably good solution, they are quite inappropriate for our type of global optimisation problems, too; not to mention again that they generally lack a guarantee of optimality, which is very desirable from a theoretical point of view. Notwithstanding, though, note that we compare our developed method in Chapter 4 with a multi-start strategy and a differential evolution heuristic on the (inexpensive) fitting of the Nelson-Siegel and Svensson models, as these present two popular ways of doing so.

1.2.3 Response Surface Methods

Most exact methods and especially the heuristic methods require a large number of objective function evaluations, which makes them somewhat unsuitable for solving the problems we have in mind. In fact, when dealing with functions that are expensive to evaluate – though not necessarily prohibitively expensive – it is natural to try to avoid needless function evaluations as much as possible. A suitable class of optimisation methods for this problem is therefore given by *response surface methods*. These methods aim at approximating the underlying objective function by a sequence of response surface models that then guide the selection of new evaluation points by means of some strategy, to iteratively refine the approximating models and to eventually find a global minimum of the original objective function. The response surface models, also referred to as *meta-models* or *surrogate models*, are composed of simple basis functions and are fit to the unknown objective function at a limited number of points, either through interpolation or some other approximation scheme. This provides for cheap function evaluations of the response surfaces and a straightforward way to modify the approximation. In particular, they thus also offer a convenient framework to take into account noisy function values. Eventually, a further compelling advantage of model-based methods over algorithms from the other categories lies in their ability of including additional information on the objective function when searching for a global minimum. This is beneficial from a practical point of view when solving a series of calibration problems in which the model remains the same for all problem instances and only the input data changes. In such a case, the resulting objective functions have the same distinctive structure, which is characterised in that various minimisers, depending continuously on the data, tend to cluster in different regions of the parameter space, where each cluster is formed by those optimal solutions that come from calibrations to similar/neighbouring data sets, see, e.g., Cairns and Pritchard [2001]. Consequently, given the way response surface methods work, the available structure of the objective function can efficiently be exploited through knowledge of previous calibrations or even through local searches within the current calibration, from which then promising evaluated points may be incorporated to improve the construction of response surfaces.

Since response surface methods appear to be the most suitable class of optimisation methods to approach the optimisation problems at hand, we will treat them more extensively in the respective chapters on global optimisation. Specifically, within this class of methods, we will predominantly focus on the so-called *radial basis function (RBF) method* by Gutmann [2001a,b], which thus shares all the abovementioned properties of this model class. In particular, the method is theoretically well-founded, has been shown to convergence for any continuous function under mild assumptions, and has proven to be a powerful tool for performing well on the majority of well-behaved expensive optimisation problems.

1.3 Outline of Thesis

The thesis is structured as follows. In Chapter 2, we deal with the use of Monte Carlo methods to approximately solve the original calibration problem (1.1) and derive essential almost sure convergence properties for the (optimal) estimators of the SAA and VSAA strategies. To obtain the respective results, we resort to available limit theorems in a Banach space setting, which will be reviewed in the first part of the chapter. We then address each of the two strategies individually, where in the case of the SAA strategy we complement the well-established strong consistency of optimal estimators by their corresponding almost sure rates of convergence and deduce several important implications. In the case of the VSAA strategy, we derive the strong uniform consistency of the objective function estimators under (relatively) weak assumptions and provide uniform sample path bounds, which may be used to show convergence of a variable-sample method using the VSAA scheme.

We then turn our focus to global optimisation, where we begin in Chapter 3 by considering the minimisation of an expensive and deterministic objective function on a compact set. As response surface methods are particularly suited for this purpose, we initially describe the basic structure of this class of optimisation methods and give a brief literature overview of existing methods. We will then introduce radial basis functions, provide conditions that guarantee the uniqueness of a radial basis function interpolant to a given function, and review related properties. This is necessary to eventually outline Gutmann's RBF method, which will serve as a main reference for our developments in the subsequent chapters. In particular, we will describe the individual steps of the method, summarise main convergence results, and discuss relevant issues regarding the implementation of the method.

Gutmann's original RBF method is theoretically well-established and works well on reasonably smooth objective functions. Yet, it has been shown in various sources to converge only very slowly to a global minimum of objective functions with a more complex structure, due to a weakly performing local search. Some improvements and extensions have already been suggested on that account but still lack efficiency when applied to solve calibration problems, or data-fitting problems in general. We thus address this issue in Chapter 4, where we present a modified RBF method that exploits the particular structure of these problems and complements the inherent search mechanism of the original RBF method by an extended local search technique. In line with the original method, we then show convergence of the modification and derive several convergence results. Eventually, we discuss the applicability of the modified method on relevant test problems and real world examples, where we calibrate the Hull-White one-factor model under the SAA strategy (see Appendix A.2) and fit the well-known Nelson-Siegel and Svensson models via an enhanced approach (see Appendix A.3 for further details).

In Chapter 5, we address the problem of optimising an expensive and noisy objective function on a compact set. Similar to the deterministic case, we first review the available literature of response surface methods in the presence of noise, and then consider the problem of approximating a given (noisy) function by means of radial basis functions when error bounds on the function values are available. Unlike interpolation, there are several possibilities to undertake this, depending on the designated use of the resulting approximant. We discuss the most common approaches for integration into a response surface method and eventually argue in favour of a regularised (weighted) least-squares approach. Based on Gutmann's original RBF method and on latter means of approximation, we then develop a RBF method that is able to deal with noisy objective function values, establish its convergence under some simplistic assumption on the error bounds, and state several convergence results. Concluding the chapter, we discuss important issues considering the implementation of the method and assess its practicability on test problems and by calibrating the Hull-White model under the VSAA strategy, where we test different schedules of sample sizes.

Finally, Chapter 6 contains our conclusions and discusses future directions of research arising from this work, whereas Appendix A provides the reader with all relevant material on the calibration of the Hull-White model as well as the enhanced fitting of the Nelson-Siegel and the Svensson models, used to additionally assess the practicability of our developed methods.

To the best of our knowledge, we are the first to investigate the use of standard Monte Carlo methods for the calibration of financial models within a global optimisation context. Arising from this investigation, we are able to make the following main contributions.

In Chapter 2, our extensive convergence analysis of (optimal) estimators closes several gaps in the literature on the SAA and VSAA approaches, see Banholzer et al. [2018a,b], respectively. As far as we are aware of, neither have almost sure rates of convergence been provided for the strong consistency of optimal estimators in the SAA approach, nor have any of our subsequent derivations been previously addressed. Similarly, we have not found any uniformly derived pathwise bounds on the error of objective functions in the VSAA approach, while the strong uniform consistency has already been stated but under very strong exponential moment conditions.

In Chapter 4, a further novelty is presented in form of the modified RBF method with extended local search (cf. Banholzer et al. [2017a]), which is particularly suited to solve deterministic (but not necessarily expensive) data-fitting problems. We establish convergence of the method, and provide numerical evidence that the suggested method works well and considerably improves on existing methods when applied to data-fitting problems.

In Chapter 5, our main contribution is the development of the RBF method for expensive and noisy objective functions, cf. Banholzer et al. [2017b]. In particular, we provide a proof of convergence of the method (albeit under some simplified assumptions) and show its practical usability.

Eventually, even though not directly related to Monte Carlo calibration and therefore only included aside, Appendix A contains a novel approach for fitting the Nelson-Siegel and Svensson models by means of zero rates, also to be found in Banholzer et al. [2017c].

Chapter 2

Monte Carlo Sampling Strategies

In this chapter, we investigate essential convergence properties of the Monte Carlo sampling strategies, as motivated earlier in the introductory chapter. To this end, we resume working in a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and assume the existence of an equivalent martingale measure $\mathbb{Q}$, under which pricing takes place. We further recall that we are fundamentally interested in solving the original optimisation problem

$$\min_{x \in \mathcal{X}} \Big\{ f(x) = g\big(\Pi(x) - C^{\mathrm{mkt}}\big) \Big\}, \tag{2.1}$$

where $\mathcal{X} \subset \mathbb{R}^d$ denotes a nonempty finite-dimensional[5] compact set with the usual (Euclidean) metric, and $g : \mathbb{R}^l \to \mathbb{R}_{\geq 0}$ is a nonnegative continuous function that measures the difference between the vector of model prices $\Pi(x)$ and corresponding market prices $C^{\mathrm{mkt}} \in \mathbb{R}^l$, with $l > d$. Specifically, by virtue of the mathematical theory on derivatives pricing, it further follows that Π can be written as

$$\Pi(x) = \mathbb{E}^{\mathbb{Q}}\big[h(x, Z)\big], \tag{2.2}$$

where $h : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}^l$ describes the discounted payoff vector of all involved calibration instruments as a function depending on the set of model parameters $x \in \mathcal{X}$ and some random vector Z whose distribution $\mathbb{Q}^Z$ is supported on a set $\mathcal{Z} \subset \mathbb{R}^{d_Z}$. In particular, to ensure that Π is well-defined, we assume that h is a random function and that $h(x, \cdot)$ is $\mathbb{Q}^Z$-integrable for each $x \in \mathcal{X}$.

By formulating the pricing function Π as an expected value, problem (2.1) becomes a (transformed) stochastic programming problem. Its optimal value and the set of optimal solutions shall be denoted by f^* and $\mathcal{X}^*$, respectively, where for a nonempty $\mathcal{X}^*$ an optimal solution is indicated by x^*. Unfortunately, however, in many situations the distribution of the random function $h(\cdot, Z)$ is not known exactly, such that the expected value in (2.2) and the objective function in (2.1) cannot be evaluated readily and therefore need to be

[5] Due to the considered application, the set $\mathcal{X}$ is assumed to be finite-dimensional but may be extended to general compact and metrisable spaces, subject to minor modifications.

approximated in some way. Using Monte Carlo methods, one frequently applied approach to do so is by the *sample average approximation (SAA) strategy*, which by use of a single sample of i.i.d. random vectors $Z_1, \dots, Z_N$ from the same distribution as Z leads to a self-contained random problem. Its optimal value and an optimal solution are thus well-defined estimators for f^* and an x^*, respectively, whose convergence properties may be analysed for increasing sample size N. However, while the strong consistency of these quantities is well-established in the literature, the involved rates of almost sure convergence have yet not been investigated. Based on Banholzer et al. [2018a], we will address this issue within the calibration setup as a first main objective of this chapter. In particular, to be able to quantify the rates at which convergence takes place almost surely, we will resort to the law of the iterated logarithm (LIL) for sequences of random variables in a Banach space setting. This is similar to the case of the functional central limit theorem (CLT), which has already served to derive the asymptotic distributions of optimal estimators.

The basic idea of using Monte Carlo methods to approximately solve (2.1) also yields other possible approaches, among them the *variable sample average approximation (VSAA) strategy*. As opposed to the SAA strategy, this strategy gives the user the flexibility to employ for a prespecified sequence of Monte Carlo sample sizes $\{N_k\}_{k\in\mathbb{N}}$, $N_k \in \mathbb{N}$, different sets of i.i.d. random vectors $Z_1^k, \dots, Z_{N_k}^k$ from the same distribution as Z and independent of previous samples, along the optimisation procedure, which then results in a sequence of newly constructed random objective functions estimating f. In particular, since the approximating objective function changes with each new sample, the convergence analysis in this approach is in fact tied to the sequential optimisation algorithm that incorporates the VSAA sampling scheme and selects new evaluation points. Yet, irrespective of this dependence, some general, although incomplete, results on the almost sure convergence of the sequence of objective functions are available in the literature, which may aid to establish convergence of such a variable-sample method in the almost sure sense. As our second main objective in this chapter, we will show according to Banholzer et al. [2018b], yet within the calibration context, how these findings can be improved and completed, making use of existing results on the strong law of large numbers (SLLN) and the LIL for general arrays of Banach space valued random variables.

To make this chapter as self-contained as possible, we briefly review in Section 2.1 basic concepts of Banach space valued random variables and related limit theorems to be used throughout this chapter. We then describe our main findings within the calibration context for each of the two Monte Carlo sampling strategies in greater detail. Specifically, in Section 2.2, we first address the SAA approach to review the strong consistency of optimal estimators, to derive the involved almost sure rates of convergence, and to draw further conclusions regarding universal confidence sets, convergence in mean and rates of error probabilities for these estimators. Subsequently, we study in Section 2.3 the VSAA approach to give conditions that ensure the strong uniform consistency of the objective function estimators and allow to derive almost sure uniform error bounds.

2.1 Probability in Banach Spaces

Let us first introduce some basic concepts of random variables taking values in separable Banach spaces and corresponding limit theorems in these spaces, as well as the related delta method. For a more detailed discussion on these subjects, we refer to Ledoux and Talagrand [1991], Shapiro et al. [2014], Section 7.2.8, and individually given references hereinafter. Note that we assume the separability of the spaces as the required limit theorems and the delta method seem to be only formulated for separable Banach spaces, as far as we know. It is an open question whether these results may be extended to non-separable Banach spaces.

2.1.1 Banach Space Valued Random Variables

Let B denote a separable Banach space, i.e. a vector space over the field of real numbers equipped with a norm $\|\cdot\|$ for which it is complete and which contains a countable dense subset. Its topological dual is denoted by B' and duality is given by $\vartheta(\mu) = \langle \vartheta, \mu \rangle$ for $\vartheta \in B'$, $\mu \in B$. For convenience, the dual norm of $\vartheta \in B'$ shall also be denoted by $\|\vartheta\|$.

A random variable X on B, or B-valued random variable in short, is a measurable mapping from some probability space $(\Omega, \mathcal{F}, \mathbb{Q})$ into B equipped with its Borel σ-algebra $\mathcal{B}(B)$ generated by the open sets of B. Thus, for every Borel set $U \in B$, we have $X^{-1}(U) \in \mathcal{F}$. A random variable X with values in B is said to be strongly (or Bochner) integrable if the real-valued random variable $\|X\|$ is integrable, i.e. $\mathbb{E}^{\mathbb{Q}}[\|X\|] < \infty$. The variable is said to be weakly (or Pettis) integrable if for any $\vartheta \in B'$ the real-valued random variable $\vartheta(X)$ is integrable and there exists a unique element $\mu \in B$ such that $\vartheta(\mu) = \mathbb{E}^{\mathbb{Q}}[\vartheta(X)] = \int \vartheta(X)\,\mathrm{d}\mathbb{Q}$. If this is the case, then the element μ is denoted by $\mathbb{E}^{\mathbb{Q}}[X]$ and called the expected value of X. Note that, in separable Banach spaces, weak integrability already implies the strong measurability such that the fine line between weak and strong integrability is the existence of the first strong moment $\mathbb{E}^{\mathbb{Q}}[\|X\|]$. In particular, if X is strongly integrable, then it is also weakly integrable and the integrals coincide. Given that $\mathbb{E}^{\mathbb{Q}}[\vartheta(X)] = 0$ and $\mathbb{E}^{\mathbb{Q}}[\vartheta^2(X)] < \infty$ for all $\vartheta \in B'$, the covariance function of X is defined by $(\mathbb{C}\mathrm{ov}^{\mathbb{Q}}\, X)(\vartheta_1, \vartheta_2) := \mathbb{E}^{\mathbb{Q}}[\vartheta_1(X)\vartheta_2(X)]$, $\vartheta_1, \vartheta_2 \in B'$, which is a nonnegative symmetric bilinear form on B'.

The familiar notions of convergence of random variables on the real line extend in a straightforward manner to Banach spaces. As such, a sequence $\{X_N\}$ of random variables with values in B converges in distribution (or weakly) to a random variable X, denoted by $X_N \xrightarrow{d} X$, if for any bounded and continuous function $\varphi : B \to \mathbb{R}$, $\mathbb{E}^{\mathbb{Q}}[\varphi(X_N)] \to \mathbb{E}^{\mathbb{Q}}[\varphi(X)]$ as $N \to \infty$. Moreover, $\{X_N\}$ converges in probability to X, in brief $X_N \xrightarrow{p} X$, if for each $\epsilon > 0$, $\lim_{N\to\infty} \mathbb{Q}(\|X_N - X\| > \epsilon) = 0$. The sequence is said to be bounded in probability if, for each $\epsilon > 0$, there exists $M_\epsilon > 0$ such that $\sup_N \mathbb{Q}(\|X_N\| > M_\epsilon) < \epsilon$. Similarly, $\{X_N\}$ is said to converge $\mathbb{Q}$-almost surely to a B-valued random variable X if $\mathbb{Q}(\lim_{N\to\infty} X_N = X) = 1$, and it is $\mathbb{Q}$-almost surely bounded if $\mathbb{Q}(\sup_N \|X_N\| < \infty) = 1$. Further, the sequence $\{X_N\}$ is said to converge completely to X as $N \to \infty$ if $\sum_{N=1}^{\infty} \mathbb{Q}(\|X_N - X\| > \epsilon) < \infty$ for all $\epsilon > 0$.

Finally, denoting by $L_1(B) = L_1(\Omega, \mathcal{F}, \mathbb{Q}; B)$ the space of all B-valued random variables X on $(\Omega, \mathcal{F}, \mathbb{Q})$ such that $\mathbb{E}^{\mathbb{Q}}[\|X\|] < \infty$ (which again is a Banach space), we say that the sequence $\{X_N\}$ converges to X in $L_1(B)$ if X_N, X are in $L_1(B)$ and $\mathbb{E}^{\mathbb{Q}}[\|X_N - X\|] \to 0$ as $N \to \infty$.

For a sequence $\{X_N\}_{N\in\mathbb{N}}$ of i.i.d. B-valued random variables with the same distribution as X (i.e. $X_i \overset{d}{=} X$ in short), we define $S_N := \sum_{i=1}^{N} X_i$, and for a general array $\{\{X_{ki}\}_{i=1}^{N_k}\}_{k\in\mathbb{N}}$ of i.i.d. B-valued random variables with the same distribution of X and deterministic $N_k \in \mathbb{N}$, we set $S_k := \sum_{i=1}^{N_k} X_{ki}$. We write $\mathrm{Log}(x)$ to denote the function $\max\{1, \log x\}$,[6] $x \geq 0$, and let $\mathrm{LLog}(x)$ stand for $\mathrm{Log}(\mathrm{Log}(x))$. Further, we set $a_N := \sqrt{2N\,\mathrm{LLog}(N)}$ for $N \in \mathbb{N}$. For any sequence $\{\mu_N\}$ in B, the set of all limit points is denoted by $\mathrm{CP}(\{\mu_N\})$. If $\mu \in B$, $U_1, U_2 \subset B$, the distance from the point μ to the set U_1 is given by $\mathrm{dist}(\mu, U_1) := \inf_{\mu_1\in U_1}\|\mu - \mu_1\|$, and the deviation (or excess) of the set U_1 from the set U_2 by $\mathrm{dev}(U_1, U_2) := \sup_{\mu_1\in U_1} \mathrm{dist}(\mu_1, U_2)$.

2.1.2 Basic Limit Theorems

Based on the notions of convergence of random variables, the SLLN, the CLT and the LIL on the real line can be extended, subject to minor modifications, to random variables taking values in a separable Banach space. However, the necessary and sufficient conditions for these to hold are fundamentally different from those for the real line.

For the sake of generality, the following discussion is phrased in terms of a generic separable Banach space B, irrespective of the underlying geometry. However, to establish our main convergence results within the SAA setup in Section 2.2, we will work in the separable Banach space $C(\mathcal{X}, \mathbb{R}^l)$ of continuous functions $\varphi : \mathcal{X} \to \mathbb{R}^l$, endowed with the supremum norm $\|\varphi\|_\infty = \sup_{x\in\mathcal{X}}\|\varphi(x)\|$, and in the separable Banach space $C^1(\mathcal{X}, \mathbb{R}^l)$ of continuously differentiable vector-valued functions φ, defined on an open neighbourhood of the compact set $\mathcal{X}$ and equipped with the norm

$$\|\varphi\|_{1,\infty} = \sup_{x\in\mathcal{X}}\|\varphi(x)\| + \sup_{x\in\mathcal{X}}\|\mathrm{D}\varphi(x)\|_{L(\mathbb{R}^d,\mathbb{R}^l)},$$

where $\mathrm{D}\varphi(x)$ denotes the differential of the function $\varphi \in C^1(\mathcal{X}, \mathbb{R}^l)$ at the point x and $\|\cdot\|_{L(\mathbb{R}^d,\mathbb{R}^l)}$ is the operator norm of linear mappings from $\mathbb{R}^d$ into $\mathbb{R}^l$. Instead of the generic B-valued random variables X and the i.i.d. copies X_i, $i = 1, \ldots, N$, we will then consider the random variables $\widetilde{X} := h(\,\cdot\,, Z) - \mathbb{E}^{\mathbb{Q}}[h(\,\cdot\,, Z)]$ and $\widetilde{X}_i := h(\,\cdot\,, Z_i) - \mathbb{E}^{\mathbb{Q}}[h(\,\cdot\,, Z_i)]$, respectively, to which the respective limit theorems for sequences of B-valued random variables are applied. In a similar way, our main results within the VSAA setup in Section 2.3 are derived by using the $C(\mathcal{X}, \mathbb{R}^l)$-valued random variables $\widetilde{X}_{ki} := h(\,\cdot\,, Z_i^k) - \mathbb{E}^{\mathbb{Q}}[h(\,\cdot\,, Z_i^k)]$, $\widetilde{X}_{ki} \overset{d}{=} \widetilde{X}$, instead of the i.i.d. B-valued general array $\{X_{ki}\}$. Our minor results are deduced by considering the finite-dimensional Banach space $\mathbb{R}^l$ with $\|\cdot\|$. Note that if the limit theorems were formulated for non-separable Banach spaces, then also other function spaces than $C(\mathcal{X}, \mathbb{R}^l)$ and $C^1(\mathcal{X}, \mathbb{R}^l)$ could be considered for our purposes, such as the Sobolev space $W^{1,\infty}(\mathcal{X})$ of

[6] To avoid any confusion, note that throughout this thesis we indicate the natural logarithm by ‘log’.

Lipschitz continuous functions on $\mathcal{X}$ equipped with the naturally induced norm or the space $L_\infty(\mathcal{X})$ of essentially bounded measurable functions on $\mathcal{X}$ with the essential supremum norm. In particular, in such a case, the underlying assumption on the continuity of g might then be further relaxed to the measurability.

In both approaches, the obtained convergence statements are further transferred by the transformation g to the respective objective functions in Banach spaces of scalar-valued functions, i.e. to $C(\mathcal{X}) = C(\mathcal{X}, \mathbb{R})$, $C^1(\mathcal{X}) = C^1(\mathcal{X}, \mathbb{R})$, and $\mathbb{R}$. Since this, however, requires stronger assumption on g (and the involved Π and $\widehat{\Pi}_N$) than continuity, we state the following main assumptions, which will then be used throughout the remainder of this chapter when needed:

(G1) There exists a compact set $\mathcal{Y} \subset \mathbb{R}^l$ such that $\Pi(x)$ and $\widehat{\Pi}_N(x)$ are interior to $\mathcal{Y}$ for all $x \in \mathcal{X}$ and $N \in \mathbb{N}$, and such that g is Lipschitz continuous on $\mathcal{Y}$ with constant L_g.

(G2) The function g is directionally differentiable at $\Pi(x)$ and $\widehat{\Pi}_N(x)$ for all $x \in \mathcal{X}$ and $N \in \mathbb{N}$.

Assumptions (G1) and (G2) pose mild conditions on the smoothness of a function g for measuring the discrepancy between model and market data. In fact, since most financial derivatives (or at least any of the equivalent products linked through a parity relation) have a bounded discounted payoff h, the images of $\Pi(x)$ and $\widehat{\Pi}_N(x)$ are typically contained in some compact set $\mathcal{Y} \subset \mathbb{R}^l$ for all $x \in \mathcal{X}$ and $N \in \mathbb{N}$. For caplets and payer swaptions, for instance, the discounted payoffs in form of (A.29) and (A.30), respectively, are bounded if the underlying model yields bounded zero-coupon bond prices at time $t = 0$, such as in the Hull-White model (by the put-call parity for bond options, e.g., Musiela and Rutkowski [2005], p. 515, the discounted payoffs of floorlets and receiver swaptions can be shown to be bounded, too). In any case, note that since the set of no-arbitrage prices of a financial derivative is nonempty and bounded from below and above by a finite number, see, e.g., Föllmer and Schied [2004], Theorem 5.30, the true pricing function Π is always interior to a compact set.

Assumption (G1) obviously holds if g constitutes a p-norm $\|\cdot\|_p$, $1 \leq p \leq \infty$, since all norms are Lipschitz continuous with constant one. In case g is itself a transformation of a p-norm, then the transformation is required to be (globally) Lipschitz continuous on $\mathcal{Y}$, or rather only locally Lipschitz continuous, as $\mathcal{Y}$ is compact and explicit knowledge of the constant L_g is not needed, see, e.g., Hildebrandt [2003], Proposition 11.3. Moreover, any p-norm is also convex and thus directionally differentiable at any point $y \in \mathbb{R}^l$ with a finite value $\|y\|_p$, see, e.g., Bonnans and Shapiro [2000], p. 86. Hence, the function g certainly satisfies assumption (G2) if it is a p-norm itself or a directionally differentiable (and continuous monotone) transformation of a p-norm.

2.1.2.1 Sequences of Random Variables

We first provide results on the CLT and the LIL for sequences of i.i.d. B-valued random variables, which will be used in the SAA setup to review results on the asymptotic distribution of optimal estimators and to derive their almost sure rates of convergence, respectively.

The Central Limit Theorem

A random variable X with values in B is said to satisfy the *CLT* if for i.i.d. B-valued random variables $\{X_N\}$ with the same distribution as X, there exists a mean zero Gaussian random variable Y with values in B such that

$$\frac{S_N}{\sqrt{N}} \xrightarrow{d} Y, \text{ as } N \to \infty,$$

cf., for instance, Ledoux and Talagrand [1991], Chapter 10. Here, by definition, a B-valued random variable Y is Gaussian if for any $\vartheta \in B'$, $\vartheta(Y)$ is a real-valued Gaussian random variable. In particular, note that all weak moments of Y thus exist for any $\vartheta \in B'$, and it follows from Fernique's theorem (see Fernique [1970]) that Y also has finite strong moments of all orders, i.e. $\mathbb{E}^{\mathbb{Q}}[\|Y\|^p] < \infty$ for $p > 0$. If X satisfies the CLT in B, then for any $\vartheta \in B'$ the real-valued random variable $\vartheta(X)$ satisfies the CLT with limiting Gaussian distribution of variance $\mathbb{E}^{\mathbb{Q}}[\vartheta^2(X)] < \infty$. Hence, the sequence $\{S_N/\sqrt{N}\}$ converges in distribution to a Gaussian random variable Y with the same covariance function as X, i.e. for $\vartheta_1, \vartheta_2 \in B'$, we have $(\mathbb{C}\mathrm{ov}^{\mathbb{Q}} X)(\vartheta_1, \vartheta_2) = (\mathbb{C}\mathrm{ov}^{\mathbb{Q}} Y)(\vartheta_1, \vartheta_2)$ (and especially $\mathbb{E}^{\mathbb{Q}}[\vartheta_1^2(X)] = \mathbb{E}^{\mathbb{Q}}[\vartheta_1^2(Y)]$).

For general Banach spaces, no necessary and sufficient conditions such that a random variable X satisfies the CLT seem to be known. In particular, as mentioned e.g. by Kuelbs [1976a], the moment conditions $\mathbb{E}^{\mathbb{Q}}[X] = 0$ and $\mathbb{E}^{\mathbb{Q}}[\|X\|^2] < \infty$ are neither necessary nor sufficient for the CLT, as opposed to real-valued random variables (see Gnedenko and Kolmogorov [1954] for the equivalence). Nevertheless, sufficient conditions can be given for certain classes of random variables, such as for mean zero Lipschitz random variables X with square-integrable (random) Lipschitz constant on the spaces $C(\mathcal{X}, \mathbb{R}^l)$ and $C^1(\mathcal{X}, \mathbb{R}^l)$, see Araujo and Giné [1980], Chapter 7.

The Law of the Iterated Logarithm

For the LIL in Banach spaces, essentially two definitions may be distinguished. The first definition naturally arises from Hartman and Wintner's LIL for real-valued random variables, see Hartman and Wintner [1941], and says that a random variable X satisfies the *bounded LIL* if the sequence $\{S_N/a_N\}$ with $a_N = \sqrt{2N\,\mathrm{LLog}(N)}$ is $\mathbb{Q}$-almost surely bounded in B, or equivalently, if the nonrandom limit (due to Kolmogorov's zero-one law)

$$\Lambda(X) := \limsup_{N\to\infty} \frac{\|S_N\|}{a_N}$$

is finite, $\mathbb{Q}$-almost surely (cf. Ledoux and Talagrand [1991], Section 8.2).

Strassen's sharpened form of the LIL for random variables on the real line, see Strassen [1964], however, suggests a second natural definition of the LIL in Banach spaces, which is known as the *compact LIL*. Accordingly, X satisfies the compact LIL if the sequence $\{S_N/a_N\}$ is not only $\mathbb{Q}$-almost surely bounded in B, but $\mathbb{Q}$-almost surely relatively compact in B. While coinciding in finite dimensions, both definitions clearly differ from each other in the case of infinite-dimensional Banach spaces. Kuelbs [1976a] further showed that when the sequence $\{S_N/a_N\}$ is $\mathbb{Q}$-almost surely relatively compact in B, then there is a convex symmetric and necessarily compact set K in B such that

$$\lim_{N\to\infty} \operatorname{dist}\left(\frac{S_N}{a_N}, K\right) = 0, \qquad \text{and} \qquad \operatorname{CP}\left(\left\{\frac{S_N}{a_N}\right\}\right) = K, \tag{2.3}$$

each $\mathbb{Q}$-almost surely, which in fact may be seen as an equivalent definition of the compact LIL (e.g., Ledoux and Talagrand [1991], Theorem 8.5). In particular, we then have $\Lambda(X) = \sup_{\mu\in K}\|\mu\|$, $\mathbb{Q}$-almost surely.

The limit set $K = K_X$ in (2.3) is known to be the unit ball of the reproducing kernel Hilbert space $H = H_X \subset B$ associated to the covariance of X, and can briefly be described as follows, see Kuelbs [1976a] and Goodman et al. [1981] for further details. Assume that for each $\vartheta \in B'$, $\mathbb{E}^{\mathbb{Q}}[\vartheta(X)] = 0$ and $\mathbb{E}^{\mathbb{Q}}[\vartheta^2(X)] < \infty$, and consider for the Hilbert space $L_2 = L_2(\Omega, \mathcal{F}, \mathbb{Q})$ the operator $A = A_X$ defined by $A : B' \to L_2$, $A\vartheta = \vartheta(X)$. We then have

$$\|A\| = \sup_{\|\vartheta\|\leq 1} \left(\mathbb{E}^{\mathbb{Q}}[\vartheta^2(X)]\right)^{1/2} =: \sigma(X), \tag{2.4}$$

and by a closed graph argument that A is bounded. Moreover, since X has a separable range, the adjoint $A' = A'_X$ of the operator A with $A'\xi = \mathbb{E}^{\mathbb{Q}}[\xi X]$ for $\xi \in L_2$ maps L_2 into $B \subset B''$, and it holds $A(B')^{\perp} = \ker A'$ and thus $\overline{A(B')} = \ker(A')^{\perp}$. The operator A' induces a bijection of $\ker(A')^{\perp}$ onto the image of A', and considering the space $A'(L_2) \subset B$ equipped with the scalar product $\langle\cdot,\cdot\rangle_X$ transferred from L_2 and given by $\langle A'\xi_1, A'\xi_2\rangle_X = \langle \xi_1, \xi_2\rangle_{L_2} = \mathbb{E}^{\mathbb{Q}}[\xi_1\xi_2]$ for $\xi_1, \xi_2 \in L_2$, it determines a (separable) Hilbert space H. Latter space reproduces the covariance structure of X in that for $\vartheta_1, \vartheta_2 \in B'$ and any element $\mu = A'(\vartheta_2(X)) \in H$, we have $\vartheta_1(\mu) = \mathbb{E}^{\mathbb{Q}}[\vartheta_1(X)\vartheta_2(X)]$. In particular, if X_1 and X_2 are two random variables with the same covariance function, it follows from the reproducing property that $H_{X_1} = H_{X_2}$. Further note that H is the completion, with respect to the scalar product $\langle\cdot,\cdot\rangle_X$, of the image of B' by the composition $A'A : B' \to B$. Eventually, the closed unit ball K of H, i.e. $K = \{\mu \in B : \mu = \mathbb{E}^{\mathbb{Q}}[\xi X], (\mathbb{E}^{\mathbb{Q}}[\|\xi\|^2])^{1/2} \leq 1\}$, is a bounded and convex symmetric subset of B, and it can be shown that

$$\sup_{\mu\in K}\|\mu\| = \sigma(X).$$

As the image of the (weakly compact) unit ball of L_2 under A', the set K is weakly compact. It is compact if $\mathbb{E}^{\mathbb{Q}}[\|X\|^2] < \infty$, since the latter entails the existence of a sequence of finite partitions of the measure space such that A may be approximated by finite-dimensional (thus compact) operators under the operator norm, which in turn implies the norm compactness of A' by Schauder's theorem. Moreover, K is compact if and only if the family of random variables $\{\vartheta^2(X) : \vartheta \in B', \|\vartheta\| \leq 1\}$ is uniformly integrable, which immediately follows from the Dunford-Pettis property for L_1-spaces, see, e.g., Diestel and Uhl [1977], Corollary III.2.14. For a summary of conditions that equivalently describe the compactness of K, we refer to Ledoux and Talagrand [1991], Lemma 8.4.

While for a real-valued or, more generally, finite-dimensional random variable X the LIL is satisfied if and only if $\mathbb{E}^{\mathbb{Q}}[X] = 0$ and $\mathbb{E}^{\mathbb{Q}}[\|X\|^2] < \infty$ (see Strassen [1966] and Pisier and Zinn [1978]), the moment conditions are neither necessary nor sufficient for a B-valued random variable to satisfy the LIL in an infinite-dimensional setting, see Kuelbs [1976a]. Yet, conditions for the bounded LIL to hold were initially given by Kuelbs [1977], asserting that under the hypothesis $\mathbb{E}^{\mathbb{Q}}[X] = 0$ and $\mathbb{E}^{\mathbb{Q}}[\|X\|^2] < \infty$, the sequence $\{S_N/a_N\}$ is $\mathbb{Q}$-almost surely bounded if and only if $\{S_N/a_N\}$ is bounded in probability. Similarly, Kuelbs also showed under the same assumptions that $\{S_N/a_N\}$ is $\mathbb{Q}$-almost surely relatively compact in B (and thus (2.3) holds for the unit ball K of the reproducing kernel Hilbert space associated to the covariance of X) if and only if

$$\frac{S_N}{a_N} \xrightarrow{p} 0, \quad \text{as } N \to \infty, \tag{2.5}$$

which holds if and only if

$$\mathbb{E}^{\mathbb{Q}}\big[\|S_N\|\big] = o(a_N). \tag{2.6}$$

An immediate consequence of this result is that, given the moment conditions, X satisfying the CLT implies that X also satisfies the compact LIL (Pisier, 1975), but not vice versa (Kuelbs, 1976b). Specifically, the former statement holds since convergence in distribution of $\{S_N/\sqrt{N}\}$ to a mean zero Gaussian random variable in B entails that the sequence is bounded in probability, from which then (2.5) follows directly.

Considering the necessary conditions for the random variable X to satisfy the LIL in Banach spaces, however, it turns out that the moment condition $\mathbb{E}^{\mathbb{Q}}[\|X\|^2] < \infty$ is unnecessarily restrictive in infinite dimensions and can hence be further relaxed. This leads to the following characterisation of the LIL in Banach spaces, providing optimal necessary and sufficient conditions, cf. Ledoux and Talagrand [1988], Theorems 1.1 and 1.2. In this regard, note that since the boundedness in probability of $\{S_N/a_N\}$ comprises $\mathbb{E}^{\mathbb{Q}}[X] = 0$, cf. Ledoux and Talagrand [1988], Proposition 2.3, the latter property is already omitted in condition (ii) of both respective statements.

Theorem 2.1. *Let X be a random variable with values in a separable Banach space.*

(a) The sequence $\{S_N/a_N\}$ is $\mathbb{Q}$-almost surely bounded if and only if (i) $\mathbb{E}^{\mathbb{Q}}[\|X\|^2/\operatorname{LLog}(\|X\|)] < \infty$, (ii) for each $\vartheta \in B'$, $\mathbb{E}^{\mathbb{Q}}[\vartheta^2(X)] < \infty$, and (iii) $\{S_N/a_N\}$ is bounded in probability.

(b) The sequence $\{S_N/a_N\}$ *is* $\mathbb{Q}$*-almost surely relatively compact if and only if* (i) $\mathbb{E}^{\mathbb{Q}}[\|X\|^2/\operatorname{LLog}(\|X\|)] < \infty$*,* (ii) $\{\vartheta^2(X) : \vartheta \in B', \|\vartheta\| \leq 1\}$ *is uniformly integrable, and* (iii) $S_N/a_N \xrightarrow{p} 0$ *as* $N \to \infty$*.*

To highlight the relation between the CLT and the compact LIL in Banach spaces by means of Theorem 2.1, note that if the CLT holds, then condition (iii) of assertion (b) is fulfilled, as described above. Also, condition (ii) follows from the CLT, as the limiting Gaussian random variable Z with the same covariance as X has a strong second moment, due to the integrability properties of Gaussian random variables. This implies that K, the unit ball of the reproducing kernel Hilbert space associated to X, is compact and that the family $\{\vartheta^2(X) : \vartheta \in B', \|\vartheta\| \leq 1\}$ is uniformly integrable, as remarked previously. Hence, necessary and sufficient conditions for the compact LIL in the presence of the CLT reduce to condition (i) of Theorem 2.1(b), cf. Ledoux and Talagrand [1988], Corollary 1.3.

Eventually, note that for deriving almost sure rates of convergence in Section 2.2.2, we will use the compact LIL, even though the bounded LIL, guaranteeing the $\mathbb{Q}$-almost sure finiteness of $\Lambda(X)$, would be sufficient to establish most of our results. However, working with the compact LIL has the advantages of finding ourselves in the same setup in which the CLT and thus convergence rates in distribution have already been established. Further, it provides the ability to describe the set of limit points K by the $\mathbb{Q}$-almost sure relation $\Lambda(X) = \sup_{\mu \in K} \|\mu\| = \sigma(X)$, allowing for a better interpretation. Finally, the compact LIL will also lead to slightly better rates of convergence in probability.

2.1.2.2 General Arrays of Random Variables

We now also state results on the SLLN and the LIL for general arrays of i.i.d. B-valued random variables, which we require to establish the strong uniform consistency and uniform sample path bounds for objective function estimators in the VSAA setup, respectively. To derive these statements, we use available results for triangular arrays $\{\{X_{ki}\}_{i=1}^{k}\}_{k \in \mathbb{N}}$ of B-valued random variables, which may then be extended to general arrays $\{\{X_{ki}\}_{i=1}^{N_k}\}_{k \in \mathbb{N}}$ by considering strictly monotonically increasing sequences $\{N_k\}$.

The Strong Law of Large Numbers

By definition, the *SLLN* for a triangular array $\{\{X_{ki}\}_{i=1}^{k}\}_{k \in \mathbb{N}}$ of i.i.d. B-valued random variables with the same distribution as X and mean zero asserts that

$$\frac{\sum_{i=1}^{k} X_{ki}}{k} \to 0, \text{ as } k \to \infty, \tag{2.7}$$

$\mathbb{Q}$-almost surely, see, e.g., Ledoux and Talagrand [1991], Chapter 7. By the Borel-Cantelli lemma[7], it follows that for i.i.d. triangular arrays the almost sure convergence in (2.7) is

[7] See, e.g., Klenke [2008], Theorem 2.7.

equivalent to the *complete convergence* of the sequence $\{\sum_{i=1}^{k} X_{ki}/k\}$ to zero (note that complete convergence always implies the almost sure convergence), which states that

$$\sum_{k=1}^{\infty} \mathbb{Q}\bigg(\frac{\|\sum_{i=1}^{k} X_{ki}\|}{k} > \epsilon\bigg) < \infty, \qquad \text{for all } \epsilon > 0. \tag{2.8}$$

Now, addressing required conditions for the complete convergence of arrays to apply, Hu et al. [1989], Theorem 4, provide that assertion (2.8) holds if and only if X has a strong finite second moment. Hence, by their Proposition, p. 155, (or, alternatively, by argumentation via subsequences such that $\{\{X_{ki}\}_{i=1}^{N_k}\}$ forms a subset of $\{\{X_{ki}\}_{i=1}^{k}\}$), this result may be used to derive the following sufficient statement for general arrays $\{\{X_{ki}\}_{i=1}^{N_k}\}$ with a strictly monotonically increasing sequence $\{N_k\}$, from which then the SLLN for these arrays follows immediately.

Theorem 2.2. *Let $\{X_{ki}\}$, $X_{ki} \overset{d}{=} X$, be a general array of i.i.d. B-valued random variables with strictly monotonically increasing sequence $\{N_k\}$. If (i) $\mathbb{E}^{\mathbb{Q}}[\|X\|^2] < \infty$ and (ii) $\mathbb{E}^{\mathbb{Q}}[X] = 0$, then $\{S_k/N_k\}$ converges completely to 0 as $k \to \infty$, i.e.*

$$\sum_{k=1}^{\infty} \mathbb{Q}\bigg(\frac{\|S_k\|}{N_k} > \epsilon\bigg) < \infty, \qquad \textit{for all } \epsilon > 0.$$

Note that Hu et al. [1999], Corollary 4.1, provide a readily available result on the complete convergence for general arrays $\{X_{ki}\}$ of rowwise independent random vectors in Banach spaces (rowwise independence means that the random variables within each row are independent but that no independence is assumed between the rows), which thus also applies to the present case of i.i.d. random vectors. In particular, under suitable strong moment conditions linked to the sequence $\{N_k\}$ and the weak law of large numbers for $\{X_{ki}\}$, the corollary allows to consider also sequences $\{N_k\}$ that only require the constraint $\sum_{k=1}^{\infty} N_k^{-\alpha} < \infty$ for some $\alpha > 0$, see also Section 2.3.1. However, since this approach is technically more involved and to remain consistent with the subsequent presentation of the compact LIL, we omit such a more general treatment here and instead refer to Banholzer et al. [2018b].

The Law of the Iterated Logarithm

Under stronger assumptions on the B-valued mean zero random variable X, the characterisation of the compact LIL in Theorem 2.1(b) may be extended to triangular arrays $\{\{X_{ki}\}_{i=1}^{k}\}_{k\in\mathbb{N}}$ of i.i.d. B-valued copies, see Li et al. [1995], Theorem 2.1. In particular, this extension then allows to describe the sequence $\{\sum_{i=1}^{k} X_{ki}/\sqrt{2k\,\mathrm{Log}(k)}\}$ by

$$\limsup_{k\to\infty} \frac{\|\sum_{i=1}^{k} X_{ki}\|}{\sqrt{2k\,\mathrm{Log}(k)}} = \sigma(X), \tag{2.9}$$

$\mathbb{Q}$-almost surely, where $\sigma(X)$ is given by (2.4).

Since the limit superior of a subsequence is at most equal to the limit superior of the entire sequence, we may further infer from relation (2.9) that for a general array $\{\{X_{ki}\}_{i=1}^{N_k}\}_{k\in\mathbb{N}}$ of i.i.d. B-value random variables with a strictly monotonically increasing sequence $\{N_k\}$, the corresponding left-hand side of (2.9) must be at most equal to $\sigma(X)$. Hence, we obtain the following result, where we use the assumptions on X and the i.i.d. sequence $\{X_N\}$ as provided by Li et al. [1995] for triangular arrays (noting that convergence in probability of a sequence implies convergence in probability of a subsequence).

Theorem 2.3. *Let $\{X_{ki}\}$, $X_{ki} \overset{d}{=} X$, be a general array of i.i.d. B-valued random variables with strictly monotonically increasing sequence $\{N_k\}$. If (i) $\mathbb{E}^{\mathbb{Q}}[\|X\|^4/(\mathrm{Log}(\|X\|))^2] < \infty$, (ii) $\mathbb{E}^{\mathbb{Q}}[X] = 0$, and (iii) $\sum_{i=1}^{N_k} X_i/\sqrt{2N_k \,\mathrm{Log}(N_k)} \xrightarrow{p} 0$ as $k \to \infty$, where $\{X_N\}$ is a sequence of i.i.d. B-valued random variables following the distribution of X, then*

$$\limsup_{k\to\infty} \frac{\|S_k\|}{\sqrt{2N_k \,\mathrm{Log}(N_k)}} \leq \sigma(X),$$

$\mathbb{Q}$-almost surely.

In particular, thus note that the extension of the compact LIL to an array of B-valued random variables comes along with a change of the iterated LLog-rate to the simple Log.

Eventually, it is to be remarked that the above theorem may also be formulated under the same assumptions $(i) - (iii)$ for a general array $\{X_{ki}\}$ of identically distributed and rowwise independent B-valued random variables with a strictly monotonically increasing sequence $\{N_k\}$, using Li et al. [1995], Theorem 3.1. However, unlike in the preceding case of complete convergence, there is at the present time no readily available result on the compact LIL for general arrays of B-valued random variables, to the best of our knowledge.

2.1.3 Delta Method

When applying the CLT to derive asymptotic statements of optimal estimators, it is convenient to work with the delta method. Its main idea is to convert convergence in distribution of a sequence of random variables into convergence in distribution of a transformation thereof, where a suitable Taylor expansion is employed for approximation. If the method is intended to be used on Banach spaces, a sufficiently strong notion of directional differentiability is required, which is given by the concept of Hadamard and thus departs from those of Gâteaux and Fréchet.

Considering two Banach spaces B_1 and B_2 and a mapping $\varphi : B_1 \to B_2$, we first recall that φ is said to be *directionally differentiable* at a point $\mu \in B_1$ if the limit

$$\varphi'_\mu(v) := \lim_{t\searrow 0} \frac{\varphi(\mu + tv) - \varphi(\mu)}{t}$$

exists for all $v \in B_1$. The mapping φ is then called *Hadamard directionally differentiable* at μ if the directional derivative $\varphi'_\mu(v)$ exists for all $v \in B_1$ and, moreover, the following limit holds:

$$\varphi'_\mu(v) = \lim_{\substack{t \searrow 0 \\ \tilde{v} \to v}} \frac{\varphi(\mu + t\tilde{v}) - \varphi(\mu)}{t}. \tag{2.10}$$

In particular, on the directional differentiability in the sense of Hadamard, the following useful properties can be provided, cf. Shapiro [1990], Propositions 3.1 and 3.5.

Proposition 2.4. *Let B_1 and B_2 be Banach spaces, $\varphi : B_1 \to B_2$ a mapping and $\mu \in B_1$.*

(a) If $\varphi(\,\cdot\,)$ is Hadamard directionally differentiable at μ, then the directional derivative $\varphi'_\mu(\,\cdot\,)$ is continuous.

(b) If $\varphi(\,\cdot\,)$ is Lipschitz continuous in a neighbourhood of μ and directionally differentiable at μ, then $\varphi(\,\cdot\,)$ is Hadamard directionally differentiable at μ.

In a similar way to (2.10), it is possible to formulate second order directional differentiability in the Hadamard sense for a mapping $\varphi : B_1 \to B_2$. That is, φ is said to be *second order Hadamard directionally differentiable* at μ if the second order directional derivative φ''_μ, defined by

$$\varphi''_\mu(v) := \lim_{\substack{t \searrow 0 \\ \tilde{v} \to v}} \frac{\varphi(\mu + t\tilde{v}) - \varphi(\mu) - t\varphi'_\mu(\tilde{v})}{\frac{1}{2}t^2}, \tag{2.11}$$

exists for all $v \in B_1$. Concerning the second order Hadamard directional differentiability, the following assertions can be made, see Bonnans and Shapiro [2000], pp. 43.

Proposition 2.5. *Let B_1 and B_2 be Banach spaces, $\varphi : B_1 \to B_2$ be a mapping and $\mu \in B_1$.*

(a) If φ is twice continuously differentiable at μ, then φ is second order directionally differentiable at μ.

(b) If φ is Lipschitz continuous in a neighbourhood of μ and second order directionally differentiable at μ, then φ is second order Hadamard directionally differentiable at μ and thus φ''_μ is continuous.

Finally, by means of (2.10) and (2.11), the first and second order delta method can be stated as follows, cf. Shapiro [2000], Theorems 2.1 and 2.3.

Theorem 2.6. *Let B_1 and B_2 be Banach spaces equipped with their Borel σ-algebras, $\{X_N\}$ be a sequence of B_1-valued random variables, $\varphi : B_1 \to B_2$ be a mapping, and $\{b_N\}$ be a sequence of positive numbers tending to infinity as $N \to \infty$.*

(a) Suppose that B_1 is separable, that φ is first order Hadamard directionally differentiable at μ, and that $b_N(X_N - \mu) \xrightarrow{d} X$ as $N \to \infty$, for a B_1-valued random variable X. Then,

$$b_N\big(\varphi(X_N) - \varphi(\mu)\big) \xrightarrow{d} \varphi'_\mu(X), \quad \text{as } N \to \infty.$$

(b) *Suppose that B_1 is separable, that φ is first and second order Hadamard directionally differentiable at μ, and that $b_N(X_N - \mu) \xrightarrow{d} X$ as $N \to \infty$, for a B_1-valued random variable X. Then,*

$$b_N^2\big(\varphi(X_N) - \varphi(\mu) - \varphi'_\mu(X_N - \mu)\big) \xrightarrow{d} \tfrac{1}{2}\varphi''_\mu(X), \quad \text{as } N \to \infty.$$

2.2 Sample Average Approximation Strategy

In this section, we study almost sure convergence properties of the SAA strategy for approximately solving the original optimisation problem (2.1). For this purpose, recall that for an i.i.d. random sample $Z_1, \dots, Z_N$, drawn from the same distribution as Z, the (transformed) SAA problem is given by

$$\min_{x \in \mathcal{X}} \Big\{ \hat{f}_N(x) = g\big(\widehat{\Pi}_N(x) - C^{\mathrm{mkt}}\big) \Big\}, \tag{2.12}$$

where $g : \mathbb{R}^l \to \mathbb{R}_{\geq 0}$ denotes a continuous function measuring the difference between the Monte Carlo estimator $\widehat{\Pi}_N(x) = \frac{1}{N}\sum_{i=1}^N h(x, Z_i)$ and some given vector of market prices C^{mkt}. In particular, the optimal value $\hat{f}_N^*$ and an optimal solution $\hat{x}_N^*$ from the set of optimal solutions $\widehat{\mathcal{X}}_N^*$ of problem (2.12) are estimators for the optimal value f^* and an optimal solution $x^* \in \mathcal{X}^*$ of the original problem (2.1), respectively.

The convergence properties of the SAA approach have been discussed in a number of publications regarding different aspects, mostly within the original setup, i.e. where $\hat{f}_N = \frac{1}{N}\sum_{i=1}^N h(\,\cdot\,, Z_i)$ for a scalar-valued random function h. Under some relatively mild regularity assumptions on h, the strong consistency of the sequences of optimal estimators is shown in various publications, see, e.g., Domowitz and White [1982] and Bates and White [1985] for an approach via uniform convergence, or Dupačová and Wets [1988] and Robinson [1996] for a general approach based on the concept of epi-convergence.

Given consistency, it is meaningful to analyse the rate of convergence at which the optimal estimators approach their original counterparts as N tends to infinity. In this regard, Shapiro [1989, 1991, 1993] and King and Rockafellar [1993], for instance, provide necessary and sufficient conditions for the asymptotic distribution of the estimators, from which it immediately follows that $\{\hat{f}_N^*\}$ and $\{\hat{x}_N^*\}$ converge in distribution to their deterministic counterparts at a rate of $1/\sqrt{N}$. Whereas the findings of the former author are essentially based on the central limit theorem in Banach spaces to which the delta method with a first and second order expansion of the optimal value function is applied, the latter use a generalised implicit function theorem to achieve these results.

Rates of convergence for stochastic optimisation problems of above type are also studied by Kaniovski et al. [1995], Dai et al. [2000], Shapiro and Homem-de-Mello [2000] and Homem-de-Mello [2008]. Specifically, they use large deviation theory to show that, under the rather strong but unavoidable assumption of an existing moment generating function with a finite value in a neighbourhood of zero, the probability of deviation of the optimal solution $\hat{x}_N^*$ from the true solution x^* converges exponentially fast to zero for $N \to \infty$. This, in turn, allows to establish rates of convergence in probability and to derive conservative estimates for the sample size required to solve the original problem to given accuracy, see, e.g., Shapiro [2003] and Shapiro et al. [2014], Section 5.3, for further details.

In view of these findings, it has to be noted that all rates of convergence which have been established within the SAA framework so far are for convergence in probability or convergence in distributions. To the best of our knowledge, rates of convergence that hold almost surely and thus complement the strong consistency of optimal estimators with its corresponding rate have yet not been considered in this framework (barring very few exceptions using particular assumptions, see, e.g., Homem-de-Mello [2003]). Further, not many results are available on convergence in mean, nor have any convergence rates been established to this extent. This is an important issue, as convergence in mean is the basis to derive reasonable statements on the average biasedness of estimators.

In this section, we close this gap in the existing literature on the SAA approach and provide almost sure rates of convergence of the approximating objective function $\{\hat{f}_N\}$ as well as for the optimal estimators $\{\hat{f}_N^*\}$ and $\{\hat{x}_N^*\}$ (within the calibration setup). As already pointed out, our results can be derived by means of the LIL in Banach spaces, which is similar to the technique that has already been applied in the form of the functional CLT to obtain asymptotic distributions of optimal values and solutions, see, e.g., Shapiro [1991]. Considering the LIL, however, does not only provide almost sure results (instead of results in distribution) under comparable assumptions, but also yields useful results on rates of convergence in mean and in probability, as well as on rates of error probabilities. In particular, while rates in mean allow to quantify the asymptotic bias of the optimal values by their corresponding rate, (weak) rates of error probabilities may be derived via the Markov inequality and via an inequality of Einmahl and Li [2008], thus decreasing the gap between rates obtained from first or second moments and rates obtained by assuming exponential moments.

To initially provide the reader with the essential convergence properties of optimal estimators within the calibration setup, we begin in Section 2.2.1 by defining the underlying probability space and summarising well-known results on their measurability and their strong consistence. Following this, we first outline in Section 2.2.2 available results on convergence in distribution of the optimal estimators and their corresponding rates, for the purpose of better comparison. In analogy to these results, we then derive within the same setup by virtue of the LIL almost sure rates of convergence for the estimators and corresponding universal confidence sets. Eventually, in Section 2.2.3, we use the established almost sure rates of convergence to infer convergence in mean and in probability under particularly mild assumptions, along with their orders of convergence.

2.2.1 Probabilistic Setup and Strong Consistency

To be able to work with the SAA strategy and make probabilistic statements, let the sequence of i.i.d. $\mathcal{Z}$-valued random vectors $\{Z_i\}$, each Z_i with the same distribution as Z, be defined on a common probability space[8] $(\Omega_{\mathrm{S}}, \mathcal{F}_{\mathrm{S}}, \mathbb{Q}_{\mathrm{S}})$.

Measurability of Optimal Estimators

We first address the existence and measurability of the optimal estimators $\hat{f}_N^*$ and $\widehat{\mathcal{X}}_N^*$ in order to guarantee that they are indeed well-defined random quantities. In particular, note that both properties essentially hinge on the structure of h, which as a random function already provides that $h(x,\,\cdot\,)$ is measurable with respect to $\mathcal{B}(\mathcal{Z})$ and $\mathcal{B}(\mathbb{R}^l)$ for each $x \in \mathcal{X}$. It thus follows that the estimators $\widehat{\Pi}_N(x) = \widehat{\Pi}_N(x, \omega_{\mathrm{S}})$ and $\hat{f}_N(x) = g(\widehat{\Pi}_N(x, \omega_{\mathrm{S}}) - C^{\mathrm{mkt}})$, $\omega_{\mathrm{S}} \in \Omega_{\mathrm{S}}$, are measurable with respect to $\mathcal{F}_{\mathrm{S}}/\mathcal{B}(\mathbb{R}^l)$ and $\mathcal{F}_{\mathrm{S}}/\mathcal{B}(\mathbb{R})$, respectively, as sum and composition of measurable functions. Also, by the assumed $\mathbb{Q}^Z$-integrability of $h(x,\,\cdot\,)$ for each $x \in \mathcal{X}$, note that they are finite-valued, $\mathbb{Q}_{\mathrm{S}}$-almost surely.

If, in addition, $h(\,\cdot\,, Z)$ is assumed to be continuous on $\mathcal{X}$, $\mathbb{Q}$-almost surely, then h constitutes a so-called Carathéodory function, i.e. a function $h : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}^l$ such that $h(x,\,\cdot\,)$ is measurable for each $x \in \mathcal{X}$ and $h(\,\cdot\,, Z)$ continuous on $\mathcal{X}$ for almost every $Z \in \mathcal{Z}$, see, e.g., Aliprantis and Border [2006], Definition 4.50. Consequently, both $\widehat{\Pi}_N$ and $\hat{f}_N$ as defined on $\mathcal{X} \times \Omega_{\mathrm{S}}$ and with values in $\mathbb{R}^l$ and $\mathbb{R}$, respectively, are also Carathéodory functions. We are thus able to conclude the following result regarding the measurability of an optimal value $\hat{f}_N^* = \hat{f}_N^*(\omega_{\mathrm{S}})$ and the set of optimal solutions $\widehat{\mathcal{X}}_N^* = \widehat{\mathcal{X}}_N^*(\omega_{\mathrm{S}})$, cf. Aliprantis and Border [2006], Theorem 18.19.

Theorem 2.7. *Let $\mathcal{X} \subset \mathbb{R}^d$ be a nonempty compact set, and suppose that the objective function $\hat{f}_N : \mathcal{X} \times \Omega_S \to \mathbb{R}$ is a Carathéodory function. Then, the optimal value $\hat{f}_N^*(\omega_S) = \min_{x \in \mathcal{X}} \hat{f}_N(x, \omega_S)$ is measurable with respect to $\mathcal{F}_S/\mathcal{B}(\mathbb{R})$, and the set of optimal solutions $\widehat{\mathcal{X}}_N^*(\omega_S) = \operatorname{argmin}_{x \in \mathcal{X}} \hat{f}_N(x, \omega_S)$ is nonempty and compact for $\mathbb{Q}_S$-almost every $\omega_S \in \Omega_S$. In particular, the set $\widehat{\mathcal{X}}_N^*$ is measurable with respect to $\mathcal{F}_S/\mathcal{B}(\mathbb{R}^d)$, and there exists a $\mathcal{F}_S/\mathcal{B}(\mathbb{R}^d)$-measurable selection $\hat{x}_N^* = \hat{x}_N^*(\omega_S)$ from $\widehat{\mathcal{X}}_N^*$ such that $\hat{x}_N^*$ minimises $\hat{f}_N$ on $\mathcal{X}$, $\mathbb{Q}_S$-almost surely.*

Strong Consistency of Optimal Estimators

Having clarified the existence and measurability of optimal values and solutions to problem (2.12), we next turn to the strong consistency of these estimators. To this end, recall that a general sequence of estimators $\{\hat{\theta}_N\}$ is said to be *strongly consistent* for a parameter θ if it converges almost surely to θ, as N tends to infinity. Thus, strong consistency is typically considered a minimal requirement if one is to investigate an estimation procedure in the almost sure sense.

[8] For instance, the probability space $(\Omega_{\mathrm{S}}, \mathcal{F}_{\mathrm{S}}, \mathbb{Q}_{\mathrm{S}})$ may be chosen as the product space $(\Omega^{\mathbb{N}}, \mathcal{F}^{\otimes\mathbb{N}}, \mathbb{Q}^{\otimes\mathbb{N}})$, see, e.g., Klenke [2008], Chapter 14.

To show the strong consistency of the sequences of optimal values and solutions to their deterministic equivalents within the present calibration setup, we use the concept of uniform convergence as suggested by Shapiro et al. [2014], Section 5.1.1. Accordingly, we first provide weak conditions on the random function h (and the subsequent transformation g), such that the sequence of approximating objective functions $\{\hat{f}_N\}$ converges uniformly to f on $\mathcal{X}$, $\mathbb{Q}_{\mathrm{S}}$-almost surely, i.e. such that

$$\sup_{x\in\mathcal{X}}\left|\hat{f}_N(x) - f(x)\right| \to 0, \qquad \text{as } N \to \infty, \tag{2.13}$$

$\mathbb{Q}_{\mathrm{S}}$-almost surely, cf. Shapiro et al. [2014], p. 458. Note that the strong notion of uniform convergence is required in this context as the optimal estimators $\hat{f}_N^*$ and $\hat{x}_N^*$ are inferred over the entire parameter space $\mathcal{X}$. By contrast, the pointwise convergence of the objective functions, stating that $\hat{f}_N(x)$ converges to $f(x)$ for any $x \in \mathcal{X}$, $\mathbb{Q}_{\mathrm{S}}$-almost surely, as $N \to \infty$, depends on the parameter x and is therefore too weak for our purpose.

To arrive at (2.13), the following theorem allows to initially deduce almost sure uniform convergence of the pricing functions $\{\widehat{\Pi}_N\}$ under relatively weak assumptions. Note that the same assumptions also guarantee the finite-valuedness and continuity of Π on $\mathcal{X}$, which shall be used at later stages. For a proof of the statements, we refer to the version given by Shapiro et al. [2014], Theorem 7.53, for a scalar-valued function h, which can be extended straightforwardly to the present multivariate setup.

Theorem 2.8. *Let $\mathcal{X} \subset \mathbb{R}^d$ be a nonempty compact set and suppose that:*

(i) For any $x \in \mathcal{X}$, the function $h(\,\cdot\,, Z)$ is continuous at x, $\mathbb{Q}$-almost surely, i.e. $\mathbb{Q}(\{h(\,\cdot\,, Z)$ is continuous at $x\}) = 1$, and

(ii) there exists a measurable function $G : \mathcal{Z} \to \mathbb{R}_{\geq 0}$ such that $\mathbb{E}^{\mathbb{Q}}[G(Z)] < \infty$ and $\|h(x, Z)\| \leq G(Z)$ for all $x \in \mathcal{X}$, $\mathbb{Q}$-almost surely.

Then, the pricing function Π is finite-valued and continuous on $\mathcal{X}$, and $\{\widehat{\Pi}_N\}$ converges uniformly to Π on $\mathcal{X}$, $\mathbb{Q}_S$-almost surely, as $N \to \infty$, i.e. it holds

$$\sup_{x\in\mathcal{X}}\left\|\widehat{\Pi}_N(x) - \Pi(x)\right\| \to 0, \qquad \text{as } N \to \infty,$$

$\mathbb{Q}_S$-almost surely.

Remark 2.9. *The statement of Theorem 2.8 only requires $h(\,\cdot\,, Z)$ to be continuous at any given point $x \in \mathcal{X}$, $\mathbb{Q}$-almost surely. This, however, does not mean that $h(\,\cdot\,, Z)$ needs to be continuous on $\mathcal{X}$, $\mathbb{Q}$-almost surely. In fact, $h(\,\cdot\,, Z)$ may be discontinuous at some $x \in \mathcal{X}$ if the discontinuity occurs with probability zero, see, e.g., Shapiro et al. [2014], Remark 56.*

Given the continuity of g and the existence of a compact set $\mathcal{Y} \subset \mathbb{R}^l$ with $\Pi(x) \in \mathcal{Y}$ for all $x \in \mathcal{X}$, an immediate consequence of Theorem 2.8 is the almost sure uniform convergence of the sequence of objective functions $\{\hat{f}_N\}$ to f, cf. Gallant and White [1988], Lemma 3.4. In particular, note that this statement clearly also holds if assumption (ii) below is strengthened to assumption (G1) on the Lipschitz continuity of g.

Proposition 2.10. *Let $\mathcal{X} \subset \mathbb{R}^d$ be a nonempty compact set and suppose that:*

(i) $\{\widehat{\Pi}_N\}$ *converges uniformly to* Π *on* $\mathcal{X}$, $\mathbb{Q}_S$*-almost surely, as* $N \to \infty$, *and*

(ii) there exists a compact set $\mathcal{Y} \subset \mathbb{R}^l$ *such that for all* $x \in \mathcal{X}$, $\Pi(x)$ *is interior to* $\mathcal{Y}$.

Then, $\{\hat{f}_N\}$ *converges uniformly to* f *on* $\mathcal{X}$, $\mathbb{Q}_S$*-almost surely, as* $N \to \infty$.

Eventually, by the uniform convergence of $\{\hat{f}_N\}$ to f, we are able to conclude the strong consistency of the sequences of optimal estimators, cf. Shapiro et al. [2014], Proposition 5.2 and Theorem 5.3. Recall that $\operatorname{dev}(\widehat{\mathcal{X}}_N^*, \mathcal{X}^*)$ denotes the deviation of the set $\widehat{\mathcal{X}}_N^*$ from $\mathcal{X}^*$.

Theorem 2.11. *Let $\mathcal{X} \subset \mathbb{R}^d$ be a nonempty compact set and suppose that:*

(i) $\{\hat{f}_N\}$ *converges uniformly to* f *on* $\mathcal{X}$, $\mathbb{Q}_S$*-almost surely, as* $N \to \infty$.

Then, $\hat{f}_N^* \to f^*$, $\mathbb{Q}_S$*-almost surely, as* $N \to \infty$. *Moreover, if*

(ii) f *is continuous on* $\mathcal{X}$, *and*

(iii) for sufficiently large N, *the set* $\widehat{\mathcal{X}}_N^*$ *is nonempty,* $\mathbb{Q}_S$*-almost surely,*

then $\operatorname{dev}(\widehat{\mathcal{X}}_N^*, \mathcal{X}^*) \to 0$, $\mathbb{Q}_S$*-almost surely, as* $N \to \infty$.

Note that the first assertion of Theorem 2.11 obviously holds under the assumptions of Theorem 2.8 and assumption (*ii*) of Proposition 2.10. In that case, assumption (*ii*) of Theorem 2.11 is also met, cf. Theorem 2.8, such that the second assertion only requires assumption (*iii*) to be additionally satisfied. The latter is ensured, for instance, if the $\hat{f}_N$ are Carathéodory functions.

The assertion that $\operatorname{dev}(\widehat{\mathcal{X}}_N^*, \mathcal{X}^*) \to 0$, $\mathbb{Q}_\text{S}$-almost surely, as $N \to \infty$, implies that for any measurable global minimiser $\hat{x}_N^* \in \widehat{\mathcal{X}}_N^*$ of the approximating problem (2.12), it holds $\operatorname{dist}(\hat{x}_N^*, \mathcal{X}^*) \to 0$, $\mathbb{Q}_\text{S}$-almost surely, as $N \to \infty$. Hence, if the original problem (2.1) admits a unique optimal solution x^*, then $\hat{x}_N^* \to x^*$, $\mathbb{Q}_\text{S}$-almost surely, as $N \to \infty$.

2.2.2 Rates of Convergence

The previously outlined strong consistency of optimal estimators provides us with the assurance that the estimators approach their deterministic counterparts almost surely as N grows to infinity. Yet, no indication about the rate of convergence is given. This issue is now addressed in the present section in which we establish almost sure rates of convergence for optimal estimators. Our results are closely related to rates of convergence in distribution, which are mainly investigated for the basic SAA setup within the asymptotic analysis of optimal values and solutions by Shapiro [1989, 1991, 1993]. Therefore, we first review the main results of these studies applied to the calibration setup in Subsection 2.2.2.1, before establishing our findings on almost sure convergence rates in Section 2.2.2.2.

2.2.2.1 Rates of Convergence in Distribution

On the Banach space $C(\mathcal{X}, \mathbb{R}^l)$, we initially make the following assumptions with respect to the random function h:

(A1) For some $x_0 \in \mathcal{X}$ we have $\mathbb{E}^{\mathbb{Q}}[\|h(x_0, Z)\|^2] < \infty$.

(A2) There exists a measurable function $G : \mathcal{Z} \to \mathbb{R}_{\geq 0}$ such that $\mathbb{E}^{\mathbb{Q}}[G^2(Z)] < \infty$ and

$$\big\|h(x_1, Z) - h(x_2, Z)\big\| \leq G(Z)\|x_1 - x_2\|, \quad \forall x_1, x_2 \in \mathcal{X},$$

$\mathbb{Q}$-almost surely.

Assumptions (A1) and (A2) imply that $\mathbb{E}^{\mathbb{Q}}[\|h(x, Z)\|]$ and $\mathbb{E}^{\mathbb{Q}}[\|h(x, Z)\|^2]$ are finite-valued for all $x \in \mathcal{X}$. Moreover, assumption (A2) provides that Π is Lipschitz continuous on $\mathcal{X}$ and, as $\mathcal{X}$ is assumed to be compact and g continuous, thus guarantees that the set of minimisers $\mathcal{X}^*$ of the original problem (2.1) is nonempty. Further, it also follows from the compactness of $\mathcal{X}$, assumption (A2) and the continuity of g that $\hat{f}_N^*$ and $\widehat{\mathcal{X}}_N^*$ are measurable with respect to $\mathcal{F}_\mathrm{S}/\mathcal{B}(\mathbb{R})$ and $\mathcal{F}_\mathrm{S}/\mathcal{B}(\mathbb{R}^d)$, respectively, and that the latter set is nonempty, $\mathbb{Q}$-almost surely, with a measurable selection $\hat{x}_N^* \in \widehat{\mathcal{X}}_N^*$, cf. Theorem 2.7.

Most notably, assumptions (A1) and (A2) are sufficient to ensure that the $C(\mathcal{X}, \mathbb{R}^l)$-valued random variable $\widetilde{X} = h(\,\cdot\,, Z) - \mathbb{E}^{\mathbb{Q}}[h(\,\cdot\,, Z)]$ satisfies the CLT in this Banach space, see Araujo and Giné [1980], Corollary 7.17. We thus have

$$\sqrt{N}(\widehat{\Pi}_N - \Pi) \xrightarrow{d} \widetilde{Y}, \qquad \text{as } N \to \infty, \tag{2.14}$$

where $\widetilde{Y}$ denotes a $C(\mathcal{X}, \mathbb{R}^l)$-valued mean zero Gaussian random variable which is completely defined by the covariance of $\widetilde{X}$, that is by $(\mathbb{C}\mathrm{ov}^{\mathbb{Q}}\,\widetilde{X})(\vartheta_1, \vartheta_2) = \mathbb{E}^{\mathbb{Q}}[\vartheta_1(\widetilde{X})\vartheta_2(\widetilde{X})]$ for $\vartheta_1, \vartheta_2 \in C(\mathcal{X}, \mathbb{R}^l)'$. In particular, for any fixed $x \in \mathcal{X}$, we have that $\{\sqrt{N}(\widehat{\Pi}_N(x) - \Pi(x))\}$ converges in distribution to a real-valued multivariate normal distributed random variable $\widetilde{Y}(x)$ with mean zero and covariance matrix $\Sigma(x) = \mathbb{E}^{\mathbb{Q}}[(h(x, Z) - \Pi(x))(h(x, Z) - \Pi(x))^\top]$.

By further assuming (G1) and (G2), the mapping $g : C(\mathcal{X}, \mathbb{R}^l) \to C(\mathcal{X})$ is Lipschitz continuous and directionally differentiable at Π. It thus follows from Proposition 2.4(b) that g is Hadamard directionally differentiable at Π, such that the delta method for Banach spaces may be applied to (2.14), see Theorem 2.6(a). We therefore have the following lemma, extending the convergence in distribution of the scaled sequence of pricing functions to that of the respective objective functions.

Lemma 2.12. *Suppose that assumptions (A1)–(A2) and (G1)–(G2) hold. Then,*

$$\sqrt{N}(\hat{f}_N - f) \xrightarrow{d} g'_\Pi(\widetilde{Y}), \qquad \text{as } N \to \infty, \tag{2.15}$$

where $\widetilde{Y}$ denotes the $C(\mathcal{X}, \mathbb{R}^l)$-valued mean zero Gaussian random variable as obtained by (2.14) in $C(\mathcal{X}, \mathbb{R}^l)$.

Assertions (2.14) and (2.15) imply that both $\{\widehat{\Pi}_N\}$ and $\{\hat{f}_N\}$ converge in distribution to their deterministic counterparts Π and f, respectively, at a rate of $1/\sqrt{N}$. Moreover, since g is Hadamard directionally differentiable at Π, note that the directional derivative $g'_\Pi : C(\mathcal{X}, \mathbb{R}^l) \to C(\mathcal{X})$ is continuous and thus measurable with respect to $\mathcal{B}(C(\mathcal{X}, \mathbb{R}^l))$ and $\mathcal{B}(C(\mathcal{X}))$, cf. Proposition 2.4(a).

Rate of Convergence of Optimal Values

Provided that $\{\sqrt{N}(\hat{f}_N - f)\}$ converges in distribution to a random variable $g'_\Pi(\widetilde{Y})$ with values in $C(\mathcal{X})$, the convergence in distribution of $\{\sqrt{N}(\hat{f}^*_N - f^*)\}$ can be assessed using a first order expansion of the optimal value function, see Shapiro [1991]. To this end, let the optimal value function $\varkappa : C(\mathcal{X}) \to \mathbb{R}$ be defined by $\varkappa(\varphi) := \inf_{x\in\mathcal{X}} \varphi(x)$, $\varphi \in C(\mathcal{X})$, i.e. $\hat{f}^*_N = \varkappa(\hat{f}_N)$ and $f^* = \varkappa(f)$. Since $\mathcal{X}$ is compact, the mapping $\varkappa$ is continuous and hence measurable with respect to the Borel σ-algebras $\mathcal{B}(C(\mathcal{X}))$ and $\mathcal{B}(\mathbb{R})$. Moreover, $\varkappa$ is Lipschitz continuous with constant one, i.e. $|\varkappa(\varphi_1) - \varkappa(\varphi_2)| \leq \|\varphi_1 - \varphi_2\|_\infty$ for any $\varphi_1, \varphi_2 \in C(\mathcal{X})$, and it can be shown that $\varkappa$ is directionally differentiable at f with

$$\varkappa'_f(\varphi) = \inf_{x\in\mathcal{X}^*} \varphi(x), \quad \varphi \in C(\mathcal{X}), \tag{2.16}$$

see Danskin's theorem (e.g., Danskin [1966]), where $\mathcal{X}^*$ denotes the set of optimal solutions of the original problem (2.1). By the Lipschitz continuity and directional differentiability, it then follows that $\varkappa$ is also directionally differentiable at f in the Hadamard sense, see Proposition 2.4(b). Hence, an application of the first order delta method for Banach spaces with $\varkappa$ to (2.15) yields the following result, cf. Shapiro [1991], Theorem 3.2.

Theorem 2.13. *Suppose that assumptions (A1)–(A2) and (G1)–(G2) hold. Then,*

$$\sqrt{N}(\hat{f}^*_N - f^*) \xrightarrow{d} \varkappa'_f\big(g'_\Pi(\widetilde{Y})\big), \quad \text{as } N \to \infty, \tag{2.17}$$

where $\widetilde{Y}$ denotes the $C(\mathcal{X}, \mathbb{R}^l)$-valued mean zero Gaussian random variable as obtained by (2.14) *in $C(\mathcal{X}, \mathbb{R}^l)$, and $\varkappa'_f$ is given by* (2.16)*. In particular, if $\mathcal{X}^* = \{x^*\}$ is a singleton, then*

$$\sqrt{N}(\hat{f}^*_N - f^*) \xrightarrow{d} g'_\Pi\big(\widetilde{Y}(x^*)\big), \quad \text{as } N \to \infty. \tag{2.18}$$

From formulas (2.17) and (2.18), we are able to deduce that the convergence in distribution of $\{\hat{f}^*_N\}$ to its corresponding counterpart may also be quantified by the rate $1/\sqrt{N}$. Further, if $\varkappa'_f$ or g'_Π are linear[9] in their argument, then $\{\sqrt{N}(\hat{f}^*_N - f^*)\}$ is asymptotically normally distributed in (2.17) or (2.18), respectively.

[9] Considering, for instance, the least-squares formulation of g with $g(\varphi) = \|\varphi - C^{\text{mkt}}\|_2^2$ for $\varphi \in C(\mathcal{X}, \mathbb{R}^l)$ and $C^{\text{mkt}} \in \mathbb{R}^l$, we have $g'_\Pi(\varphi) = 2(\Pi - C^{\text{mkt}})^\top \varphi$.

Rate of Convergence of Optimal Solutions

Under more restrictive assumptions, it is possible to specify the rate of convergence of optimal solutions as well. The derivation of this result is essentially based on the CLT in the Banach space $C^1(\mathcal{X}, \mathbb{R}^l)$, to which the delta method with a second order expansion of the optimal value function $\varkappa$ is applied. This then provides a first order expansion for optimal solutions of the approximating problem (2.12).

For keeping our exposition on convergence of optimal solutions in this and the related section on almost sure rates as comprehensive as possible, we follow the general approach of Shapiro [2000] applied to the present calibration setup. In particular, we begin by making the following additional assumptions on the underlying random function h and its differential $\mathrm{D}_x h$, facilitating convergence in distribution in $C^1(\mathcal{X}, \mathbb{R}^l)$:

(A3) The function $h(\,\cdot\,, Z)$ is continuously differentiable on $\mathcal{X}$, $\mathbb{Q}$-almost surely.

and

(A1') For some $x_0 \in \mathcal{X}$ we have $\mathbb{E}^{\mathbb{Q}}[\|\mathrm{D}_x h(x_0, Z)\|^2] < \infty$.

(A2') The differential $\mathrm{D}_x h(\,\cdot\,, Z)$ is Lipschitz continuous with constant $G(Z)$ on $\mathcal{X}$, $\mathbb{Q}$-almost surely, and $\mathbb{E}^{\mathbb{Q}}[G^2(Z)] < \infty$.

Assumption (A3) implies that $\widehat{\Pi}_N$ is a random variable with values in $C^1(\mathcal{X}, \mathbb{R}^l)$, and assumptions (A1)–(A3) together imply that Π is continuously differentiable on $\mathcal{X}$ with $\mathrm{D}\Pi(x) = \mathbb{E}^{\mathbb{Q}}[\mathrm{D}_x h(x, Z)]$ for $x \in \mathcal{X}$. Specifically, while the differentiability of Π and the interchange of differential operator and expectation is shown, for instance, in Shapiro et al. [2014], Theorem 7.49, the continuity of $\mathrm{D}\Pi$ is provided by applying Theorem 2.8 to $\mathrm{D}_x h$. Moreover, all assumptions (A1)–(A3) and (A1')–(A2') entail that $\widetilde{X} = h(\,\cdot\,, Z) - \mathbb{E}^{\mathbb{Q}}[h(\,\cdot\,, Z)]$ also satisfies the CLT in the Banach space $C^1(\mathcal{X}, \mathbb{R}^l)$, such that (2.14) holds for a $C^1(\mathcal{X}, \mathbb{R}^l)$-valued mean zero Gaussian random variable $\widetilde{Y}$.

Given the convergence (2.14) in $C^1(\mathcal{X}, \mathbb{R}^l)$, it can further be transferred to the level of objective functions on the separable Banach space $C^1(\mathcal{X})$. Accordingly, assuming that the mapping $g : C^1(\mathcal{X}, \mathbb{R}^l) \to C^1(\mathcal{X})$ satisfies (G1) and (G2), it is Hadamard directionally differentiable at Π and $\widehat{\Pi}_N$ by Proposition 2.4(b). Hence, the directional derivatives g'_Π and $g'_{\widehat{\Pi}_N}$ from $C^1(\mathcal{X}, \mathbb{R}^l)$ into $C^1(\mathcal{X})$ are continuous and measurable with respect to $\mathcal{B}(C^1(\mathcal{X}, \mathbb{R}^l))$ and $\mathcal{B}(C^1(\mathcal{X}))$, see Proposition 2.4(a), such that the compositions $f = g(\Pi)$ and $\hat{f}_N = g(\widehat{\Pi}_N)$ are elements of and with values in $C^1(\mathcal{X})$, respectively. Under all assumptions (A1)–(A3), (A1')–(A2') and (G1)–(G2), the first order delta method thus guarantees the validity of (2.15) in $C^1(\mathcal{X})$.

Note that by considering the class $C^1(\mathcal{X}, \mathbb{R}^l)$ of continuously differentiable functions and assumptions (A1')–(A2'), we implicitly assume that the pricing functions Π and $\widehat{\Pi}_N$ as well as their differentials are sufficiently well-behaved. This also holds for the related objective functions f and $\hat{f}_N$ in $C^1(\mathcal{X})$, and presents a reasonable regularity condition in order to derive general rates of convergence. If a respective function does not meet these criteria, a similar deduction becomes considerably more difficult.

Aside from conditions on h and $\mathrm{D}_x h$ and the usual assumptions on g, let us further consider the following regularity assumptions for the original problem (2.1):

(B1) The problem (2.1) has a unique optimal solution $x^* \in \mathcal{X}$.

(B2) The function f satisfies the second-order growth condition at x^*, i.e. there exists $\alpha > 0$ and a neighbourhood V_{x^*} of x^* such that

$$f(x) \geq f(x^*) + \alpha \|x - x^*\|^2, \quad \forall x \in \mathcal{X} \cap V_{x^*}. \tag{2.19}$$

(B3) The set $\mathcal{X}$ is second order regular at x^*.

(B4) The function f is twice continuously differentiable in a neighbourhood of the point x^*.

Assumptions (B1)–(B4) represent standard second order optimality conditions to be found in common literature on perturbation analysis of optimisation problems, see, e.g., Bonnans and Shapiro [2000]. While assumptions (B1) and (B4) are self-explanatory, the growth condition (2.27) involves that x^* is locally optimal and that f increases at least quadratically near x^*. This condition can be ensured to hold in several ways by assuming second order sufficient conditions as given, for instance, in Section 3.3 of Bonnans and Shapiro [2000]. Finally, the second order regularity of $\mathcal{X}$ in (B3) concerns the tangent set $T^2_{\mathcal{X}}(x^*, v)$ to $\mathcal{X}$ at x^* in direction v and guarantees that it is a sufficient good second order approximation to $\mathcal{X}$ in direction v.

By imposing (B1)–(B4), a second order expansion of the minimal value function $\varkappa$, now mapping $C^1(\mathcal{X})$ into $\mathbb{R}$, can be calculated, along with a first order expansion of the associated optimal solution function $\chi : C^1(\mathcal{X}) \to \mathbb{R}^d$, where $\chi(\varphi) \in \operatorname{argmin}_{x \in \mathcal{X}} \varphi(x)$, $\varphi \in C^1(\mathcal{X})$. More precisely, under (B1)–(B4), $\varkappa$ is shown to be first and second order Hadamard directionally differentiable at f, with $\varkappa'_f(\varphi) = \varphi(x^*)$ and

$$\varkappa''_f(\varphi) = \inf_{v \in C_{x^*}} \Big\{ 2v^\top \nabla \varphi(x^*) + v^\top \nabla^2 f(x^*) v + \inf_{z \in T^2_{\mathcal{X}}(x^*, v)} z^\top \nabla f(x^*) \Big\}, \tag{2.20}$$

for $\varphi \in C^1(\mathcal{X})$, and where C_{x^*} is the critical cone of problem (2.1), $\nabla^2 f(x^*)$ the Hessian matrix of f at x^*, and $T^2_{\mathcal{X}}(x^*, v)$ denotes the second order tangent set to $\mathcal{X}$ at x^* in direction v (see, e.g., Shapiro [2000], Theorem 4.1). Moreover, if the problem on the right-hand side of (2.20) admits a unique solution $v^*(\varphi)$, then the mapping χ is also Hadamard directionally differentiable at f, and $\chi'_f(\varphi) = v^*(\varphi)$ holds.

Eventually, we observe that an application of the second order delta method to the convergence (2.15) in $C^1(\mathcal{X})$ requires $g'_\Pi(\widetilde{Y})$ to be an element of the same space, cf. Theorem 2.6(b). This, however, can only be guaranteed if we additionally make the following assumption on the transforming function g:

(G3) The mapping g is twice continuously differentiable at Π.

By Proposition 2.5, assumption (G3) then implies together with (G1) that g is second order directionally differentiable and that g''_Π is continuous, thus providing $g'_\Pi(\widetilde{Y}) \in C^1(\mathcal{X})$. In particular, (G3) also reinforces the regularity assumption (B4), such that the latter can be dispensed with in order to apply the second order delta method with $\varkappa$ to the convergence (2.15) in $C^1(\mathcal{X})$, cf. Shapiro [2000], Theorems 4.2 and 4.3.

Theorem 2.14. *Suppose that assumptions (A1)–(A3), (A1')–(A2'), (B1)–(B3) and (G1)–(G3) hold. Then,*

$$N\big(\hat{f}_N^* - \hat{f}_N(x^*)\big) \xrightarrow{d} \tfrac{1}{2}\varkappa''_f\big(g'_\Pi(\widetilde{Y})\big), \qquad \text{as } N \to \infty, \tag{2.21}$$

where $\widetilde{Y}$ denotes the $C^1(\mathcal{X},\mathbb{R}^l)$-valued mean zero Gaussian random variable as obtained by (2.14) *in $C^1(\mathcal{X},\mathbb{R}^l)$, and $\varkappa''_f$ is given by* (2.20)*. Further, suppose that for any $\varphi \in C^1(\mathcal{X})$ the problem on the right-hand side of* (2.20) *has a unique solution $v^*(\varphi)$. Then,*

$$\sqrt{N}(\hat{x}_N^* - x^*) \xrightarrow{d} v^*\big(g'_\Pi(\widetilde{Y})\big), \qquad \text{as } N \to \infty. \tag{2.22}$$

Theorem 2.14 yields the usual convergence rate for an optimal solution, concerning convergence in distribution. Nevertheless, it does not directly imply any result on convergence in mean, nor on the bias. Although the expectation of the right-hand side of (2.22) is finite, this is not necessarily the case for the upscaled difference of the optimal solutions on the left. The limit of the expectations of the left-hand side only exists and equals the expectation of the right-hand side if and only if the upscaled sequence is uniformly integrable, see, e.g.,Serfling [1980], Theorem 1.4(A). However, making such an assumption for $\{\sqrt{N}(\hat{x}_N^* - x^*)\}$ is actually already equivalent to imposing a convergence order of $\mathcal{O}(1/\sqrt{N})$ for $\{\hat{x}_N^* - x^*\}$ to zero in the L_1-sense.

2.2.2.2 Almost Sure Rates of Convergence

We now turn to almost sure convergence and first observe that in the specific case of $C(\mathcal{X},\mathbb{R}^l)$-valued random variables, the compact LIL is satisfied under exactly the same assumptions as the CLT in the Banach space setting, see Kuelbs [1976a], Theorem 4.4. Accordingly, given assumptions (A1) and (A2), we have for the $C(\mathcal{X},\mathbb{R}^l)$-valued random variable $\widetilde{X} = h(\,\cdot\,,Z) - \mathbb{E}^{\mathbb{Q}}[h(\,\cdot\,,Z)]$ and the related sequence of i.i.d. copies $\{\widetilde{X}_i\}$ that

$$\lim_{N\to\infty} \operatorname{dist}\left(\frac{\sqrt{N}(\widehat{\Pi}_N - \Pi)}{\sqrt{2\,\mathrm{LLog}(N)}}, K_{\widetilde{X}}\right) = 0, \qquad \text{and} \qquad \mathrm{CP}\left(\left\{\frac{\sqrt{N}(\widehat{\Pi}_N - \Pi)}{\sqrt{2\,\mathrm{LLog}(N)}}\right\}\right) = K_{\widetilde{X}},$$

each $\mathbb{Q}_{\mathrm{S}}$-almost surely, where $K_{\widetilde{X}}$ denotes the unit ball of the reproducing kernel Hilbert space $H_{\widetilde{X}}$ associated to the covariance of $\widetilde{X}$ and $K_{\widetilde{X}}$ is compact. In line with Subsection 2.1.2.1, it follows from this result that

$$\Lambda(\widetilde{X}) = \limsup_{N\to\infty} \frac{\sqrt{N}\big\|\widehat{\Pi}_N - \Pi\big\|_\infty}{\sqrt{2\,\mathrm{LLog}(N)}} = \sigma(\widetilde{X}), \tag{2.23}$$

$\mathbb{Q}_S$-almost surely, where $\sigma(\widetilde{X}) = \sup_{\|\vartheta\|\leq 1}(\mathbb{E}^{\mathbb{Q}}[\vartheta^2(\widetilde{X})])^{1/2}$, $\vartheta \in C(\mathcal{X},\mathbb{R}^l)'$. Now, by virtue of Singer's extension of the Riesz respresentation theorem to vector-valued functions (e.g., Singer [1970], Chapter II, Lemma 1.6), the dual space $C(\mathcal{X},\mathbb{R}^l)'$ can be identified with the space $M^{\mathrm{reg}}(\mathcal{X},\mathbb{R}^l)$ of all regular $\mathbb{R}^l$-valued measures on $\mathcal{X}$ that are of bounded variation. Moreover, according to Singer [1970], Chapter II, Lemma 1.7, the extreme points of the set $\{m^{\mathrm{reg}} \in M^{\mathrm{reg}}(\mathcal{X},\mathbb{R}^l) : \|m^{\mathrm{reg}}\| \leq 1\}$ are of the form $m^{\mathrm{reg}} = \vartheta^\top \widetilde{X}(x)$ for an extreme point ϑ of $\{\vartheta \in \mathbb{R}^l : \|\vartheta\| \leq 1\}$, a $C(\mathcal{X},\mathbb{R}^l)$-valued random variable $\widetilde{X}$ and $x \in \mathcal{X}$. Hence, $\sigma(\widetilde{X})$ can be shown to assume the form[10]

$$\sigma(\widetilde{X}) = \sup_{x\in\mathcal{X}} \big\|\Sigma^{1/2}(x)\big\|, \tag{2.24}$$

where $\|\cdot\|$ denotes the matrix norm and $\Sigma(x)$ the symmetric and positive definite covariance matrix of the random variable $\widetilde{X}(x)$ at the point x.

By definition of the limit superior, equation (2.23) implies that for any $\epsilon > 0$, there exists a finite random variable $N^* = N^*(\epsilon) \in \mathbb{N}$ such that

$$\forall N \geq N^* : \quad \|\widehat{\Pi}_N - \Pi\|_\infty \leq \frac{\sqrt{(2+\epsilon)\,\mathrm{LLog}(N)}}{\sqrt{N}}\sigma(\widetilde{X}),$$

$\mathbb{Q}_S$-almost surely, specifying the speed of convergence of the approximating pricing function in the almost sure sense. The same almost sure rate of convergence can be taken over to the respective objective functions, provided that assumption (G1) on the Lipschitz continuity of $g : C(\mathcal{X},\mathbb{R}^l) \to C(\mathcal{X})$ with constant L_g is postulated.

Lemma 2.15. *Suppose that assumptions (A1)–(A2) and (G1) hold. Then, for any $\epsilon > 0$, there exists a finite random variable $N^* = N^*(\epsilon) \in \mathbb{N}$ such that*

$$\forall N \geq N^* : \quad \|\hat{f}_N - f\|_\infty \leq L_g\frac{\sqrt{(2+\epsilon)\,\mathrm{LLog}(N)}}{\sqrt{N}}\sigma(\widetilde{X}), \tag{2.25}$$

$\mathbb{Q}_S$-almost surely, where $\sigma(\widetilde{X})$ is given by (2.24) *for the $C(\mathcal{X},\mathbb{R}^l)$-valued random variable $\widetilde{X}$.*

In particular, inequality (2.25) reveals that the convergence of $\{\hat{f}_N\}$ to f occurs at a rate of $\mathcal{O}(\sqrt{\mathrm{LLog}(N)}/\sqrt{N})$, which is only marginally slower than the rate $1/\sqrt{N}$ obtained from convergence in distribution (to get an idea for the scale involved, note that $\sqrt{\log\log 10^{99}} \approx 2.33$). Yet, unlike the latter, the rate $\sqrt{\mathrm{LLog}(N)}/\sqrt{N}$ holds $\mathbb{Q}_S$-almost surely, which is a different notion of convergence than convergence in distribution. In any case, however, it is not possible to determine the exact N^* for which (2.25) holds, as this depends on the particular realisation of the underlying random sequence $\{Z_i\}$.

[10] To aid intuition, note that when $\|\cdot\|$ stands for the Euclidean norm, then $\|\Sigma^{1/2}(x)\|$ corresponds to the square root of the largest eigenvalue of the covariance matrix $\Sigma(x)$ at $x \in \mathcal{X}$.

Rate of Convergence of Optimal Values

Once having assertion (2.25), the rate of convergence of the optimal value $\{\hat{f}_N^*\}$ to f^* is easily obtained by recalling the Lipschitz continuity of the continuous optimal value function $\varkappa(\varphi) = \inf_{x\in\mathcal{X}} \varphi(x)$, with $\hat{f}_N^* = \varkappa(\hat{f}_N)$ and $f^* = \varkappa(f)$. We thus have the following result, in analogy to Theorem 2.13.

Theorem 2.16. *Suppose that assumptions (A1)–(A2) and (G1) hold. Then,*

$$\forall N \in \mathbb{N}: \quad |\hat{f}_N^* - f^*| \leq \|\hat{f}_N - f\|_\infty.$$

In particular, it holds that $\{\hat{f}_N^*\}$ converges to f^*, $\mathbb{Q}_S$-almost surely, at a rate of $\mathcal{O}(\sqrt{\mathrm{LLog}(N)}/\sqrt{N})$.

Rate of Convergence of Optimal Solutions

Next, we proceed with analysing the rate of convergence of optimal solutions under the almost sure criterion. Considering the space $C(\mathcal{X}, \mathbb{R}^l)$ of continuous functions on $\mathcal{X}$, we note first of all that if the random function h only satisfies the moment and Lipschitz conditions (A1) and (A2), respectively, and g is Lipschitz continuous according to (G1), then a slower rate of almost sure convergence can be obtained under the regularity conditions (B1) and (B2), as the following proposition shows.

Proposition 2.17. *Suppose that assumptions (A1)–(A2), (B1)–(B2) and (G1) hold. Then, there exists a finite random variable $N^* \in \mathbb{N}$ such that*

$$\forall N \geq N^*: \quad \|\hat{x}_N^* - x^*\|^2 \leq \frac{2}{\alpha}\|\hat{f}_N - f\|_\infty, \tag{2.26}$$

$\mathbb{Q}_S$-almost surely. In particular, it holds that $\{\hat{x}_N^*\}$ converges to x^*, $\mathbb{Q}_S$-almost surely, at a rate of $\mathcal{O}((\sqrt{\mathrm{LLog}(N)}/\sqrt{N})^{1/2})$.

Proof. By assumptions (A1)–(A2), (B1) and the continuity of g, $\{\hat{x}_N^*\}$ converges to x^*, $\mathbb{Q}_S$-almost surely, as $N \to \infty$, see Theorems 2.8 and 2.11 together with Proposition 2.10. This implies that $x_N^* \in V_{x^*}$ holds $\mathbb{Q}_S$-almost surely for $N \geq N^*$ for some finite $N^* \in \mathbb{N}$. Hence, the second-order growth condition (B2) at x^* with $\alpha > 0$ yields

$$\begin{aligned}
\|\hat{x}_N^* - x^*\|^2 &\leq \frac{1}{\alpha}\big(f(\hat{x}_N^*) - f(x^*)\big) \\
&= \frac{1}{\alpha}\Big(f(\hat{x}_N^*) - \hat{f}_N(\hat{x}_N^*) + \hat{f}_N(\hat{x}_N^*) - f(x^*)\Big) \\
&\leq \frac{1}{\alpha}\Big(f(\hat{x}_N^*) - \hat{f}_N(\hat{x}_N^*) + \hat{f}_N(x^*) - f(x^*)\Big) \\
&\leq \frac{1}{\alpha}\Big(\big|\hat{f}_N(\hat{x}_N^*) - f(\hat{x}_N^*)\big| + \big|\hat{f}_N(x^*) - f(x^*)\big|\Big) \\
&\leq \frac{2}{\alpha}\|\hat{f}_N - f\|_\infty,
\end{aligned}$$

where $\hat{f}_N(\hat{x}_N^*)$ has been added and subtracted from the first line to the second and $\hat{f}_N(\hat{x}_N^*) \leq \hat{f}_N(x^*)$ has been used from the second line to the third. This proves (2.26), and the remaining assertion then follows from Lemma 2.15. □

To achieve a faster rate of almost sure convergence, stronger assumptions on h and the differential $D_x h$ in the subspace $C^1(\mathcal{X}, \mathbb{R}^l)$ are required, as described in Section 2.2.2.1 for convergence in distribution. Specifically, if, in addition to assumptions (A1) and (A2), we assume that h is also continuously differentiable on $\mathcal{X}$, i.e. assumption (A3) holds, then Π is an element of the Banach space $C^1(\mathcal{X}, \mathbb{R}^l)$ and $\widehat{\Pi}_N$ is $C^1(\mathcal{X}, \mathbb{R}^l)$-valued. Consequently, on condition that the moment and Lipschitz assumptions of $D_x h$ in (A1') and (A2'), respectively, are also fulfilled, $\widetilde{X}$ satisfies the compact LIL in $C^1(\mathcal{X}, \mathbb{R}^l)$. Together with assumptions (G1) and (G2), providing that the directional derivatives of $g : C^1(\mathcal{X}, \mathbb{R}^l) \to C^1(\mathcal{X})$ at Π and $\widehat{\Pi}_N$ are continuous (see Proposition 2.4) and thus ensure that f and $\hat{f}_N$ are well-defined elements of and with values in $C^1(\mathcal{X})$, respectively, we can state the following, cf. Lemma 2.15.

Lemma 2.18. *Suppose that assumptions (A1)–(A3), (A1')–(A2') and (G1)–(G2) hold. Then, for any $\epsilon > 0$, there exists a finite random variable $N^* = N^*(\epsilon) \in \mathbb{N}$ such that*

$$\forall N \geq N^* : \quad \|\hat{f}_N - f\|_{1,\infty} \leq L_g \frac{\sqrt{(2+\epsilon)\,\mathrm{LLog}(N)}}{\sqrt{N}} \sigma(\widetilde{X}),$$

$\mathbb{Q}_S$-almost surely, where $\sigma(\widetilde{X})$ is given in general form by (2.4) for the $C^1(\mathcal{X}, \mathbb{R}^l)$-valued random variable $\widetilde{X}$.

Moreover, we further consider the regularity assumptions (B1) and (B2) on the original problem (2.1), where we marginally strengthen the latter according to:

(B2') The function f satisfies the second-order growth condition at x^*, i.e. there exists $\alpha > 0$ and a neighbourhood V_{x^*} of x^* such that

$$f(x) \geq f(x^*) + \alpha \|x - x^*\|^2, \quad \forall x \in \mathcal{X} \cap V_{x^*}. \tag{2.27}$$

Further, V_{x^*} can be chosen such that $\mathcal{X} \cap V_{x^*}$ is star-shaped with centre x^*.

We are then able to derive the following result on the speed of convergence of optimal solutions of the SAA problem. Note that this result holds in parallel to Theorem 2.14 in the almost sure case.

Theorem 2.19. *Suppose that assumptions (A1)–(A3), (A1')–(A2'), (B1) and (B2'), and (G1)–(G2) hold. Then, there exists a finite random variable $N^* \in \mathbb{N}$ such that*

$$\forall N \geq N^* : \quad \|\hat{x}_N^* - x^*\| \leq \frac{1}{\alpha} \|\hat{f}_N - f\|_{1,\infty}, \tag{2.28}$$

$\mathbb{Q}_S$-almost surely. In particular, it holds that $\{\hat{x}_N^*\}$ converges to x^*, $\mathbb{Q}_S$-almost surely, at a rate of $\mathcal{O}(\sqrt{\mathrm{LLog}(N)}/\sqrt{N})$.

Proof. Again, by postulating (A1)–(A2), (B1) and the continuity of g, $\{\hat{x}_N^*\}$ converges to x^*, $\mathbb{Q}_\text{S}$-almost surely, as $N \to \infty$. This implies that $\hat{x}_N^* \in V_{x^*}$ holds $\mathbb{Q}_\text{S}$-almost surely for $N \geq N^*$ for some finite random $N^* \in \mathbb{N}$. Hence, the second-order growth condition (B2') at x^* with $\alpha > 0$ yields

$$\begin{aligned}
\|\hat{x}_N^* - x^*\|^2 &\leq \frac{1}{\alpha}\big(f(\hat{x}_N^*) - f(x^*)\big) \\
&\leq \frac{1}{\alpha}\Big(f(\hat{x}_N^*) - \hat{f}_N(\hat{x}_N^*) + \hat{f}_N(\hat{x}_N^*) - f(x^*)\Big) \\
&\leq \frac{1}{\alpha}\Big(f(\hat{x}_N^*) - \hat{f}_N(\hat{x}_N^*) + \hat{f}_N(x^*) - f(x^*)\Big) \\
&\leq \frac{1}{\alpha}\Big(f(\hat{x}_N^*) - \hat{f}_N(\hat{x}_N^*) - \big(f(x^*) - \hat{f}_N(x^*)\big)\Big),
\end{aligned}$$

and therefore

$$\|\hat{x}_N^* - x^*\| \leq \frac{\big|f(\hat{x}_N^*) - \hat{f}_N(\hat{x}_N^*) - \big(f(x^*) - \hat{f}_N(x^*)\big)\big|}{\alpha\|\hat{x}_N^* - x^*\|}.$$

Since $(f - \hat{f}_N)$ is assumed to be differentiable on $\mathcal{X}$, $\mathbb{Q}_\text{S}$-almost surely, and $\mathcal{X} \cap V_{x^*}$ is star-shaped with centre x^*, it further holds by the mean value theorem (e.g., Dieudonné [1960], Theorem 8.5.4) that

$$\begin{aligned}
&\frac{\big|f(\hat{x}_N^*) - \hat{f}_N(\hat{x}_N^*) - \big(f(x^*) - \hat{f}_N(x^*)\big)\big|}{\alpha\|\hat{x}_N^* - x^*\|} \\
&\qquad \leq \frac{1}{\alpha} \sup_{0 \leq t \leq 1} \big\|\nabla\big(f(\hat{x}_N^* + t(x^* - \hat{x}_N^*)) - \hat{f}_N(\hat{x}_N^* + t(x^* - \hat{x}_N^*))\big)\big\| \\
&\qquad \leq \frac{1}{\alpha} \sup_{x \in \mathcal{X} \cap V_{x^*}} \big\|\nabla\big(f(x) - \hat{f}_N(x)\big)\big\|.
\end{aligned}$$

Thus, by definition of the norm $\|\cdot\|_{1,\infty}$, the latter then provides inequality (2.28), and applying Lemma 2.18 eventually yields the assertion on the rate of convergence. □

It is to be remarked that the results established in Proposition 2.17 and Theorem 2.19 require fewer assumptions on the objective function f than the corresponding Theorem 2.14 on convergence in distribution, while providing almost sure convergence instead of convergence in distribution. This becomes most notable in that the former results are able to dispense with assumptions (B3) and (G3), while these are necessary to ensure $g'_\Pi(\widetilde{Y}) \in C^1(\mathcal{X})$ and the second order Hadamard directional derivative $\varkappa''_f$ in the latter.

2.2.3 Further Results

In addition to the previous section on almost sure rates of convergence, we now derive some further results from our analysis of the LIL in Banach spaces, concerning convergence in mean and in probability. Specifically, in Section 2.2.3.1, we infer from a characterisation of the compact LIL that the optimal estimators also convergence in mean and derive the corresponding rates of convergence. In particular, these rates can be used to quantify the asymptotic bias of optimal estimators without the additional (strong) assumption of uniform integrability. Based on the obtained results on convergence in mean, Section 2.2.3.2 then provides rates of convergence in probability for optimal estimators as well as (slow) rates of error probabilities under considerably mild conditions. The latter is opposed to other approaches yielding (fast) exponential rates of convergence but relying on a strong exponential moment condition (or boundedness condition).

2.2.3.1 Rates of Convergence in Mean

By recalling that the $C(\mathcal{X},\mathbb{R}^l)$-valued random variable $\widetilde{X} = h(\cdot\,,Z) - \mathbb{E}^{\mathbb{Q}}[h(\cdot\,,Z)]$ satisfies the compact LIL under assumptions (A1) and (A2), we can apply Kuelbs's equivalence (2.6) (cf. also Kuelbs [1977], Theorem 4.1) to obtain

$$\mathbb{E}^{\mathbb{Q}_{\mathrm{S}}}\Big[\big\|\sum_{i=1}^{N}\widetilde{X}_i\big\|_\infty\Big] = o(a_N).$$

This, in turn, implies

$$\lim_{N\to\infty}\frac{\sqrt{N}}{\sqrt{2\,\mathrm{LLog}(N)}}\,\mathbb{E}^{\mathbb{Q}_{\mathrm{S}}}\Big[\|\widehat{\Pi}_N - \Pi\|_\infty\Big] = 0, \tag{2.29}$$

which, by the Lipschitz continuity of g, lets us state the following main proposition for the involved objective functions.

Proposition 2.20. *Suppose that assumptions (A1)–(A2) and (G1) hold. Then,*

$$\lim_{N\to\infty}\frac{\sqrt{N}}{\sqrt{2\,\mathrm{LLog}(N)}}\,\mathbb{E}^{\mathbb{Q}_S}\Big[\|\hat{f}_N - f\|_\infty\Big] = 0,$$

i.e. $\mathbb{E}^{\mathbb{Q}_S}[\|\hat{f}_N - f\|_\infty] = o(\sqrt{\mathrm{LLog}(N)}/\sqrt{N})$*, and in particular* $\{\hat{f}_N\}$ *converges to* f *in* $L_1(C(\mathcal{X}))$*, at a rate of* $o(\sqrt{\mathrm{LLog}(N)}/\sqrt{N})$*.*

Rates of Convergence of Optimal Values and Biasedness

By the Lipschitz continuity of the optimal value function $\varkappa(\varphi) = \inf_{x\in\mathcal{X}}\varphi(x)$, $\varphi \in C(\mathcal{X})$, and Proposition 2.20, we immediately arrive at the following result for the convergence of optimal values.

Theorem 2.21. *Suppose that assumptions (A1)–(A2) and (G1) hold. Then,* $\{\hat{f}_N^*\}$ *converges to* f^* *in* L_1*, and* $\mathbb{E}^{\mathbb{Q}_S}[|\hat{f}_N^* - f^*|] = o(\sqrt{\mathrm{LLog}(N)}/\sqrt{N})$*. In particular, one has that the bias vanishes at the same rate, i.e.* $|\mathbb{E}^{\mathbb{Q}_S}[\hat{f}_N^*] - f^*| = o(\sqrt{\mathrm{LLog}(N)}/\sqrt{N})$*.*

As we have seen, Theorem 2.21 implies that $\hat{f}_N^*$ is an asymptotically unbiased estimator of f^* and that the bias $\mathbb{E}^{\mathbb{Q}_{\mathrm{S}}}[\hat{f}_N^*] - f^*$ is of order $o(\sqrt{\mathrm{LLog}(N)}/\sqrt{N})$. In contrast to classical results for a standard SAA setup, cf. Shapiro et al. [2014], pp. 185, these results on the bias do not need the additional (strong) assumption of uniform integrability of the sequence $\{\sqrt{N}(\hat{f}_N^* - f^*)\}$. Instead one deduces here directly that $\{\sqrt{N}/\sqrt{2\,\mathrm{LLog}(N)}(\hat{f}_N^* - f^*)\}$ is uniformly integrable, as it is convergent in L_1, see, e.g., Bauer [2001], Theorem 21.4.

Remark 2.22. *In the classical SAA approach, where the objective function f is approximated by $\hat{f}_N = \frac{1}{N}\sum_{i=1}^N h(\,\cdot\,, Z_i)$ with a scalar-valued random function h, Theorem 2.21 can be used to further bound the unknown optimal value f^* by an interval of known size and to precise the speed at which $\mathbb{E}^{\mathbb{Q}_S}[\hat{f}_N^*]$ approaches f^*, see Banholzer et al. [2018a] for details.*

Rates of Convergence of Optimal Solutions

Finally, if assumptions (A1)–(A2) and (G1) are met together with (B1)–(B2), then convergence of optimal solutions $\{\hat{x}_N^*\}$ to x^* in any L_p, $1 \le p < \infty$, is easily obtained.

Proposition 2.23. *Suppose that assumptions (A1)–(A2), (B1)–(B2) and (G1) hold. Then, $\{\hat{x}_N\}$ converges to x^* in any L_p, $1 \le p < \infty$, i.e.*

$$\mathbb{E}^{\mathbb{Q}_S}\big[\|\hat{x}_N^* - x^*\|^p\big] \to 0, \quad \textit{as } N \to \infty.$$

In particular, this implies that $\hat{x}_N^*$ is an asymptotically unbiased estimator for x^* and that the mean squared error $\mathbb{E}^{\mathbb{Q}_S}[\|\hat{x}_N^* - x^*\|^2]$ vanishes asymptotically.

Proof. From Proposition 2.17, we know that $\{\hat{x}_N^*\}$ converges to x^*, $\mathbb{Q}_\mathrm{S}$-almost surely, i.e. for each $1 \le p < \infty$, we have $\|\hat{x}_N^* - x^*\|^p \to 0$, $\mathbb{Q}_\mathrm{S}$-almost surely. Further, due to compactness of $\mathcal{X}$, we have $\|\hat{x}_N^* - x^*\|^p \le (\mathrm{diam}(\mathcal{X}))^p$ for the finite diameter $\mathrm{diam}(\mathcal{X}) := \sup\{\|x_1 - x_2\| : x_1, x_2 \in \mathcal{X}\}$. The main statement now follows directly from Lebesgue's dominated convergence theorem (e.g., Serfling [1980], Theorem 1.3.7). The remaining statements are easy consequences. □

Remark 2.24. *The above proposition relies on the initially made assumption that the set $\mathcal{X}$ is compact. In the case of an unbounded $\mathcal{X}$, it is quite easy to construct a counterexample to the above result. More specifically, one can construct a uniformly convex quadratic objective function, where optimal solutions still converge almost surely but not in mean.*

To further derive the corresponding rates for the convergence of $\{\hat{x}_N^*\}$ to x^* in L_p, we additionally require the following lemma. It quantifies the probability that $\hat{x}_N^*$ lies outside the set V_{x^*} of the second-order growth condition, in terms of the rate $\sqrt{\mathrm{LLog}(N)}/\sqrt{N}$.

Lemma 2.25. *Suppose that assumptions (A1)–(A2), (B1)–(B2) and (G1) hold. Then, there exists a $\delta > 0$ (depending on V_{x^*}), such that for all $x \in \mathcal{X}$,*

$$f(x) < f(x^*) + \delta \quad \Rightarrow \quad x \in V_{x^*}.$$

Further, it holds

$$\mathbb{Q}_S\big(\hat{x}_N^* \notin V_{x^*}\big) = o\Big(\frac{\sqrt{\mathrm{LLog}(N)}}{\sqrt{N}}\Big).$$

Proof. We prove the first statement by contradiction, assuming that there exists no such δ. Then we can find a sequence $\{\delta_N\}$ which converges monotonically to 0, together with a sequence $\{x_N\} \in \mathcal{X}\backslash V_{x^*}$ with $f(x_N) < f(x^*) + \delta_N$. As $\mathcal{X}\backslash V_{x^*}$ is compact, the sequence has a least one cluster point $\bar{x} \neq x^*$ with $f(\bar{x}) \leq f(x^*)$. This, however, yields the contradiction to the uniqueness of x^*, as assumed by (B1).

Now, let us consider the following chain of inequalities

$$\begin{aligned}
\mathbb{Q}_{\mathrm{S}}(\hat{x}_N^* \notin V_{x^*}) &\leq \mathbb{Q}_{\mathrm{S}}(f(\hat{x}_N^*) - f(x^*) \geq \delta)\\
&\leq \mathbb{Q}_{\mathrm{S}}(|f(\hat{x}_N^*) - \hat{f}_N(\hat{x}_N^*)| + |\hat{f}_N(\hat{x}_N^*) - f(x^*)| \geq \delta)\\
&\leq \mathbb{Q}_{\mathrm{S}}(\|f - \hat{f}_N\|_\infty + |\hat{f}_N^* - f^*| \geq \delta)\\
&\leq \mathbb{Q}_{\mathrm{S}}(2\|f - \hat{f}_N\|_\infty \geq \delta)\\
&\leq \frac{2\,\mathbb{E}^{\mathbb{Q}_{\mathrm{S}}}\left[\|f - \hat{f}_N\|_\infty\right]}{\delta}
\end{aligned}$$

where we have used Markov's inequality in the last step. Proposition 2.20 now yields the claim. □

Finally, we are now in position to state the following result on rates of convergence in L_p for optimal solutions.

Theorem 2.26. *Suppose that assumptions (A1)–(A2), (B1)–(B2) and (G1) hold. Then, $\{\hat{x}_N^*\}$ converges to x^* in L_1 at a rate of $o((\sqrt{\mathrm{LLog}(N)}/\sqrt{N})^{1/2})$ and in L_p, $2 \leq p < \infty$, at a rate of $o(\sqrt{\mathrm{LLog}(N)}/\sqrt{N})$, i.e.*

$$\mathbb{E}^{\mathbb{Q}_S}\left[\frac{\sqrt{N}\|\hat{x}_N^* - x^*\|}{\sqrt{2\,\mathrm{LLog}(N)}}\right] \to 0, \quad \textit{and} \quad \mathbb{E}^{\mathbb{Q}_S}\left[\frac{\sqrt{N}\|\hat{x}_N^* - x^*\|^p}{\sqrt{2\,\mathrm{LLog}(N)}}\right] \to 0,$$

respectively, as $N \to \infty$.

Moreover, if assumptions (A1)–(A3), (A1')–(A2'), (B1) and (B2'), and (G1)–(G2) are satisfied, then the rate for convergence in L_1 is $o(\sqrt{\mathrm{LLog}(N)}/\sqrt{N})$.

Proof. Under the assumptions (A1)–(A2), (B1)–(B2) and (G1), we only need to prove the statement for $p = 2$. The case $p > 2$ follows from the case $p = 2$ using $\|\hat{x}_N^* - x^*\| \leq \mathrm{diam}(\mathcal{X})$; the case $p = 1$ follows directly from Hölder's inequality. Accordingly, in analogy to the proof of Proposition 2.17, we have

$$\begin{aligned}
\mathbb{E}^{\mathbb{Q}_{\mathrm{S}}}\left[\frac{\sqrt{N}\|\hat{x}_N^* - x^*\|^2}{\sqrt{2\,\mathrm{LLog}(N)}}\right] &= \mathbb{E}^{\mathbb{Q}_{\mathrm{S}}}\left[\frac{\sqrt{N}\|\hat{x}_N^* - x^*\|^2}{\sqrt{2\,\mathrm{LLog}(N)}}\mathbf{1}_{\{\hat{x}_N^* \in V_{x^*}\}}\right] + \mathbb{E}^{\mathbb{Q}_{\mathrm{S}}}\left[\frac{\sqrt{N}\|\hat{x}_N^* - x^*\|^2}{\sqrt{2\,\mathrm{LLog}(N)}}\mathbf{1}_{\{\hat{x}_N^* \notin V_{x^*}\}}\right]\\
&\leq \frac{2}{\alpha}\mathbb{E}^{\mathbb{Q}_{\mathrm{S}}}\left[\frac{\sqrt{N}\|\hat{f}_N - f\|_\infty}{\sqrt{2\,\mathrm{LLog}(N)}}\mathbf{1}_{\{\hat{x}_N^* \in V_{x^*}\}}\right] + \mathbb{E}^{\mathbb{Q}_{\mathrm{S}}}\left[\frac{2\sqrt{N}\|\hat{x}_N^* - x^*\|^2}{\sqrt{2\,\mathrm{LLog}(N)}}\mathbf{1}_{\{\hat{x}_N^* \notin V_{x^*}\}}\right]\\
&\leq \frac{2}{\alpha}\mathbb{E}^{\mathbb{Q}_{\mathrm{S}}}\left[\frac{\sqrt{N}\|\hat{f}_N - f\|_\infty}{\sqrt{2\,\mathrm{LLog}(N)}}\right] + \mathbb{E}^{\mathbb{Q}_{\mathrm{S}}}\left[\frac{\sqrt{N}\|\hat{x}_N^* - x^*\|^2}{\sqrt{2\,\mathrm{LLog}(N)}}\mathbf{1}_{\{\hat{x}_N^* \notin V_{x^*}\}}\right].
\end{aligned}$$

The first term of the latter expression already shows the proposed rate according to Proposition 2.20. For the second term, we use

$$\mathbb{E}^{\mathbb{Q}_\mathrm{S}}\left[\frac{\sqrt{N}\|\hat{x}_N^*-x^*\|^2}{\sqrt{2\,\mathrm{LLog}(N)}}\mathbf{1}_{\{\hat{x}_N^*\notin V_{x^*}\}}\right] \leq \mathrm{diam}(\mathcal{X})^2\frac{\sqrt{N}}{\sqrt{2\,\mathrm{LLog}(N)}}\,\mathbb{Q}_\mathrm{S}(\hat{x}_N^*\notin V_{x^*})$$

which, together with Lemma 2.25 above, shows the claim for convergence in L_2.

Eventually, assuming (A1)–(A3), (A1')–(A2'), (B1) and (B2'), and (G1)–(G2), the stronger rate of $o(\sqrt{\mathrm{LLog}(N)}/\sqrt{N})$ can be obtained analogously for convergence in L_1, cf. Theorem 2.19. □

2.2.3.2 Convergence in Probability

In this section, we first consider rates for the size of the deviation corridor (i.e. inside the probability), before addressing rates of a fixed deviation probability (i.e. outside the probability). Both results provide further insights into the asymptotic behaviour of optimal estimators and the related error probabilities, by referring to results from the literature on the LIL in Banach spaces.

Rates of Convergence in Probability

From the well-known fact that almost sure convergence implies convergence in probability, all convergence rates obtained in Section 2.2.2.2 also hold in probability. However, slightly better convergence results can be obtained by making use of the rates of convergence in mean, yielding the following result.

Proposition 2.27. *Suppose that assumptions (A1)–(A2) and (G1) hold, and let $\delta > 0$ be arbitrary. Then,*

$$\mathbb{Q}_S\left(\frac{\sqrt{N}|\hat{f}_N^*-f^*|}{\sqrt{2\,\mathrm{LLog}(N)}} > \delta\right) \to 0, \quad \textit{as } N\to\infty. \tag{2.30}$$

Further, if in addition assumptions (B1)–(B2) are satisfied, then we have

$$\mathbb{Q}_S\left(\frac{\sqrt{N}\|\hat{x}_N^*-x^*\|^2}{\sqrt{2\,\mathrm{LLog}(N)}} > \delta\right) \to 0, \quad \textit{as } N\to\infty. \tag{2.31}$$

Finally, if assumptions (A1)–(A3), (A1')–(A2'), (B1) and (B2'), and (G1)–(G2) are satisfied, then it holds

$$\mathbb{Q}_S\left(\frac{\sqrt{N}\|\hat{x}_N^*-x^*\|}{\sqrt{2\,\mathrm{LLog}(N)}} > \delta\right) \to 0, \quad \textit{as } N\to\infty. \tag{2.32}$$

Proof. The proof follows in a straightforward manner from Proposition 2.20, and Theorems 2.21 and 2.26. □

Proposition 2.27 may be used to derive confidence sequences for the true optimal value f^* and an optimal solution x^* under quite weak assumptons, see Banholzer et al. [2018a] for further details.

Rates of Error Probabilities

To derive rates for the probability of lying outside a small confidence ball from convergence in mean, we reconsider inequality (2.29) under assumptions (A1) and (A2), which implies that for any $\epsilon > 0$ there exists a deterministic $N^* = N^*(\epsilon) \in \mathbb{N}$ such that

$$\forall N \geq N^* : \quad \frac{\sqrt{N}}{\sqrt{2\,\mathrm{LLog}(N)}}\mathbb{E}^{\mathbb{Q}_\mathrm{S}}\Big[\|\widehat{\Pi}_N - \Pi\|_\infty\Big] \leq \epsilon. \tag{2.33}$$

In consequence of this inequality, we are able to formulate under the additional assumption of (G1) the following probabilistic estimates for the differences in objective function values, where we distinguish between the case when no further moment conditions on the random variable $\widetilde{X} = h(\,\cdot\,, Z) - \mathbb{E}^{\mathbb{Q}}[h(\,\cdot\,, Z)]$ are available (to apply Markov's inequality) and the case when higher moment conditions on $\widetilde{X}$ are satisfied (to use an inequality by Einmahl and Li [2008]).

Theorem 2.28. *Suppose that assumptions (A1)–(A2) and (G1) hold, and let $\delta > 0$. Then, the following statements hold:*

(a) For any $\epsilon > 0$, there exists an $N^* = N^*(\epsilon) \in \mathbb{N}$ such that

$$\forall N \geq N^* : \quad \mathbb{Q}_S\Big(\|\hat{f}_N - f\|_\infty \geq \delta\Big) \leq L_g \frac{\epsilon}{\delta}\frac{\sqrt{2\,\mathrm{LLog}(N)}}{\sqrt{N}}. \tag{2.34}$$

(b) If $\mathbb{E}^{\mathbb{Q}}[\|\widetilde{X}\|_\infty^p] < \infty$ for $p > 2$, then there exists an $N^* = N^*(\delta) \in \mathbb{N}$ such that for all $N \geq N^*$:

$$\mathbb{Q}_S\Big(\|\hat{f}_N - f\|_\infty \geq \delta\Big) \leq \exp\bigg\{-\frac{N\delta^2}{12L_g^2\,\sigma^2(\widetilde{X})}\bigg\} + \frac{\bar{c}}{N^{p-1}(\delta/(2L_g))^p}\mathbb{E}^{\mathbb{Q}}\big[\|\widetilde{X}\|_\infty^p\big], \tag{2.35}$$

where $\sigma(\widetilde{X})$ is given by (2.24) for the $C(\mathcal{X}, \mathbb{R}^l)$-valued random variable $\widetilde{X}$ and $\bar{c}$ is a positive constant.

Proof. Assertion (a) follows directly from Markov's inequality, the Lipschitz continuity of g and inequality (2.33).

To show assertion (b), we first observe that the Lipschitz continuity of g gives

$$\mathbb{Q}_\mathrm{S}\Big(\|\hat{f}_N - f\|_\infty \geq \delta\Big) \leq \mathbb{Q}_\mathrm{S}\Big(\|\widehat{\Pi}_N - \Pi\|_\infty \geq \delta/L_g\Big).$$

Due to inequality (2.33), it then follows that for $\delta > 0$ and some arbitrary but fixed $0 < \eta \leq 1$, there exists an $N^* = N^*(\delta, \eta) \in \mathbb{N}$ such that for all $N \geq N^*$,

$$\begin{aligned}\mathbb{Q}_\mathrm{S}\Big(\|\widehat{\Pi}_N - \Pi\|_\infty \geq \delta/L_g\Big) &\leq \mathbb{Q}_\mathrm{S}\Big(\|\widehat{\Pi}_N - \Pi\|_\infty \geq (1+\eta)\mathbb{E}^{\mathbb{Q}_\mathrm{S}}\big[\|\widehat{\Pi}_N - \Pi\|_\infty\big] + \delta/(2L_g)\Big)\\ &\leq \mathbb{Q}_\mathrm{S}\Big(\max_{1\leq j\leq N}\|\widehat{\Pi}_j - \Pi\|_\infty \geq (1+\eta)\mathbb{E}^{\mathbb{Q}_\mathrm{S}}\big[\|\widehat{\Pi}_N - \Pi\|_\infty\big] + \delta/(2L_g)\Big),\end{aligned}$$

where the second inequality follows from $\max_{1\leq j\leq N}\|\widehat{\Pi}_j - \Pi\|_\infty \geq \|\widehat{\Pi}_N - \Pi\|_\infty$. By applying Theorem 4 of Einmahl and Li [2008] (with $\delta = 1$ and $t = \delta/2$) on the $C(\mathcal{X}, \mathbb{R}^l)$-valued random variables $\widetilde{X}_i/N$ under the moment condition $\mathbb{E}^{\mathbb{Q}}[\|\widetilde{X}\|_\infty^p] < \infty$, we then obtain $\forall N \geq N^*$,

$$\begin{aligned}\mathbb{Q}_{\mathrm{S}}\Big(\max_{1\leq j\leq N}\|\widehat{\Pi}_j - \Pi\|_\infty \geq (1+\eta)\mathbb{E}^{\mathbb{Q}_{\mathrm{S}}}\big[\|\widehat{\Pi}_N - \Pi\|_\infty\big] + \delta/(2L_g)\Big)&\\ \leq \exp\left\{-\frac{N\delta^2}{12L_g^2\,\sigma^2(\widetilde{X})}\right\} + \frac{\bar{c}}{N^{p-1}(\delta/(2L_g))^p}\mathbb{E}^{\mathbb{Q}}\big[\|\widetilde{X}\|_\infty^p\big]&,\end{aligned}$$

with the specified constants $\sigma(\widetilde{X})$ and $\bar{c}$. □

Both statements in Theorem 2.28 imply that for small values δ, $N \gg 1/\delta^2$ is mandatory to obtain a reasonably high probability for a given small accuracy δ for the sample average approximation. More interestingly, while the error probability of $\widehat{\Pi}_N$ with respect to Π in (2.34) has essentially the usual rate $\sqrt{\mathrm{LLog}(N)}/\sqrt{N}$, the rate is of order $1/N^{p-1}$ in (2.35). However, it has to be noted that in both approaches the exact number N^* needed for the validity of both estimates is not known.

Remark 2.29. *Given Theorem 2.16, both estimates* (2.34) *and* (2.35) *in Theorem 2.28 can further be used to infer rates in probability for the absolute error of the optimal values, i.e.* $\mathbb{Q}_S(|\hat{f}_N^* - f^*| > \delta$*. Moreover, Markov's inequality can be applied to obtain similar rates for the error probability of the optimal solutions* $\mathbb{Q}_S(\|\hat{x}_N^* - x^*\| > \delta)$*, based on Theorem 2.26.*

2.3 Variable Sample Average Approximation Strategy

We now discuss almost sure convergence properties of the second approach to approximately solve the original problem (2.1) by means of Monte Carlo methods, the VSAA strategy. Specifically, we assume that for a given sequence of sample sizes $\{N_k\}_{k\in\mathbb{N}}$, $N_k \in \mathbb{N}$, as set by the user (and henceforth also termed schedule), a new sample of i.i.d. random vectors $Z_1^k, \ldots, Z_{N_k}^k$ is drawn at the k-th function evaluation from the distribution of Z and independently of previous samples. This then amounts to estimating the original objective function f at that stage by

$$\hat{f}_{N_k}(x) = g\Big(\widehat{\Pi}_{N_k}(x) - C^{\mathrm{mkt}}\Big), \qquad x \in \mathcal{X}, \tag{2.36}$$

with usual transformation $g : \mathbb{R}^l \to \mathbb{R}_{\geq 0}$, Monte Carlo estimator $\widehat{\Pi}_{N_k}(x) = \frac{1}{N_k}\sum_{i=1}^{N_k} h(x, Z_i^k)$, and market prices C^{mkt}.

Due to the fact that the approximating objective function (2.36) changes with each function evaluation, the appealing convergence results as established within the SAA approach are no longer applicable. Above all, as the VSAA strategy is integrated into a sequential sampling algorithm, a complete proof of convergence essentially depends on the method under consideration and thus on the underlying mechanism to select new sample points. Nevertheless, some general results on the sequence of approximating objective functions $\{\hat{f}_{N_k}\}$

are provided by Homem-de-Mello [2003] and Royset [2013] under conditions on the random function h and the imposed schedule in a standard VSAA framework, which may then facilitate establishing convergence of a variable-sample method. In particular, these results are derived pathwisely, which allows applicability in fairly general contexts without any specific assumption on the underlying distribution; e.g., to show that a given sequential sampling algorithm converges to the original set of optimal solutions almost surely and thus preserves the convergence properties of a deterministic equivalent.

Homem-de-Mello [2003] states various conditions in different setups on the random function h and the schedule $\{N_k\}$ such that the pointwise strong consistency of the sequence of estimators $\{\hat{f}_{N_k}(x)\}$ for $f(x)$ is ensured at any $x \in \mathcal{X}$, as $k \to \infty$. More precisely, the conditions enable to apply either the theory of large deviations or Chebyshev's inequality for bounding the probability of deviation at a single iteration k, from which then the desired almost sure convergence statement can be obtained by the Borel-Cantelli lemma. In a similar fashion, but under stronger assumptions on h and $\{N_k\}$ allowing to use large deviations together with a variant of the CLT and Borel-Cantelli, Homem-de-Mello also obtains almost sure pointwise bounds of the form $\sqrt{\operatorname{Log}(N_k)}/\sqrt{N_k}$ for the estimation error $|\hat{f}_{N_k}(x) - f(x)|$ at any $x \in \mathcal{X}$. Eventually, it is exemplarily shown by means of the pointwise strong consistency that modified random-search type methods converge almost surely to the set of true optimal solutions on a finite (i.e. discrete) feasible set $\mathcal{X}$.

Royset [2013] further considers a certain class of algorithms for the VSAA approach and develops a technique to adaptively select efficient sample sizes in smooth stochastic programming problems. The technique aims at minimising the expected computing time to reach an ϵ-optimal solution, which is formulated as a discrete-time optimal control problem and approximately solved by dynamic programming. Within this analysis, Royset shows that, under a uniformly linear convergence assumption, the class of algorithms is able to produce an ϵ-optimal solution in finite time, almost surely. The proof of the latter relies on the almost sure uniform convergence of $\{\hat{f}_{N_k}\}$ to f on $\mathcal{X}$, as $k \to \infty$, which is derived under strong exponential moment conditions on h but relatively weak assumptions on $\{N_k\}$ via the theory of large deviations and the Borel-Cantelli lemma.

Considering these contributions, we point out that fairly reasonable assumptions on the random function h and the schedule of sample sizes $\{N_k\}$ are only given for pointwise results, whereas the strong uniform consistency relies on rather strong assumptions on h and no findings on uniform sample path bounds seem to be known at all. In many situations, though, this compilation of available means may not be sufficient (or the involved assumptions too restrictive) to show convergence of a variable-sample method, in particular when the feasible set $\mathcal{X}$ is infinite and uniformly derived results are required. This is also the case when the VSAA strategy is adopted by the noisy RBF method, as considered in Section 5.3.

Yet, given our previous studies on almost sure rates of convergence in the SAA framework, it stands to reason to also apply results on functional limit theorems in order to obtain relevant convergence statements in the VSAA setup. In particular, as will be shown in the subsequent part of this chapter within the calibration setup, we are able to considerably improve on and complement the existing VSAA literature in this way. Specifically, by use of the SLLN for general arrays of Banach space valued random variables, we will establish the strong uniform consistency of the estimators $\{\hat{f}_{N_k}\}$ to f under assumptions that do not rely on exponential moments and are thus substantially weaker than those made by Royset [2013]. Moreover, by applying the LIL for Banach space valued arrays in a similar manner, we are even able to derive – for the first time – sample path bounds and almost sure rates of the form $\sqrt{\operatorname{Log}(N_k)}/\sqrt{N_k}$ for the uniform convergence of the objective functions $\{\hat{f}_{N_k}\}$ to f on $\mathcal{X}$. Eventually, as an immediate consequence of our derivations, it is also possible to provide different (partially) weaker assumptions than those made by Homem-de-Mello [2003] in order to ensure the pointwise strong consistency of the estimators $\{\hat{f}_{N_k}(x)\}$ for $f(x)$ and to guarantee pointwise sample path bounds on $|\hat{f}_{N_k}(x) - f(x)|$ at any $x \in \mathcal{X}$.

In accordance with the presentation of the SAA approach, we begin in Section 2.3.1 by embedding the above VSAA setup into a suitable probability space, address the involved measurability issues and give conditions on the random function h and the schedule $\{N_k\}$ such that the estimators $\{\hat{f}_{N_k}(x)\}$ are strongly uniformly (and pointwise) consistent for $f(x)$ on $\mathcal{X}$. In Section 2.3.2, we then derive uniform (and pointwise) pathwise bounds on the deviation of $\hat{f}_{N_k}(x)$ from $f(x)$, which may be further used for establishing convergence of the noisy RBF method in Section 5.3 if the VSAA strategy is adopted.

2.3.1 Probabilistic Setup and Strong Consistency

To suitably accommodate the VSAA strategy in a probabilistic framework, let us consider for a given deterministic schedule of Monte Carlo sample sizes $\{N_k\}$ the sequence of i.i.d. $\mathcal{Z}$-valued random vectors $\{\{Z_i^k\}_{i=1}^{N_k}\}_{k\in\mathbb{N}}$, each Z_i^k following the distribution of Z, as defined on a common probability space $(\Omega_{\mathrm{V}}, \mathcal{F}_{\mathrm{V}}, \mathbb{Q}_{\mathrm{V}})$. In particular, the latter is specified by the sample space $\Omega_{\mathrm{V}} := \bigtimes_{k=1}^{\infty} \Omega^{N_k}$, where Ω^{N_k} denotes the N_k-fold Cartesian product of the space Ω, the respective σ-algebra $\mathcal{F}_{\mathrm{V}} := \bigotimes_{k=1}^{\infty} \mathcal{F}^{\otimes N_k}$, which is the smallest σ-algebra on Ω_{V} such that the coordinate maps $Z_i^k : \Omega_{\mathrm{V}} \to \mathcal{Z}$ are $\mathcal{F}_{\mathrm{V}}/\mathcal{B}(\mathcal{Z})$-measurable, and the unique product probability measure $\mathbb{Q}_{\mathrm{V}} := \bigotimes_{k=1}^{\infty} \mathbb{Q}^{\otimes N_k}$. Note that this construction is feasible as the random vectors are assumed to be independent between different samples.

Even though the underlying random vectors $\{Z_i^k\}$ are assumed to be i.i.d. in our considered case, the above probability space also allows to accommodate, by some abuse of notation, a setup in which each set $Z_1^k, \ldots, Z_{N_k}^k$ is i.i.d. and independent from previous sets, but may be drawn from a different distribution than other sample sets, cf. Homem-de-Mello [2003]. This would then enable the use of some variance reduction technique, see, e.g., Glasserman [2003], Chapter 4, along the variable-sample procedure. However, one should bear in mind that in such a modified setup one would require available results covering general arrays of rowwise i.i.d. random variables (that is i.i.d. within each row). While this is provided by Hu et al. [1999], Corollary 4.1, for the case of complete convergence, which then allows to establish the strong uniform consistency of $\{\hat{f}_{N_k}\}$ to f, no such result is currently known on the compact LIL for deriving uniform sample path bounds.

Measurability of Estimators

To take care of potential measurability issues arising within the VSAA approach, we initially recall that for any $x \in \mathcal{X}$ the vector $h(x, \cdot)$ is assumed to be $\mathcal{B}(\mathcal{Z})/\mathcal{B}(\mathbb{R}^l)$-measurable. We thus have by construction of the underlying probability space $(\Omega_\mathrm{V}, \mathcal{F}_\mathrm{V}, \mathbb{Q}_\mathrm{V})$ that for each $k \in \mathbb{N}$, the estimators $\widehat{\Pi}_{N_k}(x) = \widehat{\Pi}_{N_k}(x, \omega_\mathrm{V})$ and $\hat{f}_{N_k}(x) = g(\widehat{\Pi}_{N_k}(x, \omega_\mathrm{V}) - C^{\mathrm{mkt}})$ with $\omega_\mathrm{V} \in \Omega_\mathrm{V}$ are measurable with respect to $\mathcal{F}_\mathrm{V}/\mathcal{B}(\mathbb{R}^l)$ and $\mathcal{F}_\mathrm{V}/\mathcal{B}(\mathbb{R})$, respectively. Also, their finite-valuedness, $\mathbb{Q}_\mathrm{V}$-almost surely, follows from the fact that $h(x, \cdot)$ is assumed to be $\mathbb{Q}^Z$-integrable for each $x \in \mathcal{X}$.

Further, to ensure the measurability of the function values $\widehat{\Pi}_{N_k}(x_k) = \widehat{\Pi}_{N_k}(x_k(\omega_\mathrm{V}), \omega_\mathrm{V})$ and $\hat{f}_{N_k}(x_k) = \hat{f}_{N_k}(x_k(\omega_\mathrm{V}), \omega_\mathrm{V})$ at any point $x_k(\omega_\mathrm{V})$ as selected by a potential algorithm employing the VSAA strategy, we also assume that $h(\cdot, Z)$ is continuous on $\mathcal{X}$ for almost every $Z \in \mathcal{Z}$. Hence, the functions $\widehat{\Pi}_{N_k}$ and $\hat{f}_{N_k}$ from $\mathcal{X} \times \Omega_\mathrm{V}$ into $\mathbb{R}^l$ and $\mathbb{R}$, respectively, are Carathéodory functions, cf. Section 2.2.1. Thus, we can assert the following, see, e.g., Geletu [2006], Proposition 6.3.7, for an explicit reference.

Proposition 2.30. *Let $\mathcal{X} \subset \mathbb{R}^d$ be a nonempty compact set, and suppose that for each $k \in \mathbb{N}$, the pricing function $\widehat{\Pi}_{N_k} : \mathcal{X} \times \Omega_V \to \mathbb{R}^l$ and the objective function $\hat{f}_{N_k} : \mathcal{X} \times \Omega_V \to \mathbb{R}$ are Carathéodory functions and $x_k : \Omega_V \to \mathcal{X}$ is measurable with respect to $\mathcal{F}_V/\mathcal{B}(\mathbb{R}^d)$. Then, the function values $\widehat{\Pi}_{N_k}(x_k(\omega_V), \omega_V)$ and $\hat{f}_{N_k}(x_k(\omega_V), \omega_V)$ are measurable with respect to $\mathcal{F}_V/\mathcal{B}(\mathbb{R}^l)$ and $\mathcal{F}_V/\mathcal{B}(\mathbb{R})$, respectively.*

Note that, if required, the measurability of optimal values and solutions over all computed function values, i.e. of $\min_{1\le i\le k} \hat{f}_{N_i}(x_i(\omega_\mathrm{V}), \omega_\mathrm{V})$ and $\operatorname{argmin}_{1\le i\le k} \hat{f}_{N_i}(x_i(\omega_\mathrm{V}), \omega_\mathrm{V})$ for each $k \in \mathbb{N}$, may be easily deduced from Proposition 2.30, using standard measurability arguments.

Strong Consistency of Estimators

Eventually, we address the strong consistency of the sequence of estimators $\{\hat{f}_{N_k}\}$ and give conditions on the random function h and the schedule of sample sizes $\{N_k\}$ such that $\{\hat{f}_{N_k}\}$ converges to f uniformly (and also pointwise) on $\mathcal{X}$, $\mathbb{Q}_\mathrm{V}$-almost surely, as the number of function evaluations k tends to infinity. To establish this result, however, an application of the ordinary SLLN for a sequence of random vectors is not sufficient, as for each evaluation k a new subsample of random vectors $Z_1^k, \dots, Z_{N_k}^k$ is taken from the entire stream $\{Z_i^k\}$ but convergence is sought $\mathbb{Q}_\mathrm{V}$-almost surely. Nonetheless, it may be derived by using a version of the SLLN for general arrays of random vectors, which relies on the concept of complete convergence.

To begin with, let us make the following assumptions on the random function h. Note that these assumptions are implied, in turn, by assuming (A1) and (A2); similarly, the below assumption (C1) follows from the supposed existence of a measurable and square-integrable function $G : \mathcal{Z} \to \mathbb{R}_{\geq 0}$ such that $\|h(x,Z)\| \leq G(Z)$ for all $x \in \mathcal{X}$, $\mathbb{Q}$-almost surely, cf. Theorem 2.8(*ii*).

- **(C1)** For any $x \in \mathcal{X}$ we have $\mathbb{E}^{\mathbb{Q}}[\|h(x,Z)\|^2] < \infty$.
- **(C2)** The function $h(\,\cdot\,,Z)$ is continuous on $\mathcal{X}$, $\mathbb{Q}$-almost surely.

Both assumptions (C1) and (C2) ensure that the general array $\{\widetilde{X}_{ki}\}$ of i.i.d. random variables with $\widetilde{X}_{ki} = h(\,\cdot\,,Z_i^k) - \mathbb{E}^{\mathbb{Q}}[h(\,\cdot\,,Z_i^k)]$ takes values in the space $C(\mathcal{X},\mathbb{R}^l)$ and has a finite strong second moment $\mathbb{E}^{\mathbb{Q}}[\|\widetilde{X}\|_\infty^2] < \infty$, where $\widetilde{X} = h(\,\cdot\,,Z) - \mathbb{E}^{\mathbb{Q}}[h(\,\cdot\,,Z)]$. Since $\widetilde{X}$ has mean zero, it thus follows from Theorem 2.2 that for a strictly monotonically increasing schedule $\{N_k\}$, the sequence $\{\sum_{i=1}^{N_k} \widetilde{X}_{ki}/N_k\}$ converges completely to zero as $k \to \infty$, i.e. it holds

$$\sum_{k=1}^{\infty} \mathbb{Q}_{\mathrm{V}}\Big(\big\|\widehat{\Pi}_{N_k} - \Pi\big\|_\infty > \epsilon\Big) < \infty, \qquad \text{for all } \epsilon > 0.$$

Consequently, by the Borel-Cantelli lemma, we may state the following assertion on the strong uniform consistency of the sequence of estimators $\{\widehat{\Pi}_{N_k}\}$.

Theorem 2.31. *Suppose that assumptions (C1)–(C2) hold, and that the schedule of sample sizes $\{N_k\}$ is strictly monotonically increasing. Then,*

$$\big\|\widehat{\Pi}_{N_k} - \Pi\big\|_\infty \to 0, \qquad \textit{as } k \to \infty,$$

$\mathbb{Q}_V$*-almost surely.*

As a direct consequence of Theorem 2.31, the almost sure uniform convergence of the related sequence of objective functions can be concluded, given that the transforming function g is Lipschitz continuous, i.e. assumption (G1) is made (which also applies to $\widehat{\Pi}_{N_k}$).

Proposition 2.32. *Suppose that assumptions (C1)–(C2) and (G1) hold, and that the schedule of sample sizes $\{N_k\}$ is strictly monotonically increasing. Then,*

$$\big\|\hat{f}_{N_k} - f\big\|_\infty \to 0, \qquad \textit{as } k \to \infty,$$

$\mathbb{Q}_V$*-almost surely.*

It is worth pointing out that, besides the almost sure continuity of $h(\,\cdot\,,Z)$ on $\mathcal{X}$ and a strictly monotonically increasing choice of $\{N_k\}$, Proposition 2.32 only requires a finite strong second moment of the random function $h(\,\cdot\,,Z)$. The proposition is thus able to dispense with the strong assumptions of a finite-valued moment generating function for $h(x,Z) - \mathbb{E}^{\mathbb{Q}}[h(x,Z)]$ at any $x \in \mathcal{X}$ and for the Lipschitz constant of $h(\,\cdot\,,Z)$ in a neighbourhood of zero. These are used by Royset [2013], Theorem 1, to derive the uniform convergence of $\{\hat{f}_{N_k}\}$ to f on $\mathcal{X}$, $\mathbb{Q}_{\mathrm{V}}$-almost surely, within a standard VSAA setup where $\hat{f}_{N_k}$ is directly formed as a sample average of scalar-valued random functions h. In addition to both assumptions (and the Lipschitz continuity of $h(\,\cdot\,, Z)$ on $\mathcal{X}$, $\mathbb{Q}$-almost surely), the schedule $\{N_k\}$ is required to satisfy $\sum_{k=1}^{\infty} \tilde{\alpha}^{N_k} < \infty$ for all $\tilde{\alpha} \in (0,1)$, which obviously holds for any strictly increasing sequence $\{N_k\}$ of natural numbers and for any $\{N_k\}$ following a linear or sublinear growth where $1/N_k$ is of order $\mathcal{O}(1/k)$ or $\mathcal{O}(1/\sqrt[\tilde{c}]{k})$ with $\tilde{c} \in \mathbb{N}$ (e.g., by the integral test), respectively. However, for a sequence according to $1/N_k = \mathcal{O}(1/\log k)$, the condition is not met, see also the discussion in Homem-de-Mello [2003], p. 119.

Note that the strict monotonicity of the schedule $\{N_k\}$ in Theorem 2.31 and Proposition 2.32 could be further relaxed by applying Hu et al. [1999], Corollary 4.1, instead of Theorem 2.2, to obtain the complete convergence of $\{\sum_{i=1}^{N_k} \widetilde{X}_{ki}/N_k\}$. Under (possibly) adjusted moment conditions on $\widetilde{X}$, the schedule $\{N_k\}$ is then required to satisfy $\sum_{k=1}^{\infty} N_k^{-\tilde{\alpha}} < \infty$ for some $\tilde{\alpha} > 0$, which would yield about the same orders of growth as Royset's constraint.

In line with above derivation, it is straightforward to also deduce the pointwise strong consistency of the sequence of estimators $\{\hat{f}_{N_k}(x)\}$ for $f(x)$ at any $x \in \mathcal{X}$, yet under weaker assumptions than given in Proposition 2.32. Indeed, by assuming (C1), it follows that the array of i.i.d. $\mathbb{R}^l$-valued random vectors $\{\widetilde{X}_{ki}\}$ with $\widetilde{X}_{ki} = h(x, Z_i^k) - \mathbb{E}^{\mathbb{Q}}[h(x, Z_i^k)]$ has a finite strong second moment, such that Theorem 2.2 equally applies in the finite-dimensional space $\mathbb{R}^l$ under the stated condition on the schedule $\{N_k\}$. From the pointwise strong consistency of $\{\widehat{\Pi}_{N_k}(x)\}$ for $\Pi(x)$ and the continuous mapping theorem (e.g., Serfling [1980], Theorem 1.7) with the continuous function $g : \mathbb{R}^l \to \mathbb{R}_{\geq 0}$, we may then derive the following result.

Corollary 2.33. *Suppose that assumption (C1) holds, and that the schedule of sample sizes $\{N_k\}$ is strictly monotonically increasing. Then, for any $x \in \mathcal{X}$,*

$$\hat{f}_{N_k}(x) \to f(x), \qquad \text{as } k \to \infty,$$

$\mathbb{Q}_{\mathrm{V}}$-almost surely.

In comparison to Corollary 2.33 stands the pointwise consistency result as derived by Homem-de-Mello [2003], Proposition 3.2, for the particular case of i.i.d. random vectors $\{Z_i^k\}$ (not necessarily following the distribution of Z) and an unbiased Monte Carlo estimator $\hat{f}_{N_k}(x) = \frac{1}{N_k} \sum_{i=1}^{N_k} h(x, Z_i^k)$. In particular, his result relies on the stronger assumption of a bounded second moment for the scalar-valued $h(x, Z_i^k)$ at any $x \in \mathcal{X}$, while the schedule of sample sizes $\{N_k\}$ is supposed to meet $\sum_{k=1}^{\infty} \tilde{\alpha}^{N_k} < \infty$ for all $\tilde{\alpha} \in (0,1)$.

2.3.2 Sample Path Bounds

Given the strong consistency of the estimators $\{\hat{f}_{N_k}\}$ for f, we now provide for each sample path $\omega_{\mathrm{V}} \in \Omega_{\mathrm{V}}$, i.e. for each particular realisation of the sequence of underlying random vectors $\{Z_i^k\}$, uniform (as well as pointwise) error bounds on the deviation of $\hat{f}_{N_k}$ from its true counterpart f. By analogy with the SAA approach, these bounds can be derived by means of the LIL in a suitable Banach space setting, which then also allows us to quantify the implied rate of convergence that holds almost surely. However, since the VSAA basically requires a version of the LIL for general arrays of i.i.d. random vectors, stronger assumptions on the random function h have to be imposed in turn, complemented by conditions on the schedule $\{N_k\}$.

To enter into details, we reconsider the general array $\{\widetilde{X}_{ki}\}$ of i.i.d. $C(\mathcal{X}, \mathbb{R}^l)$-valued random vectors with entries $\widetilde{X}_{ki} = h(\,\cdot\,, Z_i^k) - \mathbb{E}^{\mathbb{Q}}[h(\,\cdot\,, Z_i^k)]$, i.e. from the same distribution as $\widetilde{X}$, and require the following assumption on the involved random function h:

(D1) For any $x \in \mathcal{X}$ we have

$$\mathbb{E}^{\mathbb{Q}}\big[\|h(x,Z) - \mathbb{E}^{\mathbb{Q}}[h(x,Z)]\|^4 / \big(\operatorname{Log}(\|h(x,Z) - \mathbb{E}^{\mathbb{Q}}[h(x,Z)]\|)\big)^2\big] < \infty.$$

By assumption (D1), the moment condition (i) of Theorem 2.3 is clearly satisfied for the mean zero $C(\mathcal{X}, \mathbb{R}^l)$-valued random variable $\widetilde{X}$. Moreover, since the assumption also implies $\mathbb{E}^{\mathbb{Q}}[\|h(x,Z)\|^2] < \infty$ for any $x \in \mathcal{X}$, we have together with assumption (A2) on the Lipschitz continuity of $h(\,\cdot\,, Z)$ that the compact LIL holds for the $C(\mathcal{X}, \mathbb{R}^l)$-valued sequence $\{\widetilde{X}_N\}$, $\widetilde{X}_N = h(\,\cdot\,, Z_N) - \mathbb{E}^{\mathbb{Q}}[h(\,\cdot\,, Z_N)]$, where the Z_i are i.i.d. random variables with the same distribution as Z, see Kuelbs [1976a], Theorem 4.4. In particular, according to Theorem 2.1(b), assertion (iii), it thus follows that for any sequence $\{N_k\}$ with $\lim_{k\to\infty} N_k = \infty$, the term $\sum_{i=1}^{N_k} \widetilde{X}_i / \sqrt{2N_k \operatorname{LLog}(N_k)}$ converges in probability to zero, as $k \to \infty$, which again provides $\sum_{i=1}^{N_k} \widetilde{X}_i / \sqrt{2N_k \operatorname{Log}(N_k)} \xrightarrow{p} 0$ as $k \to \infty$. Hence, we have by Theorem 2.3 for any strictly monotonically increasing sequence $\{N_k\}$,

$$\limsup_{k\to\infty} \frac{\sqrt{N_k}\big\|\widehat{\Pi}_{N_k} - \Pi\big\|_\infty}{\sqrt{2\operatorname{Log}(N_k)}} \le \sigma(\widetilde{X}),$$

where $\sigma(\widetilde{X})$ is given by (2.24) for the $C(\mathcal{X}, \mathbb{R}^l)$-valued random variable $\widetilde{X}$. Using the definition of the limit superior, we thus obtain that for any $\epsilon > 0$, there exists a finite random variable $k^* = k^*(\epsilon) \in \mathbb{N}$ such that

$$\forall k \ge k^* : \quad \big\|\widehat{\Pi}_{N_k} - \Pi\big\|_\infty \le \frac{\sqrt{(2+\epsilon)\operatorname{Log}(N_k)}}{\sqrt{N_k}}\sigma(\widetilde{X}), \tag{2.37}$$

$\mathbb{Q}_V$-almost surely, which bounds the difference in pricing functions uniformly. Eventually, from inequality (2.37), it immediately follows the uniform boundedness of the difference in objective functions, given the Lipschitz continuity of g with constant L_g. For the equivalent assertion in the SAA approach, see Lemma 2.15.

Proposition 2.34. *Suppose that assumptions (D1), (A2) and (G1) hold. Further, suppose that the schedule of sample sizes $\{N_k\}$ is strictly monotonically increasing. Then, for any $\epsilon > 0$, there exists a finite random variable $k^* = k^*(\epsilon) \in \mathbb{N}$ such that*

$$\forall k \ge k^* : \quad \big\|\hat{f}_{N_k} - f\big\|_\infty \le L_g \frac{\sqrt{(2+\epsilon)\operatorname{Log}(N_k)}}{\sqrt{N_k}}\sigma(\widetilde{X}), \tag{2.38}$$

$\mathbb{Q}_V$-almost surely, where $\sigma(\widetilde{X})$ is given by (2.24) *for the $C(\mathcal{X}, \mathbb{R}^l)$-valued random variable $\widetilde{X}$.*

Note that the rate at which the convergences in (2.37) and (2.38) occur is of $\mathcal{O}(\sqrt{\mathrm{Log}(N_k)}/\sqrt{N_k})$, which is marginally weaker than the rate $\mathcal{O}(\sqrt{\mathrm{LLog}(N_k)}/\sqrt{N_k})$ obtained in the SAA approach, cf. Section 2.2.2.2. This difference can be attributed to the fact that an array of random vectors has to be used instead of an ordinary sequence. In particular, Proposition 2.34 may be used to show almost sure convergence of a variable-sample method, where new evaluation points are selected over an infinite feasible set $\mathcal{X}$ and a uniform quantification of the error in objective function values is required.

Again, by considering $\{\widetilde{X}_{ki}\}$ as an array of $\mathbb{R}^l$-valued random variables instead of random variables taking values in $C(\mathcal{X}, \mathbb{R}^l)$, it is an easy task to also derive pointwise sample path bounds on the error of objective functions at any $x \in \mathcal{X}$. Specifically, given that assumption (D1) holds and that the schedule $\{N_k\}$ is strictly monotonically increasing, an application of Theorem 2.3 yields that the upper bound of (2.37) also pertains to $\|\widehat{\Pi}_{N_k}(x) - \Pi(x)\|$, $\mathbb{Q}_{\mathrm{V}}$-almost surely, if the respective quantity $\sigma(\widetilde{X})$ takes the form

$$\sigma(\widetilde{X}) = \left\|\Sigma^{1/2}(x)\right\|, \tag{2.39}$$

where $\Sigma(x)$ denotes the covariance matrix of the associated random variable $\widetilde{X}$ at the point x. Hence, under the additional assumption that g is Lipschitz continuous with constant L_g, we are able to conclude the following almost sure pointwise bounds for the objective functions.

Corollary 2.35. *Suppose that assumptions (D1) and (G1) hold. Further, suppose that the schedule of sample sizes $\{N_k\}$ is strictly monotonically increasing. Then, for each $x \in \mathcal{X}$ and any $\epsilon > 0$, there exists a finite random variable $k^* = k^*(x, \epsilon) \in \mathbb{N}$ such that*

$$\forall k^* \geq k : \quad \left|\hat{f}_{N_k}(x) - f(x)\right| \leq L_g \frac{\sqrt{(2+\epsilon)\,\mathrm{Log}(N_k)}}{\sqrt{N_k}} \sigma(\widetilde{X}),$$

$\mathbb{Q}_V$-almost surely, where $\sigma(\widetilde{X})$ is given by (2.39) *for the $\mathbb{R}^l$-valued random variable $\widetilde{X}$.*

Pathwise bounds of the form $\sqrt{\mathrm{Log}(N_k)}/\sqrt{N_k}$ for the pointwise approximation error $|\hat{f}_{N_k}(x) - f(x)|$ at any $x \in \mathcal{X}$ are also obtained by Homem-de-Mello [2003], Theorem 3.4, via the theory of large deviations. Provided that the random vectors $\{Z_i^k\}$ are i.i.d. and the estimator $\hat{f}_{N_k}(x) = \frac{1}{N_k}\sum_{i=1}^{N_k} h(x, Z_i^k)$ is unbiased, the assumptions there amount to requiring that the expression $\mathbb{E}^{\mathbb{Q}}[|h(x,Z) - \mathbb{E}^{\mathbb{Q}}[h(x,Z)]|^3]/(\mathbb{V}\mathrm{ar}^{\mathbb{Q}}[h(x,Z)])^{3/2}$ is bounded for any $x \in \mathcal{X}$ and that the schedule $\{N_k\}$ follows $1/N_k = \mathcal{O}(1/k^{\tilde{c}})$ for some constant $\tilde{c} > 2$.

Chapter 3

Global Optimisation of Expensive, Deterministic Objective Functions

We now turn our attention towards global optimisation and are predominantly interested in solving problems of the form

$$\min_{x \in \mathcal{X}} f(x), \tag{3.1}$$

where $f : \mathcal{X} \to \mathbb{R}$ denotes a continuous and nonconvex deterministic function on some nonempty compact set $\mathcal{X} \subset \mathbb{R}^d$ whose evaluation is considered to be expensive. In particular, as motivated in Chapter 1, this setup thus also includes, by abuse of notation, the case where a deterministic approximation to some underlying objective function is minimised, e.g. by solving the approximating problem (1.4) within the SAA strategy for a particular realisation of the underlying random sample.

To effectively solve problem (3.1), *response surface methods* have been developed. The basic idea behind these kind of methods is to approximate the underlying expensive objective function f at a limited number of points by some response surface models, which are cheaper to evaluate and easier to handle. Based on the approximating models, new evaluations points are then iteratively determined through different strategies, upon which the current response surface is refined to capture the global behaviour of the expensive objective function as best as possible. In this way, promising regions for a potential global minimum of f can be identified more conveniently, which can then be explored to eventually find a global minimum.

Within the class of response surface methods, various methods can be distinguished, mainly depending on the intended use. A rather general method that has become popular both from a theoretical and practical perspective is the *RBF method* by Gutmann [2001a,b], which will be of central interest in this chapter. As the name suggests, the method relies on the use of *radial basis functions*, which are theoretically well-established and provide a robust and numerically efficient framework for constructing suitable response surfaces. The involved strategy for determining new evaluation points is based on a mathematically sound mechanism, which not only facilitates establishing convergence of the method but also offers a solid basis for possible extensions. On the practical side, the method has proven to be a powerful tool for performing well on well-behaved expensive optimisation problems, see, e.g., Gutmann [2001a], Björkman and Holmström [2000] or Regis and Shoemaker [2007b].

To initially provide the reader with the necessary background on response surface methods, Section 3.1 is devoted to this class of methods. We describe the common structure and give an overview of existing methods for the global optimisation of deterministic and expensive objective functions. In Section 3.2, we introduce radial basis functions and discuss their main properties. In particular, we address the use of radial basis functions for the purpose of interpolation and state conditions that ensure the existence and uniqueness of radial basis function interpolants. We further examine the linear spaces generated by these functions and show how their induced semi-norm serves as a measure of smoothness. All these features define Gutmann's RBF method, which is then reviewed in the required detail in Section 3.3. This involves a description of the individual steps of the method and main convergence results, as well as a summary of relevant practical aspects concerning its implementation.

3.1 Response Surface Methods

Unlike most approaches from other classes of global optimisation methods, response surface methods intend to make use of the available information in form of already sampled points and their function values as much as possible, due to the expensiveness of objective function evaluations. This is accomplished by constructing response surface models to the expensive objective function, which are substantially cheaper to evaluate and analytically tractable. For most well-behaved objective functions such approximating models are able to provide a good global picture after relatively few function evaluations, which can then be used to decide at which point the objective function should be evaluated next. A naive idea would be to simply choose the global minimum of the approximating model as the next point for evaluation. However, if this approach is iterated, the method neglects any error made by the approximating model and tends to reduce to a local search algorithm in which the global minimum might be missed. A key role is thus not only played by the choice of a response surface model, but also by a careful strategy on how to determine new evaluation points.

In the following subsection, we first present the general structure of response surface methods in further detail. In Subsection 3.1.2, we then give a comprehensive, albeit not exhaustive, overview of relevant methods within this class of optimisation methods for minimising deterministic objective functions.

3.1.1 Structure of Response Surface Methods

Assuming that a procedure to evaluate the expensive objective function f at any $x \in \mathcal{X}$ is in place, the structure of response surface methods for solving problem (3.1) may be described in elementary terms by Algorithm 3.1, cf. Locatelli and Schoen [2013], Section 3.2.

Algorithm 3.1. *(General Framework of Response Surface Methods).*

```
0. Initial step:
    - Generate a set of initial points $\{x_1, \ldots, x_{n_0}\} \subset \mathcal{X}$.
    - Evaluate $f$ at the points $x_1, \ldots, x_{n_0}$, and set $n = n_0$.
1. Iteration step:
    while a suitable stopping criterion is not satisfied do
        - Construct a response surface model $s_n$ to $f$ at the set of points $\{x_1, \ldots, x_n\}$.
        - Determine the next evaluation point $x_{n+1}$ by optimising a utility function $u_n$.
        - Evaluate $f$ at $x_{n+1}$, and set $n = n + 1$.
    end while
```

Accordingly, response surface methods essentially consist of three main components: an initialisation, the construction of suitable approximating models and the determination of new evaluation points, each of which shall briefly be described hereinafter.

Most response surface methods are initialised by generating a set of sample points $\{x_1, \ldots, x_{n_0}\} \subset \mathcal{X}$, $n_0 \in \mathbb{N}$, at which f is first evaluated. The set of initial points is typically defined either manually or through experimental design, where available strategies are plentiful and range from Latin hypercube designs over low-discrepancy sequences to different optimal designs, see, e.g., Pukelsheim [2006], or the more condensed overviews of Fowkes [2011], Chapter 4, or Vu et al. [2017], Section 2, and the references therein. A desirable criterion for a strategy is to spread the initial sample points equally well over the parameter space. Moreover, to guarantee that the initial model fit is well-defined, the points are supposed to satisfy some requirements regarding their arrangement, such as being unisolvent.

Following their initialisation, response surface methods then proceed iteratively, by fitting a response surface model s_n to the available data $(x_1, f(x_1)), \ldots, (x_n, f(x_n))$, in order to find one or more data points for successive function evaluation and updating the approximating model. To the most common basis functions for constructing response surfaces belong *low-degree polynomials*, as used in the traditional response surface methodology (e.g., Myers and Montgomery [1995]), *multivariate adaptive regression splines* (e.g., Friedman [1991]), *kriging* (e.g., Sacks et al. [1989] and Cressie [1991]) and *radial basis functions* (e.g., Powell [1992] and Buhmann [2003]). Besides, neural networks and support vector machines (e.g., Schölkopf and Smola [2002]), and, more recently, *kernels* as generalisations of radial basis functions (e.g., Schaback and Wendland [2006]) have also been considered. Closely related to the choice of a suitable basis function is the technique by which response surfaces are constructed from available function values. Typically, if observed function values are known to be exact, then an *interpolation* scheme is adopted to build the surfaces, while for inexact observations, i.e. function values that are contaminated by random noise or other forms of inaccuracies, an *approximation* (or *regression*) technique is sought. Still, however, there are cases when even in a setup with exact function values an approximation is more appropriate, such as for smoothing out unnecessary oscillations of the true function f.

Once a response surface model is constructed in an iteration, it serves as a basis for selecting one or more points[11] at which the objective function is evaluated next. The determination of new evaluation points is the most crucial step in the procedure for which various strategies are available, which either depend on the approximating model or work in general with any surface. Most commonly, the strategies are set up as a subproblem in which a suitable *utility function* u_n, also termed *merit* or *loss function*, is specified and optimised to determine the next point for evaluation. Conforming to Jones [2001], the strategies may broadly be categorised into two main families, *two-stage* and *one-stage* methods. Strategies of the former family are exclusively based on the sample data and deem the derived approximating model as sufficiently good for the objective function. In the first stage, it is thus used to estimate any further required quantities, which are then employed in the second stage together with the model to determine new evaluation points. The most classical example here is the minimisation of the response surface itself. Moreover, all criteria given for Bayesian and regression-based methods in Subsection 3.1.2 below are two-stage strategies. By contrast, one-stage strategies omit the first stage of estimating model-dependent quantities and directly determine a new point under the additional assumption of a target value for the objective function. Since it is considered desirable to achieve this value, the location of the next evaluation can then be chosen as the point which makes the occurrence of this target value based on previous sample data the ‘most likely’, as defined by some suitable criterion. The most representative example of this family is the RBF method, where the ‘bumpiness’ of a surface is used as a measure of likeliness, see Section 3.3. In any case, an ideal strategy is endowed with an adjustable parameter or the like, allowing to balance between selecting points in rather unexplored regions of the parameter space to improve the accuracy of the model there and trusting the approximating model in regions with many function evaluations to find a minimum thereof. In this way, promising regions for a global minimum of the objective function are first identified and subsequently exploited by the methods.

Eventually, response surface methods are terminated if some stopping criterion is met. In most cases, this is when a maximum number of function evaluations is reached, or when the improvement over the current best solution/function value in a certain number of iterations is below a prescribed threshold.

3.1.2 Overview of Related Literature

The response surface framework outlined above is very general and simplistic, without any closer specification of the individual components. We now give a more thorough overview of

[11] Note that the iteration step of Algorithm 3.1 is designed for single function evaluations but can easily be amended to include more evaluations.

how these components are specified within the most relevant global response surface methods for deterministic objective functions. For a comparative survey of various response surface approaches regarding their suitability for global optimisation, we refer to Jones [2001] and, more recently, Vu et al. [2017]. A more practical overview is given, for instance, in Forrester and Keane [2009]. Eventually, note that a summary of response surface methods dealing with noisy objective functions is presented in Section 5.1.

Very generally, one may say that there are three main methodologies according to which traditional response surface methods may be classified. Due to their setup, the first two approaches, *Bayesian* and *regression-based methods*, provide an appealing probabilistic framework to account for the uncertainty in the objective function, but nevertheless result in deterministic methods. In contrast, methods based on *Jones's general technique* follow a purely deterministic approach using linear function spaces. Finally, we also list some response surface methods that do not seem to have originated from any of these methodologies.

Bayesian Methods

According to Gutmann [2001b], Section 2.1, underlying idea of Bayesian methods is to understand the objective function f as a realisation of a stochastic process $F : \mathcal{X} \times \Omega \to \mathbb{R}$, $(x, \omega) \mapsto F(x, \omega)$, on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Hence, for each $x \in \mathcal{X}$, $F(x, \cdot)$ is a real-valued random variable (denoted by $F(x)$ in short), and there exists an $\omega \in \Omega$ with $f(x) = F(x, \omega)$, $x \in \mathcal{X}$.

Now, given the observations $f(x_1), \dots, f(x_n)$, the objective function f may be considered as a realisation of the process $\{F(x)\}$ conditioned on the information $F(x_1) = f(x_1), \dots, F(x_n) = f(x_n)$. In particular, it is thus reasonable to choose the process $\{F(x)\}$ such that the conditional mean function

$$s_n(x) = \mathbb{E}^{\mathbb{P}}\big[F(x) \,\big|\, F(x_1) = f(x_1), \dots, F(x_n) = f(x_n)\big], \quad x \in \mathcal{X},$$

and the conditional variance

$$v_n(x) = \mathbb{V}\mathrm{ar}^{\mathbb{P}}\left[F(x) \,\middle|\, F(x_1) = f(x_1), \dots, F(x_n) = f(x_n)\right], \quad x \in \mathcal{X},$$

supposed to act as a response surface and the involved error, respectively, are easily computable. This, in fact, is given for Gaussian processes[12], which have the appealing property that their conditioning on a set of observed function values is again Gaussian. Specifically, if $\{F(x)\}$ is a Gaussian process with mean function $\mu_F(x) = \mathbb{E}^{\mathbb{P}}[F(x)]$ and covariance function $\Sigma_F(x, y) = \mathbb{C}\mathrm{ov}^{\mathbb{P}}(F(x), F(y))$, $x, y \in \mathcal{X}$, the distributional properties imply that the conditional process has the mean function

$$s_n(x) = \mu_F(x) + \big(f(x_1) - \mu_F(x_1), \dots, f(x_n) - \mu_F(x_n)\big)\widetilde{\Sigma}^{-1}k(x), \qquad x \in \mathcal{X},$$

[12] A real-valued stochastic process is called Gaussian if the distribution of any finite number of its random variables is multivariate normal, see, e.g., Klenke [2008], Definition 9.7.

and the variance

$$v_n(x) = \Sigma_F(x,x) - k(x)^\top \widetilde{\Sigma}^{-1} k(x), \quad x \in \mathcal{X},$$

where the matrix $\widetilde{\Sigma} \in \mathbb{R}^{n \times n}$ is given by $\widetilde{\Sigma}_{ij} = \Sigma_F(x_i, x_j)$, $i, j = 1, \ldots, n$, and $k(x) = (\Sigma_F(x, x_1), \ldots, \Sigma_F(x, x_n))^\top \in \mathbb{R}^n$. In particular, since $k(x_i)$ corresponds to the i-th column of $\widetilde{\Sigma}$, it thus follows that $s_n(x_i) = f(x_i)$ and $v_n(x_i) = 0$, $i = 1, \ldots, n$, such that s_n and v_n indeed provide a response surface to the objective function f and a measure of the associated error, respectively.

As for determining a new evaluation point x_{n+1} by means of an utility function u_n in a Bayesian method, there are two main strategies. The first one is to maximise the probability of achieving a certain target function value below the current minimum of the surface $\min_{x \in \mathcal{X}} s_n(x)$. Accordingly, by letting $\rho_n > 0$ denote a positive threshold, the related utility function can be formulated as

$$u_n(x) = \mathbb{P}\big(F(x) \leq \min_{y \in \mathcal{X}} s_n(y) - \rho_n \,\big|\, F(x_1) = f(x_1), \ldots, F(x_n) = f(x_n)\big), \quad x \in \mathcal{X},$$

whose maximisation then yields a point close the current best sample point for sufficiently small ρ_n and a point far away from the sampled points for a considerably large value. Due to the use of a probability in the derivation, this method is known in the literature as *P-algorithm*, dating back to Kushner [1962, 1964]. In particular, under the assumption that the underlying process $\{F(x)\}$ is Gaussian, the utility function may be written in analytical form as

$$u_n(x) = \mathcal{N}\left(\frac{\min_{y \in \mathcal{X}} s_n(y) - \rho_n - s_n(x)}{\sqrt{v_n(x)}}\right),$$

where $\mathcal{N}(\,\cdot\,)$ denotes the standard normal cumulative distribution function.

The second strategy is to determine a new point x_{n+1} by maximising the expected improvement over the current best function value $f_n^{\min} := \min\{f(x_1), \ldots, f(x_n)\}$. The utility function thus becomes

$$u_n(x) = \mathbb{E}^{\mathbb{P}}\big[\max\{f_n^{\min} - F(x), 0\} \,\big|\, F(x_1) = f(x_1), \ldots, F(x_n) = f(x_n)\big], \quad x \in \mathcal{X}, \qquad (3.2)$$

which is maximised to balance between local and global search irrespective of any additional parameter. This approach is known as the *Expected Improvement Algorithm* and has its origin in Mockus et al. [1978]. In particular, if $\{F(x)\}$ is conditionally Gaussian with mean $s_n(x)$ and variance $v_n(x)$, the utility function u_n may be simplified (using integration by parts) as

$$u_n(x) = \big(f_n^{\min} - s_n(x)\big)\mathcal{N}\left(\frac{f_n^{\min} - s_n(x)}{\sqrt{v_n(x)}}\right) + \sqrt{v_n(x)}\,\mathcal{N}'\left(\frac{f_n^{\min} - s_n(x)}{\sqrt{v_n(x)}}\right), \qquad (3.3)$$

for $v_n(x) > 0$, and $u_n(x) = 0$ otherwise, where $\mathcal{N}'(\,\cdot\,)$ denotes the standard normal probability density function.

Regression-Based Methods

Closely related to Bayesian methods are regression-based methods, which still consider the objective function f as a realisation of a stochastic process but use a linear regression to fit a response surface model. The regression technique, which is commonly also referred to as kriging (Matheron [1963]), will then generally lead to different methods and results except in the special case of a Gaussian process, see, e.g., Fowkes [2011], Sections 3.1 and 3.2, for a detailed derivation of this equivalence.

Following Sacks et al. [1989], the underlying model treats the objective function f as a realisation of a stochastic process in form of

$$F(x) = \sum_{j=1}^{\widetilde{m}} c_j p_j(x) + Z(x), \quad x \in \mathcal{X}, \tag{3.4}$$

where p_j, $j = 1, \dots, \widetilde{m}$, are known regression functions with unknown parameters c_j, and $\{Z(x)\}$ is a stochastic process. The process $\{Z(x)\}$ is assumed to have mean zero and covariance $\mathbb{C}\mathrm{ov}^{\mathbb{P}}(Z(x), Z(y)) = \sigma_Z^2 \Gamma_Z(x, y)$, $x, y \in \mathcal{X}$, for a scaling factor σ_Z^2 (also called process variance) and some correlation function $\Gamma_Z(\cdot, \cdot)$. Given the observed function values $f(x_1), \dots, f(x_n)$, the specification (3.4) then amounts to the linear regression model

$$f_X = Pc + z,$$

where $f_X = (f(x_1), \dots, f(x_n))^\top$, $P \in \mathbb{R}^{n \times \widetilde{m}}$ denotes the design matrix with entries $P_{ij} = p_j(x_i)$, $i = 1, \dots, n$, $j = 1, \dots, \widetilde{m}$, $c = (c_1, \dots, c_{\widetilde{m}})^\top \in \mathbb{R}^{\widetilde{m}}$ the parameter vector, and $z = (Z(x_1), \dots, Z(x_n))^\top$.

Now, to find the best linear unbiased predictor of the form $s_n(x) = a(x)^\top f_X$ for f at any untried $x \in \mathcal{X}$, the unknown vector of coefficients $a(x) \in \mathbb{R}^n$ needs to be determined such that it minimises the mean squared error (MSE)

$$\mathbb{E}^{\mathbb{P}}\left[\Big(\sum_{i=1}^{n} a_i(x) F(x_i) - F(x)\Big)^2\right], \tag{3.5}$$

subject to the unbiased constraint $\mathbb{E}^{\mathbb{P}}[\sum_{i=1}^{n} F(x_i) - F(x)] = 0$. Using standard statistical estimation techniques, this can be readily derived as

$$s_n(x) = \pi(x)^\top \hat{c} + r(x)^\top \widetilde{\Gamma}^{-1}\big(f_X - P\hat{c}\big), \quad x \in \mathcal{X},$$

where $\pi(x) = (p_1(x), \dots, p_{\widetilde{m}}(x))^\top \in \mathbb{R}^{\widetilde{m}}$, $\hat{c} = (P^\top \widetilde{\Gamma}^{-1} P)^{-1} P^\top \widetilde{\Gamma}^{-1} f_X$ is the usual generalised least-squares estimate of c, $\widetilde{\Gamma} \in \mathbb{R}^{n \times n}$ denotes the correlation matrix with $\widetilde{\Gamma}_{ij} = \Gamma_Z(x_i, x_j)$, and $r(x) = (\Gamma_Z(x, x_1), \dots, \Gamma_Z(x, x_n))^\top \in \mathbb{R}^n$. The corresponding MSE in x is then the function

$$v_n(x) = \sigma_Z^2 \left[1 - \begin{pmatrix} r(x) \\ \pi(x) \end{pmatrix}^\top \begin{pmatrix} \widetilde{\Gamma} & P \\ P^\top & 0 \end{pmatrix}^{-1} \begin{pmatrix} r(x) \\ \pi(x) \end{pmatrix}\right], \quad x \in \mathcal{X},$$

where the parameter σ_Z^2 (and possibly any correlation parameters) is usually determined in each iteration by maximum likelihood estimation to the sample. Similar to the case of Bayesian methods, it is easy to verify for the prediction that $s_n(x_i) = f(x_i)$ and for the MSE that $v_n(x_i) = 0$, $i = 1, \ldots, n$. Thus, the predictor s_n interpolates the data and the MSE v_n at the data points is zero.

The most popular method embedded within above methodology is the *Efficient Global Optimisation* (EGO) algorithm by Jones et al. [1998], which uses a constant c as mean function and specifies the correlation according to

$$\Gamma_Z(x, y) = \prod_{j=1}^{d} \mathrm{e}^{-\theta_j |x_j - y_j|^2}, \quad x, y \in \mathcal{X}, \tag{3.6}$$

for some nonnegative parameters θ_j, $j = 1, \ldots, d$. Hence, this leads to expressions for s_n and v_n with $P = \mathbf{1}$ and $\pi(x) = 1$, and where the matrix $\widetilde{\Gamma}$ is specified accordingly through (3.6). Moreover, it is assumed that the random process $\{Z(x)\}$ in (3.4) is Gaussian, such that the normally distributed random variable $F(x)$ has mean value $s_n(x)$ and variance $v_n(x)$. This enables that EGO finds the next evaluation point by using the *expected improvement* criterion (3.2), as earlier suggested for Bayesian methods.

The expected improvement criterion balances between local and global search without the need to specify any explicit parameter. For poorly estimated correlation parameters, however, this may result in a search that is too local, as observed by Schonlau [1997]. Therefore, he suggests using a *generalised expected improvement*, which maximises the utility function

$$u_n(x) = \mathbb{E}^{\mathbb{P}}\Big[\max\big\{\big(f_n^{\min} - F(x)\big)^q, 0\big\}\Big], \quad x \in \mathcal{X},$$

where $q \in \mathbb{N}_0$ is an additionally introduced parameter that determines how global versus local the search will be. In particular, for $q = 0$, the criterion yields the probability of improvement $\mathcal{N}((f_n^{\min} - s_n(x))/\sqrt{v_n(x)})$, while for higher values the emphasis shifts more and more towards global search. Another modification allowing to exogenously control local and global search is the *weighted expected improvement*, due to Sóbester et al. [2005], which maximises

$$u_n(x) = w^{\mathrm{EI}}\big(f_n^{\min} - s_n(x)\big)\mathcal{N}\left(\frac{f_n^{\min} - s_n(x)}{\sqrt{v_n(x)}}\right) + \big(1 - w^{\mathrm{EI}}\big)\sqrt{v_n(x)}\mathcal{N}'\left(\frac{f_n^{\min} - s_n(x)}{\sqrt{v_n(x)}}\right),$$

for a weighting factor $w^{\mathrm{EI}} \in [0, 1]$. Here, a small value of w^{EI} will enforce a global search in unexplored regions of the parameter space, while a large value will concentrate the search around the current best sample point.

The EGO approach to construct suitable response surfaces and estimate the related correlation parameters is also used by Villemonteix et al. [2009] in their method, called *Informational Approach to Global Optimisation* (IAGO). However, instead of expected improvement, they use the conditional entropy of a minimiser as a criterion to iteratively determine new evaluation points, which is then minimised accordingly. In particular, to compute the entropy, the authors make explicit use of a Gaussian random process model for f, which in turn allows to estimate the entropy via regression-based techniques and conditional simulations.

The probabilistic interpretation of s_n as best predictor and of v_n as involved error also facilitates the derivation of other utility functions. Besides the rigorous (but often naive) choices of minimising $u_n(x) = s_n(x)$ and maximising $u_n(x) = v_n(x)$, Cox and John [1997], for instance, suggest to define the next evaluation point as the minimum of a lower confidence bounding function, which constitutes a compromise between the minimum of the response surface and the maximum error. Specifically, this approach minimises

$$u_n(x) = s_n(x) - \tau\sigma_z\sqrt{v_n(x)},$$

where τ is a suitably chosen positive constant balancing between local and global search.

Response Surface Methods Based on Jones's General Technique

For methods that do not model the objective function by means of stochastic processes, a general response surface technique for finding a new evaluation point is proposed by Jones [1996] (see also Gutmann [2001b], Section 2.3). It is inspired by the P-Algorithm in that it employs a target value to decide where to evaluate next. However, since no underlying probabilistic interpretation is possible, the notion of the 'bumpiness' of a response surface is used as a criterion to base this decision on. Accordingly, a new point should then be chosen such that the interpolating surface through previous function values and the chosen target value at the new point becomes the least 'bumpy'.

More formally, Jones considers a linear function space $\mathcal{A}$, which is left unspecified but admits a measure of 'bumpiness' for its elements $s \in \mathcal{A}$. Once a response surface model is fit to a set of available function values $f(x_1), \dots, f(x_n)$ in an iteration through interpolation, a target value f^* is chosen which may be considered as a rough estimate of the global minimum of f. By this choice, a new evaluation point x_{n+1} is then determined as the value of $y \in \mathcal{X}\backslash\{x_1, \dots, x_n\}$ such that the augmented response surface $s_y \in \mathcal{A}$ minimises the 'bumpiness' of s_y on $\mathcal{A}$, subject to the interpolation conditions

$$\begin{aligned} s_y(x_i) &= f(x_i), \qquad i = 1, \dots, n, \\ s_y(y) &= f^*, \end{aligned}$$

and provided that for any $y \in \mathcal{X}\backslash\{x_1, \dots, x_n\}$, the interpolant $s_y \in \mathcal{A}$ is uniquely defined.

The most popular response surface method using Jones's general concept is developed in the *RBF method* by Gutmann [2001a,b], showing how radial basis function can effectively be employed in this technique to minimise a continuous nonconvex and expensive black-box objective function. Specifically, by considering interpolants s that are composed of radial basis functions ϕ, a 'bumpiness' measure in form of the semi-norm of s can be introduced on the linear space $\mathcal{A}_\phi$ of all interpolants, which is then used to determine new evaluation points as suggested by Jones, see Section 3.3 for details. For most types of radial basis functions considered, Gutmann shows convergence of the method to a global minimum without any further assumptions on the objective function. Moreover, he establishes that the method is closely related to a statistical optimisation method, the P-Algorithm, even though coming from a different underlying model. Eventually, the method is compared with other global optimisation methods on a standard set of test functions for which it yields promising results.

Based on Gutmann's original RBF method, there have been several improvements and extensions, which shall be briefly outlined in what follows. Björkman and Holmström [2000] present an improved numerical implementation of Gutmann's RBF method, called *rbfSolve*, which is part of their TOMLAB optimisation environment[13] in MATLAB. By applying their implementation to a set of standard test problems, they show that it can be very efficient compared to other global solvers such as DIRECT (Jones et al. [1993]), EGO (Jones et al. [1998]) or MCS (Huyer and Neumaier [1999]).

Regis and Shoemaker [2007b] observe that for certain problems, e.g. problems in which the global minimum of the objective function is located in steep valleys, the RBF method converges only slowly, due to the failure of performing local search. To circumvent this issue and improve the performance of the method, they suggest to restrict the minimisation of the 'bumpiness' measure for determining a new point to a subregion around a minimiser of the current response surface, whose size varies for better distinction between local and global search. Moreover, they also suggest using a complete restart strategy whenever the RBF method does not make substantial progress within a specified number of iterations. Eventually, they show through computational results that both strategies are able to significantly improve the performance of Gutmann's RBF method on some of the test instances for which the original method exhibits a slow convergence.

A further extension to Gutmann's RBF method is given by the *adaptive radial basis function (ARBF) method* of Holmström [2008]. Based on numerical tests, he points out that the performance of Gutmann's method notably depends on the scaling of the problem and is very sensitive to the static selection of target values determining new evaluation points. In particular, points are frequently sampled on the boundary of the domain, hindering its practical convergence. To make the choice of target values more flexible and improve the convergence, Holmström suggests to perform an extensive search for suitable target values in each iteration of the method, resulting in a sequence of utility functions to be solved per iteration. It is shown that the ARBF method has similar theoretical convergence properties as Gutmann's RBF method but achieves better numerical results on a standard set of test problems, due to the modification.

In Cassioli and Schoen [2013], a modified version of Gutmann's RBF method is derived for problems in which a lower bound of the objective function is known, such as in data-fitting problems. Specifically, the knowledge of a lower bound is used to introduce additional constraints in the construction of radial basis function interpolants, such that the result-

[13] See `http://tomopt.com/tomlab/` for details.

ing response surface models become more accurate and avoid that the method selects rather unrealistic target values. By extensive numerical experiments on a large class of small dimensional yet challenging test problems, both authors then show that the suggested modification can improve Gutmann's RBF method if small extra computational effort is invested.

Eventually, Costa and Nannicini [2014] propose two further modifications of Gutmann's RBF method in order to improve its practical performance. Their first suggestion addresses the choice of a suitable radial basis function for any given black-box optimisation problem, which is often difficult to determine beforehand and to which end they provide an automated model selection methodology by means of a cross-validation scheme. Their second modification concerns the practical convergence of the RBF method to a global minimum of the underlying unknown function. In particular, they show that in case cheaper but noisy objective function values are additionally available and lie within some error bounds around their true values, both types of function values can be combined to achieve a noticeable speed-up in the convergence of the method.

Further Response Surface Methods

Aside from the above response surface methods, there are further approaches which seemingly do not rely on any of the described underlying methodologies and are thus designed to work with any kind of response surface model. We now also outline some of these contributions.

In Regis and Shoemaker [2005], the authors introduce the *constrained optimisation using response surfaces (CORS) method* for minimising a black-box objective function. Given the construction of a response surface in an iteration of their method, the next evaluation point is obtained by minimising the surface model subject to the constraints that the new point is of some distance from previously evaluated points, where the balance between local and global search is established by cycling through multiple distances. The so-defined method is shown to converge to the global minimiser of any continuous function on a compact set. Moreover, unlike usual response surface methods, it is extendable to general nonlinear constraints. Finally, both authors demonstrate by means of numerical experiments that the method, when used with radial basis functions, is competitive with other global optimisation methods in terms of function evaluations.

Regis and Shoemaker [2007a] also present a *stochastic response surface (SRS) method* which may be used with any surface model. The method selects new points for function evaluation from a set of randomly generated candidate points according to their response surface value and the distance to previously sampled points, where different weights are set cyclically on both measures to balance between local and global search. Under some mild technical assumptions on the probability distribution generating the candidate points, the SRS method can be shown to converge to a global minimum of any continuous function on a compact set with probability one.

A framework for parallel global optimisation of expensive functions with response surface models is provided in Regis and Shoemaker [2007c]. Although, in general, any response surface method may be used within their framework, the authors primarily focus on parallelising Gutmann's RBF method and the CORS method, where in each iteration multiple points are generated for simultaneous function evaluation in parallel. Using radial basis functions, both parallelisations are then compared with other parallelised versions for global optimisation, i.e. differential evolution, a multi-start SQP and UOBYQA (Powell [2002]), where results indicate that the former outperform their counterparts on chosen test problems.

Regis and Shoemaker [2013] further develop a *quasi multi-start response surface (AQUARS) framework* where, upon the construction of a response surface model in an iteration of the algorithm, either a global search strategy is applied to determine a new point from a previously unexplored region or a local strategy is adopted to run local searches near the local minima of the response surface model. By using response surface models, the framework is further able to prevent unnecessary local searches in neighbourhoods of already explored local minima. Moreover, it is supposed to work with any global and local search procedures. While no theoretical convergence results of AQUARS are available, it can be shown to perform very well with radial basis functions on a number of test problems, especially for those problems in which global minima are located in narrow valleys.

A similar framework for combining global and a local search procedures is presented by Ji et al. [2013]. Essentially, it consists of a global search procedure which maintains a response surface model throughout the entire optimisation in order to find unexplored regions, and a local one which is used for exploiting promising search regions up to optimality whenever detected. To adaptively switch between both procedures, a set of candidate sample points is generated in both, whose potential for successive function evaluation and further execution is then assessed by a performance measure called modified expected improvement. This measure extends the expected improvement criterion to general models, and relies on the response surface model value as well as on the minimum distance to previously evaluated points. Under mild regularity conditions, the algorithm is shown to converge to the global optimum of any continuous function on a compact set. Using radial basis functions and kriging-based response surface models, the algorithm is ultimately tested against rbfSolve, SRS and EGO, among others, indicating that the enforced local search enhances the performance.

3.2 Radial Basis Functions

Response surface methods require the construction of approximating models to a multivariate function at a limited number of points in general position, for which radial basis function methods offer a particularly well-suited and powerful framework. Above all, they are successfully employed in Gutmann's RBF method, see Section 3.3, which serves as a basis for our modification and extension to noise in Chapter 4 and Section 5.3, respectively. In view of these methods, the aim of this section is to provide the reader with all relevant material on radial basis functions needed for the description and the involved analysis of these methods for global optimisation. For the time being, our interest lies in radial basis function interpolation and the definition of a suitable measure of smoothness, both being essential ingredients of Gutmann's method and our first modification. Later on, in Section 5.2, we will additionally address the issue of radial basis function approximation as required by the RBF method for noisy objective functions. Comprehensive surveys on properties of radial basis functions are given, for instance, by Powell [1992] and Buhmann [2003].

For a given continuous function $\phi : [0, \infty) \to \mathbb{R}$, radial basis function[14] approximants are in their most generic form given by

$$s(x) := \sum_{i=1}^{n} \lambda_i \phi(\|x - x_i\|_2) + p(x), \qquad x \in \mathbb{R}^d, \tag{3.7}$$

where $\{\lambda_i\}_{i=1}^n$ are real coefficients, $\{x_i\}_{i=1}^n \subset \mathbb{R}^d$ are distinct centre points, and $p \in \mathcal{P}_m^d$ is a polynomial from the linear space of all real-valued polynomials of total degree at most $m-1$ in d variables, where $\mathcal{P}_0^d = \{0\}$, i.e. the zero polynomial.

Employing approximants in form of (3.7) has several appealing features. Firstly, by using a finite linear combination of basis functions that are radially symmetric, i.e. whose value at a point depends on its distance to the centres, the approximants are conceptually simple and can easily be used in a multivariate setup. In particular, their construction as well as their evaluation can take place often and efficiently, provided that the number of centres does not increase exceedingly. Secondly, due to the use of radial basis functions, approximants of the form (3.7) encode certain kind of smoothness and thus have the inherent ability to model well the curvature of smooth target functions. This characteristic property is not only beneficial from the point of view of approximation theory, but also plays a significant role in some response surface methods. Eventually, the addition of (low-degree) polynomials to the sum of radial basis function provides the user with more freedom to construct suitable approximants, guaranteeing their existence and uniqueness under mild conditions for a much wider class of radial basis functions, as shall be seen below.

Some of the most common choices of radial basis functions, along with their shape and smoothing parameters, are given in Table 3.1, cf. Gutmann [2001a], Table 3.2. Note that the listed radial basis functions are all of *global support* as they infer their information over the entire domain. In addition to these, families of *compactly supported* radial basis functions have been developed more recently and gained increasing popularity, e.g., Wendland [1995], Wu [1995] and Buhmann [1998]. For an extensive list of radial basis function, especially from a practical point of view, we refer to Fasshauer [2007].

In the following Subsection 3.2.1, we consider the problem of radial basis function interpolation to summarise well-established results guaranteeing the existence and uniqueness of an interpolant. Subsection 3.2.2 then examines the linear spaces generated by functions of the form (3.7), which in fact are semi-inner product spaces, to introduce the induced semi-norm as a measure of smoothness for its elements s.

[14] Note that most of the results presented in the remainder of this chapter are not necessarily restricted to radial basis functions but may be generalised to conditionally definite basis functions, mapping $\mathbb{R}^d \times \mathbb{R}^d$ into $\mathbb{R}$, see, e.g., Gutmann [2001b], Section 4.4.

Radial basis function	$\phi(r)$	Specification
Surface splines	r^ν	$\nu \in \mathbb{N}$, ν odd
	$r^\nu \log r$	$\nu \in \mathbb{N}$, ν even
Multiquadrics	$(r^2 + \zeta^2)^\nu$	$\nu > 0$, $\nu \notin \mathbb{N}$
Inverse multiquadrics	$(r^2 + \zeta^2)^\nu$	$\nu < 0$
Gaussians	$e^{-\zeta r^2}$	

Table 3.1: Common choices of radial basis functions, with smoothing parameter ν and shape parameter $\zeta > 0$.

3.2.1 Radial Basis Function Interpolation

When using approximants of the form (3.7) to reconstruct an unknown function $f : \mathbb{R}^d \to \mathbb{R}$ from a set of function values $f(x_1), \ldots, f(x_n)$ observed at the pairwise distinct points $x_1, \ldots, x_n \in \mathbb{R}^d$, a crucial role is played by the way in which these function values are available. If the values are known to be exact, then an interpolation scheme is usually adopted, where the coefficients of the resulting interpolant s are determined by imposing the conditions

$$s(x_i) = f(x_i), \qquad i = 1, \ldots, n, \tag{3.8}$$

and the centres of the radial basis functions coincide with the interpolation points.

In view of representation (3.7), it would seem obvious to set $m = 0$ and thus omit any polynomials in the linear combination of radial basis functions. This, however, may have the undesirable effect that for some choices of radial basis functions, e.g. surface splines, and a particular arrangement of $x_1, \ldots, x_n$, the interpolation matrix $\Phi \in \mathbb{R}^{n \times n}$ arising from (3.8) and given by

$$\Phi_{ij} := \phi(\|x_i - x_j\|_2), \qquad i, j = 1, \ldots, n, \tag{3.9}$$

becomes singular, such that no interpolant exists. Nevertheless, under certain conditions on the coefficients λ_i, $i = 1, \ldots, n$, the function values $f(x_1), \ldots, f(x_n)$ can always be interpolated if a polynomial of low degree is added to the linear combination of radial basis functions. Specifically, by letting the polynomials $\{p_j\}_{j=1}^{\widetilde{m}}$ be a basis of the space $\mathcal{P}_m^d$ with dimension[15] $\widetilde{m}$, the complementary conditions on λ_i can be formulated as

$$\sum_{i=1}^{n} \lambda_i p_j(x_i) = 0, \qquad j = 1, \ldots, \widetilde{m}, \tag{3.10}$$

thus taking up the additional degrees of freedom introduced through the use of polynomials.

[15] It is a well-known result that the dimension $\widetilde{m}$ of the polynomial space $\mathcal{P}_m^d$ is given by $\widetilde{m} = \binom{m-1+d}{d}$, see, e.g., Wendland [2005b], Theorem 2.5.

Hence, together with the side conditions (3.10), the interpolation conditions (3.8) amount to solving the linear system

$$\begin{pmatrix} \Phi & P \\ P^\top & 0 \end{pmatrix} \begin{pmatrix} \lambda \\ c \end{pmatrix} = \begin{pmatrix} f_X \\ 0 \end{pmatrix}, \tag{3.11}$$

where Φ is the matrix defined by (3.9), $P \in \mathbb{R}^{n \times \widetilde{m}}$ is the polynomial basis matrix with entries

$$P_{ij} := p_j(x_i), \qquad i = 1, \ldots, n,\ j = 1, \ldots, \widetilde{m} \tag{3.12}$$

$\lambda \in \mathbb{R}^n$ and $c \in \mathbb{R}^{\widetilde{m}}$ denote the coefficient vectors, and $f_X = (f(x_1), \ldots, f(x_n))^\top$ stands for the vector of observed function values.

Turning to the unique solvability of the linear system (3.11), the key concept are conditionally positive definite functions. Basically, these functions guarantee the positive definiteness of the interpolation matrix Φ in (3.9) by restricting the corresponding vector of coefficients λ to a linear subspace of $\mathbb{R}^n$, and in this way provide a natural generalisation to all relevant choices of basis functions. More formally, we have the following definition, cf. Wendland [2005b], Definition 8.1.

Definition 3.2. *A continuous radial function $\phi : [0, \infty) \to \mathbb{R}$ is said to be conditionally positive definite of order m on $\mathbb{R}^d$ if the quadratic form*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j \phi(\|x_i - x_j\|_2)$$

is positive for any choice of pairwise distinct points $x_1, \ldots, x_n$, $n \in \mathbb{N}$, and any $\lambda \in \mathbb{R}^n \backslash \{0\}$ satisfying

$$\sum_{i=1}^{n} \lambda_i p(x_i) = 0, \qquad p \in \mathcal{P}_m^d.$$

A radial function ϕ that is conditionally positive definite of order 0 is called positive definite. Moreover, ϕ is called conditionally definite if either ϕ or $-\phi$ is conditionally positive definite.

An immediate consequence of Definition 3.2 is the fact that, for any pair $m_1, m_2 \in \mathbb{N}_0$ with $m_1 \leq m_2$, a conditionally positive definite function of order m_1 is also conditionally positive definite of order m_2, e.g., Wendland [2005b], Proposition 8.2. Of major significance for any radial function ϕ is thus the smallest possible order $m := m(\phi) \in \mathbb{N}_0$ such that ϕ is conditionally positive definite, which will henceforth be simply referred to as *the order* of ϕ.

To verify that a given function ϕ is conditionally positive definite, there exist several approaches. Certainly the most powerful one goes back to Micchelli [1986] and uses the concept of completely monotone functions, cf. Buhmann [2003], Definition 2.1. For another useful characterisation relying on generalised Fourier transforms, we refer to Schaback and Wendland [2001], for instance.

Definition 3.3. *A function $\varphi \in C^\infty(\mathbb{R}_{>0})$ is called completely monotone on $\mathbb{R}_{>0}$ if*

$$(-1)^k \varphi^{(k)}(r) \geq 0$$

holds for all $r \in \mathbb{R}_{>0}$ and all $k \in \mathbb{N}_0$.

A sufficient condition for conditional positive definiteness is then given by the following theorem, due to Micchelli [1986]. Note that the converse of the statement is also true, as shown by Guo et al. [1993], but it is not needed for our purpose.

Theorem 3.4. *Let $\phi : [0,\infty) \to \mathbb{R}$ be a continuous radial function, and suppose that $(-1)^m \phi_{\sqrt{}}^{(m)}$, with $\phi_{\sqrt{}} := \phi(\sqrt{\cdot})$, is well-defined, completely monotone and not constant on $(0,\infty)$. Then, ϕ is conditionally positive definite of order m for all $d \geq 1$.*

By means of Theorem 3.4, it is now straightforward to conclude the conditional positive definiteness of the radial basis functions given in Table 3.1, along with their minimal order m, cf. Gutmann [2001b], Corollary 3.3. Recall that for any real argument $x \in \mathbb{R}$, the floor function $\lfloor x \rfloor$ denotes the largest integer less than or equal to x.

Corollary 3.5. *Let ϕ be a radial basis function from Table 3.1. If $m \in \mathbb{N}_0$ has the value*

$m = \lfloor \nu/2 \rfloor + 1$ *for surface splines,*

$m = \lfloor \nu \rfloor + 1$ *for multiquadrics,*

$m = 0$ *for inverse multiquadrics and Gaussians,*

then $(-1)^m \phi$ is conditionally positive definite of order m.

It is worth noting that Theorem 3.4 implies that either ϕ or $-\phi$ is conditionally positive definite of order m, or, put differently, that ϕ is conditionally definite of order m, as may been seen in Corollary 3.5. However, as the factor -1 can always be absorbed into the coefficients λ_i in the linear combination of radial basis functions (3.7), we can tacitly restrict our prospective analysis to conditionally positive definite radial basis functions, without loss of generality. This clearly simplifies our exposition.

According to Corollary 3.5, if surface splines with $\nu = 2$ or $\nu = 3$ are used, then at least linear polynomials have to be added to the linear combination of radial basis functions to ensure their conditional positive definiteness, while in the special case of multiquadrics with $\nu = 1/2$ a constant polynomial is needed. All other radial basis functions listed in Table 3.1 are already positive definite and thus do not require any additional polynomial term.

The last condition required to guarantee a unique solution of the system (3.11) is the fundamental notion of a unisolvent set, characterising the situation in which an interpolating polynomial is uniquely determined, cf. Wendland [2005b], Definition 2.6.

Definition 3.6. *A set of points* $\{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ *is called* $\mathcal{P}_m^d$*-unisolvent if the zero polynomial is the only polynomial from* $\mathcal{P}_m^d$ *that vanishes in all points* $x_1, \ldots, x_n$*, i.e. if*

$$p \in \mathcal{P}_m^d \quad \textit{and} \quad p(x_i) = 0, \quad i = 1, \ldots, n \qquad \Longrightarrow \qquad p \equiv 0. \tag{3.13}$$

From Definition 3.6, it immediately follows that a set of points is $\mathcal{P}_m^d$-unisolvent if and only if the equation $Pc = 0$ is solved by $c = 0$, where the matrix P is given by (3.12). This, in turn, is equivalent to having $\operatorname{rank}(P) = \widetilde{m}$, such that a $\mathcal{P}_m^d$-unisolvent set has at least $\widetilde{m}$ elements, see, e.g., Gutmann [2001b], Remark 3.5. In particular, note that condition (3.13) poses a rather mild requirement on the geometry of the points $x_1, \ldots, x_n$ for small m. In fact, for $m = 0$, the condition is satisfied by the zero polynomial, for $m = 1$, it is trivial as $\mathcal{P}_m^d$ only contains constants, and for $m = 2$, the points $x_1, \ldots, x_n$ are not allowed to lie on a straight line.

Finally, we are now in place to state the main theorem of this section, guaranteeing the existence of a unique solution to the interpolation system (3.11), cf. Wendland [2005b], Theorem 8.21.

Theorem 3.7. *Let* ϕ *be a conditionally positive definite radial basis function of order* m *and* $\{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ *be a* $\mathcal{P}_m^d$*-unisolvent set. Then, the interpolation system* (3.11) *is uniquely solvable.*

Proof. Consider the related homogeneous linear system

$$\Phi\lambda + Pc = 0 \tag{3.14}$$

$$P^\top \lambda = 0, \tag{3.15}$$

from which we show that it only admits the trivial solution $\lambda = 0$ and $c = 0$. Multiplying (3.14) from the left by $\lambda^\top$ yields $0 = \lambda^\top \Phi \lambda + (P^\top \lambda)^\top c = \lambda^\top \Phi \lambda$, where (3.15) has been used in the last equality. Now, due to the conditional positive definiteness of ϕ and (3.15), this condition is satisfied if and only if $\lambda = 0$. For $\lambda = 0$, however, equation (3.14) reduces to $Pc = 0$ and since $\{x_1, \ldots, x_n\}$ is $\mathcal{P}_m^d$-unisolvent, it also follows that $c = 0$. □

Remark 3.8. *The assumption of* $\mathcal{P}_m^d$*-unisolvency of the set* $\{x_1, \ldots, x_n\}$ *in Theorem 3.7 is only required to show the uniqueness of a solution of the linear system but not its existence. Moreover, it is easy to verify that a uniquely solvable linear system will also remain uniquely solvable upon the successive addition of new data points, provided that the points are distinct from previous ones.*

3.2.2 Variational Theory of Radial Basis Functions

Radial basis functions have the inherent property of producing relatively smooth interpolants without too many oscillations. Of particular interest, however, are not only interpolants that satisfy given interpolation conditions but also have the least curvature among all interpolants subject to the same conditions. This rationale originates from the univariate case of spline interpolation, where the natural cubic spline minimises the curvature (given by the integral over the squared second derivative) over all interpolants (e.g., Knott [2000]). By analogy with natural cubic splines, the concept can be carried over to the multivariate case of radial basis function interpolation, where the space generated by functions of the form (3.7) constitutes a semi-inner product space and the induced semi-norm may be employed as a measure of smoothness. This construction then enables to alternatively characterise radial basis function interpolants from a variational perspective, namely as solutions of a minimisation problem. Moreover, the semi-inner product gives rise to a larger function space, the so-called *native space*, which plays an important role in the convergence theory of Gutmann's RBF method.

To outline the main results of this subsection, we follow along the lines of Gutmann [2001b], Sections 3.2 and 3.3. For a more general perspective on these topics, we refer to the books of Buhmann [2003] and Wendland [2005b], for instance.

3.2.2.1 Characterisation of Interpolants

Let ϕ be a conditionally positive definite radial basis function of order m, and let $\mathcal{D}$ be a subset of $\mathbb{R}^d$ such that it contains at least one $\mathcal{P}_m^d$-unisolvent set. Then, we define the linear function space of all possible approximants s of the generic form (3.7) by

$$\mathcal{A}_\phi(\mathcal{D}) := \mathcal{F}_\phi(\mathcal{D}) + \mathcal{P}_m^d, \tag{3.16}$$

where the linear space

$$\mathcal{F}_\phi(\mathcal{D}) := \left\{ \sum_{i=1}^{n} \lambda_i \phi(\|\cdot - x_i\|_2) : \, n \in \mathbb{N}, \lambda \in \mathbb{R}^n, \{x_i\}_{i=1}^n \subset \mathcal{D}, \sum_{i=1}^{n} \lambda_i p(x_i) = 0, p \in \mathcal{P}_m^d \right\}$$

contains all finite linear combinations of the radial basis function ϕ whose coefficients satisfy the side conditions (3.10) on the polynomial space $\mathcal{P}_m^d$. On $\mathcal{A}_\phi(\mathcal{D})$, we can introduce a semi-inner product $\langle \cdot , \cdot \rangle_\phi$ by

$$\langle s_1, s_2 \rangle_\phi := \sum_{i=1}^{n(s_1)} \lambda_i^{s_1} s_2\big(x_i^{s_1}\big), \tag{3.17}$$

for any two elements $s_1, s_2 \in \mathcal{A}_\phi(\mathcal{D})$ with

$$s_1(x) = \sum_{i=1}^{n(s_1)} \lambda_i^{s_1} \phi(\|x - x_i^{s_1}\|_2) + p^{s_1}(x) \quad \text{and} \quad s_2(x) = \sum_{j=1}^{n(s_2)} \lambda_j^{s_2} \phi(\|x - x_j^{s_2}\|_2) + p^{s_2}(x).$$

That expression (3.17) is indeed a semi-inner product[16] follows immediately from the radial symmetry and the conditional positive definiteness of the basis function ϕ. Moreover, the

[16] Recall that a semi-inner product $\langle \cdot , \cdot \rangle$ on some linear space $\mathcal{A}$ satisfies the same properties as an inner product, i.e. it is bilinear, symmetric and nonnegative, but $\langle s, s \rangle = 0$ need not imply $s = 0$ for $s \in \mathcal{A}$. The same thus also applies to its induced semi-norm $\|\cdot\|$.

semi-inner product in (3.17) naturally induces the semi-norm $\|\cdot\|_\phi$ by

$$\|s\|_\phi := \langle s, s\rangle_\phi^{1/2}, \qquad s \in \mathcal{A}_\phi(\mathcal{D}), \tag{3.18}$$

such that the square of this semi-norm can be explicitly evaluated as

$$\|s\|_\phi^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \phi(\|x_i - x_j\|_2).$$

It is to be noted that for $\langle s, s\rangle_\phi = 0$, the conditional positive definiteness of ϕ implies $\lambda_i = 0$, $i = 1, \ldots, n$, such that $s \in \mathcal{P}_m^d$. Hence, the semi-inner product $\langle \cdot, \cdot \rangle_\phi$ has the null space $\mathcal{P}_m^d$, and it thus becomes a proper inner product if restricted to $\mathcal{F}_\phi(\mathcal{D})$. Therefore, $\mathcal{F}_\phi(\mathcal{D})$ equipped with $\langle \cdot, \cdot \rangle_\phi$ constitutes a pre-Hilbert space.

An important characteristic of above construction is that radial basis function interpolants can be equivalently formulated as optimal solutions of a minimisation problem on the linear space $\mathcal{A}_\phi(\mathcal{D})$. This alternative characterisation, dating back to Schaback [1993] and being sometimes also referred to as the *variational principle*, has turned out to be particular convenient for the development of Gutmann's RBF method, as we shall see in Section 3.3.

Theorem 3.9. *Let ϕ be a conditionally positive definite radial basis function of order m. Assume that $\mathcal{D} \subset \mathbb{R}^d$, and let $\{x_1, \ldots, x_n\} \subset \mathcal{D}$ be a given $\mathcal{P}_m^d$-unisolvent set of points with corresponding function values $f(x_1), \ldots, f(x_n)$. Then, the interpolant $s^* \in \mathcal{A}_\phi(\mathcal{D})$ of the form* (3.7) *whose coefficients solve the linear system* (3.11) *is the unique solution of*

$$\begin{aligned} &\min_{s \in \mathcal{A}_\phi(\mathcal{D})} && \|s\|_\phi \\ &s.t. && s(x_i) = f(x_i), \quad i = 1, \ldots, n. \end{aligned} \tag{3.19}$$

Proof. See the proof of Schaback [1993], Theorem 4. □

As an immediate implication of the proof of Theorem 3.9, note that the centres of the optimal interpolant $s^* \in \mathcal{A}_\phi(\mathcal{D})$ solving (3.19) always correspond to the interpolation points $x_1, \ldots, x_n$ with function values $f(x_1), \ldots, f(x_n)$.

3.2.2.2 Native Space

Based on Theorem 3.9, the linear function space $\mathcal{A}_\phi(\mathcal{D})$ can be further extended to a larger function space $\mathcal{N}_\phi(\mathcal{D})$, such that the optimal interpolant $s^* \in \mathcal{A}_\phi(\mathcal{D})$ to given data minimises the semi-norm $\|s\|_\phi$ over all $s \in \mathcal{N}_\phi(\mathcal{D})$ satisfying the same interpolation conditions. For any function $f \in \mathcal{N}_\phi(\mathcal{D})$, the optimal interpolant s^* to f at any $\mathcal{P}_m^d$-unisolvent set of points then has a semi-norm which is bounded above by $\|f\|_\phi$. Consequently, a function lies in $\mathcal{N}_\phi(\mathcal{D})$ if and only if all its interpolants have a uniformly bounded semi-norm, which provides the following definition, cf. Gutmann [2001b], Definition 3.10.

Definition 3.10. *Let ϕ be a conditionally positive definite radial basis function of order m, and $\mathcal{D} \subset \mathbb{R}^d$. Then, a function $f : \mathcal{D} \to \mathbb{R}$ belongs to the native space $\mathcal{N}_\phi(\mathcal{D})$ if and only if for any $\mathcal{P}_m^d$-unisolvent set $\{x_1, \ldots, x_n\} \subset \mathcal{D}$ the optimal interpolant $s^* \in \mathcal{A}_\phi(\mathcal{D})$ to f at these points satisfies*

$$\|s^*\|_\phi \leq C_f,$$

where C_f is a nonnegative constant that only depends on f.

Since, by Theorem 3.9, the semi-norms of all optimal interpolants to any $s \in \mathcal{A}_\phi(\mathcal{D})$ are bounded above by $\|s\|_\phi$, the linear space $\mathcal{A}_\phi(\mathcal{D})$ clearly is a subspace of $\mathcal{N}_\phi(\mathcal{D})$. Using Definition 3.10, it can further be shown that the native space $\mathcal{N}_\phi(\mathcal{D})$ constitutes the natural completion of the semi-inner product space $\mathcal{A}_\phi(\mathcal{D})$ with respect to $\langle \cdot, \cdot \rangle_\phi$, see Gutmann [2001b], Proposition 3.12 and Theorem 3.16. This is also the formal definition provided by Schaback [1999], Definition 5.2, in his discussion of native spaces. In particular, by letting $f \in \mathcal{N}_\phi(\mathcal{D})$ and defining

$$C_f := \sup \Big\{ \|s^*\|_\phi : s^* \in \mathcal{A}_\phi(\mathcal{D}) \text{ optimal interpolant to } (x_1, f(x_1)), \ldots, (x_n, f(x_n)) \\ \text{for the } \mathcal{P}_m^d\text{-unisolvent set } \{x_1, \ldots, x_n\} \subset \mathcal{D},\ n \in \mathbb{N} \Big\}, \tag{3.20}$$

the space $\mathcal{N}_\phi(\mathcal{D})$ may be equipped with the semi-inner product

$$\langle f, \tilde{f} \rangle_{\mathcal{N}_\phi} := \lim_{n\to\infty} \langle s_n^*, \tilde{s}_n^* \rangle_\phi, \qquad f, \tilde{f} \in \mathcal{N}_\phi(\mathcal{D}), \tag{3.21}$$

where $\{s_n^*\}$ and $\{\tilde{s}_n^*\}$ are Cauchy sequences of optimal interpolants from $\mathcal{A}_\phi(\mathcal{D})$ to f and $\tilde{f}$, respectively, with $C_f = \lim_{n\to\infty} \|s_n^*\|_\phi$ and $C_{\tilde{f}} = \lim_{n\to\infty} \|\tilde{s}_n^*\|_\phi$.[17] The semi-inner product defined in (3.21) coincides with $\langle \cdot, \cdot \rangle_\phi$ on $\mathcal{A}_\phi(\mathcal{D})$, and its null space is $\mathcal{P}_m^d$. It induces the semi-norm $\|\cdot\|_{\mathcal{N}_\phi} := \langle \cdot, \cdot \rangle_{\mathcal{N}_\phi}^{1/2}$, which is thus given for any $f \in \mathcal{N}_\phi(\mathcal{D})$ by the constant C_f.

Latter derivations allow to generalise Theorem 3.9 to the native space $\mathcal{N}_\phi(\mathcal{D})$, by equally replacing the term $\mathcal{A}_\phi(\mathcal{D})$ and the corresponding norm $\|\cdot\|_\phi$ in the statement with the respective expressions $\mathcal{N}_\phi(\mathcal{D})$ and $\|\cdot\|_{\mathcal{N}_\phi}$, cf. Gutmann [2001b], Theorem 3.17. This suggest that the native space $\mathcal{N}_\phi(\mathcal{D})$ of a radial basis function ϕ may also be interpreted as the largest function space for which the variational principle holds.

Native spaces of radial basis functions play an important role in the derivation of error estimates, e.g., Wendland [2005b], Chapter 11, but also in the convergence analysis of Gutmann's RBF method, see Section 3.3. In particular, they provide useful information on the smoothness of the functions they contain. To determine whether a function belongs to the native space of a radial basis function, however, the above definitions are rather abstract and do not give further insight. Nevertheless, native spaces may also be characterised by generalised Fourier transforms (e.g., Gutmann [2001b], Chapter 6), which then allow to derive

[17] Note that the sequences $\{\|s_n^*\|_\phi\}$ and $\{\|\tilde{s}_n^*\|_\phi\}$ are monotonically increasing, such that their limits are consistent with (3.20).

sufficient smoothness conditions for a function to lie in the native space. Loosely speaking, one may say that a function belongs to a native space $\mathcal{N}_\phi(\mathcal{D})$ if its Fourier transform decays fast enough relative to the Fourier transform of the basis function ϕ on $\mathcal{D}$. By means of Fourier transforms, the native spaces of surface splines can be shown to be Beppo Levi spaces, whereas those of (inverse) multiquadrics and Gaussians do not coincide with any of the classical function spaces and are rather small, essentially comprising only very smooth functions. More theoretical details on the derivation and the particular spaces can be found, for instance, in Wendland [2005b], Section 10.5.

For the convergence analysis of Gutmann's original RBF method and our subsequent modification and extension, we require the following theorem giving conditions for a function to be in the native space of a radial basis function ϕ, cf. Gutmann [2001b], Theorem 3.19. Note, however, that the assertion only holds for surface spline type radial basis functions. An extension to (inverse) multiquadric and Gaussian radial basis functions is not possible since the native spaces of such basis functions do not contain any nonzero functions with compact support, see Gutmann [2001b], Section 6.4.

Theorem 3.11. *Let ϕ be a conditionally positive definite surface spline of order m from Table 3.1, and let*

$$\nu_d = \begin{cases} (d+\nu+1)/2 & \text{if } d+\nu \text{ is odd,} \\ (d+\nu)/2 & \text{if } d+\nu \text{ is even.} \end{cases}$$

If $f \in C^{\nu_d}(\mathcal{D})$, where (i) $\mathcal{D} \subset \mathbb{R}^d$ is compact, or (ii) $\mathcal{D} = \mathbb{R}^d$ and f has compact support, then $f \in \mathcal{N}_\phi(\mathcal{D})$.

Proof. See the proof of Gutmann [2001b], Theorem 3.19. □

3.3 Gutmann's Radial Basis Function Method

Gutmann's RBF method [2001a; 2001b] belongs to the class of response surface methods that aim at minimising a nonconvex deterministic objective function which is expensive to evaluate. It is based on the general surface technique proposed by Jones [1996], see Subsection 3.1.2 for a brief description, where Gutmann's main idea is to employ radial basis functions as means to construct response surface interpolants. As seen in Section 3.2, the use of radial basis functions not only ensures the uniqueness of interpolants under relatively mild conditions on the location of the sample points, but also provides in a natural way a measure of 'bumpiness' in form of the semi-norm $\|\cdot\|_\phi$. In particular, by use of the latter, the related expression for finding a new evaluation can be simplified accordingly, which in turn allows to establish convergence of the method for certain radial basis functions under reasonable assumptions on the objective function and the choice of target values f_n^*.

In what follows, we review Gutmann's original RBF method in more detail as it serves as a main references for our subsequent developments in later sections. Specifically, by following Gutmann [2001b], Chapter 4, we start with a general description of the method, then state the most important convergence results in view of our developments, and eventually address some relevant practical aspects regarding the implementation of the method.

3.3.1 Description of Method

To outline the RBF method for minimising a continuous objective function $f : \mathcal{X} \to \mathbb{R}$ on a compact set $\mathcal{X} \subset \mathbb{R}^d$, let ϕ be a conditionally positive definite radial basis function of order m, and $\{p_j\}_{j=1}^{\widetilde{m}}$ be a basis of the polynomial space $\mathcal{P}_m^d$, with $\widetilde{m} = \dim(\mathcal{P}_m^d)$. Further, assume that the method has been initialised either manually or by some experimental design, yielding a $\mathcal{P}_m^d$-unisolvent set of points $\{x_1, \dots, x_{n_0}\} \subset \mathcal{X}$ at which the values of f are known exactly. Setting $n = n_0$, a general iteration of the method, consisting of the construction of an interpolant and the determination of a new evaluation point by a suitably chosen target value, can then be described as follows, cf. also the overview at the end of this subsection.

3.3.1.1 Construction of Response Surface

Since the points $x_1, \dots, x_n$ are assumed to form a $\mathcal{P}_m^d$-unisolvent set, Theorem 3.7 implies the existence of a unique function $s_n \in \mathcal{A}_\phi(\mathcal{X})$ of the form

$$s_n(x) = \sum_{i=1}^{n} \lambda_i \phi(\|x - x_i\|_2) + p(x), \quad x \in \mathbb{R}^d, \tag{3.22}$$

interpolating the data $(x_1, f(x_1)), \dots, (x_n, f(x_n))$. Note that s_n can be equivalently formulated as the unique solution of the variation principle, minimising the semi-norm among all functions in $\mathcal{A}_\phi(\mathcal{X})$ subject to the same interpolation conditions, see Theorem 3.9.

3.3.1.2 Determination of Next Evaluation Point

Upon the construction of the interpolant s_n, the next point x_{n+1} at which the objective function f is evaluated has to be determined. To this end, Gutmann adopts Jones's general technique and assumes that a suitable target value f_n^* has been chosen, such that the new point x_{n+1} may be defined as the value of $y \in \mathcal{X} \backslash \{x_1, \dots, x_n\}$ which minimises the semi-norm $\|s_y\|_\phi$ of the augmented interpolant $s_y \in \mathcal{A}_\phi(\mathcal{X})$ through the previous data $(x_1, f(x_1)), \dots, (x_n, f(x_n))$ and the additional point (y, f_n^*).

Tackling this problem, it is convenient to work with the Lagrange representation and rewrite the optimal interpolant $s_y \in \mathcal{A}_\phi(\mathcal{X})$, $y \in \mathcal{X} \backslash \{x_1, \dots, x_n\}$, satisfying the interpolation conditions

$$\begin{aligned} s_y(x_i) &= f(x_i), \qquad i = 1, \dots, n, \\ s_y(y) &= f_n^*, \end{aligned}$$

in form of

$$s_y(x) = s_n(x) + \left[f_n^* - s_n(y)\right] l_n(y, x), \quad x \in \mathbb{R}^d, \tag{3.23}$$

where $l_n(y, \cdot\,) \in \mathcal{A}_\phi(\mathcal{X})$ is the optimal interpolant to

$$\begin{aligned} l_n(y, x_i) &= 0, \qquad i = 1, \dots, n, \\ l_n(y, y) &= 1. \end{aligned} \tag{3.24}$$

Representation (3.23) is feasible since the right-hand side is a linear combination of functions from $\mathcal{A}_\phi(\mathcal{X})$, thus belonging to the same space and interpolating the same function values as s_y. Since s_y, however, is the unique element with theses properties, the equality in (3.23) must hold.

Now to make use of representation (3.23), first observe that the Lagrange basis function $l_n(y, \cdot\,)$ can be expressed as

$$l_n(y, x) = \sum_{i=1}^{n} \alpha_i(y)\phi(\|x - x_i\|_2) + \beta(y)\phi(\|x - y\|_2) + \sum_{j=1}^{\widetilde{m}} b_j(y)p_j(x), \quad x \in \mathbb{R}^d, \tag{3.25}$$

where the coefficients $\alpha(y) = (\alpha_1(y), \dots, \alpha_n(y))^\top \in \mathbb{R}^n$, $\beta(y) \in \mathbb{R}$ and $b(y) = (b_1(y), \dots, b_{\widetilde{m}}(y))^\top \in \mathbb{R}^{\widetilde{m}}$ solve the linear system

$$\begin{pmatrix} \Phi & m_n(y) & P \\ m_n(y)^\top & \phi(0) & \pi(y)^\top \\ P^\top & \pi(y) & 0 \end{pmatrix} \begin{pmatrix} \alpha(y) \\ \beta(y) \\ b(y) \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \tag{3.26}$$

for the matrices $\Phi \in \mathbb{R}^{n\times n}$ and $P \in \mathbb{R}^{n\times\widetilde{m}}$, and the vectors $m_n(y) := (\phi(\|x_1 - y\|_2), \dots, \phi(\|x_n - y\|_2))^\top \in \mathbb{R}^n$ and $\pi(y) := (p_1(y), \dots, p_{\widetilde{m}}(y))^\top \in \mathbb{R}^{\widetilde{m}}$. By means of (3.23), the squared semi-norm of s_y can then be simplified to

$$\begin{aligned} \|s_y\|_\phi^2 &= \|s_n\|_\phi^2 + 2\big[f_n^* - s_n(y)\big]\langle s_n, l_n(y, \cdot\,)\rangle_\phi + \big[f_n^* - s_n(y)\big]^2 \|l_n(y, \cdot\,)\|_\phi^2 \\ &= \|s_n\|_\phi^2 + \beta(y)\big[f_n^* - s_n(y)\big]^2, \end{aligned} \tag{3.27}$$

as the definition of the semi-inner product (3.17) and the interpolation conditions (3.24) imply that

$$\langle s_n, l_n(y, \cdot\,)\rangle_\phi = \sum_{i=1}^{n} \lambda_i l_n(y, x_i) = 0,$$

and

$$\|l_n(y, \cdot\,)\|_\phi^2 = \sum_{i=1}^{n} \alpha_i(y) l_n(y, x_i) + \beta(y) l_n(y, y) = \beta(y).$$

Since $\|s_n\|_\phi$ is independent of y, however, equation (3.27) shows that the required minimisation of $\|s_y\|_\phi$ with respect to y boils down to minimising the nonnegative utility function

$$g_n(y) := \mu_n(y)\big[f_n^* - s_n(y)\big]^2, \qquad y \in \mathcal{X}\backslash\{x_1, \dots, x_n\}, \tag{3.28}$$

where the function $\mu_n : \mathcal{X}\backslash\{x_1, \dots, x_n\} \to \mathbb{R}$ is given by

$$\mu_n(y) := \|l_n(y, \cdot)\|_\phi^2 = \beta(y). \tag{3.29}$$

Note that μ_n is well-defined since $l_n(y, \cdot)$ is well-defined for any $y \in \mathcal{X}\backslash\{x_1, \dots, x_n\}$. Moreover, we can make the following observations on its behaviour.

Remark 3.12. *Definition* (3.29) *provides* $\mu_n(y) > 0$ *for* $y \in \mathcal{X}\backslash\{x_1, \dots, x_n\}$*, which can be seen by contradiction as follows. Assuming there is an* $y_0 \in \mathcal{X}\backslash\{x_1, \dots, x_n\}$ *with* $\mu_n(y_0) = 0$*, definition* (3.29) *implies that* $l_n(y_0, \cdot) \in \mathcal{P}_m^d$*. By the first n equations of* (3.24) *and the* $\mathcal{P}_m^d$*-unisolvency, this further entails* $l_n(y_0, \cdot) \equiv 0$*, which, however, contradicts the last equation of* (3.24)*.*

Further, by applying Cramer's rule to the linear system (3.26)*, the function* μ_n *can be calculated as*

$$\mu_n(y) = \frac{\det A_n}{\det A_n(y)}, \qquad y \in \mathcal{X}\backslash\{x_1, \dots, x_n\},$$

where A_n *and* $A_n(y)$ *are given by the nonsingular interpolation matrices on the left-hand sides of equations* (3.11) *and* (3.26)*, respectively, i.e.*

$$A_n := \begin{pmatrix} \Phi & P \\ P^\top & 0 \end{pmatrix} \qquad \text{and} \qquad A_n(y) := \begin{pmatrix} \Phi & m_n(y) & P \\ m_n(y)^\top & \phi(0) & \pi(y)^\top \\ P^\top & \pi(y) & 0 \end{pmatrix}. \tag{3.30}$$

Hence, since $\det A_n$ *is a nonzero constant and* $\lim_{y\to x_i} \det A_n(y) = 0$ *for any* $i \in \{1, \dots, n\}$*, it follows that*

$$\lim_{y\to x_i} \mu_n(y) = \infty, \qquad i = 1, \dots, n. \tag{3.31}$$

3.3.1.3 Choice of Target Value

The choice of the target value f_n^* crucially influences the location of the new point x_{n+1}, minimising g_n on $\mathcal{X}\backslash\{x_1, \dots, x_n\}$. If $f_n^* \in [\min_{y\in\mathcal{X}} s_n(y), \max_{y\in\mathcal{X}} s_n(y)]$, then $g_n(y) = 0$ is attained at every point $y \neq \{x_1, \dots, x_n\}$ for which $s_n(y) = f_n^*$, such that a global minimiser of g_n on $\mathcal{X}\backslash\{x_1, \dots, x_n\}$ need not exist (e.g. if f_n^* is equal to an observed function value). However, if $f_n^* < \min_{y\in\mathcal{X}} s_n(y)$, property (3.31) shows that $g_n(y)$ tends to infinity as y approaches any x_i, $i = 1, \dots, n$, implying that a global minimiser of g_n over $\mathcal{X}$ exists and is away from already sampled points. Hence, to guarantee a well-defined point x_{n+1}, it must hold that

$$f_n^* \in \big[-\infty, \min_{y\in\mathcal{X}} s_n(y)\big], \tag{3.32}$$

where the case $f_n^* = \min_{y\in\mathcal{X}} s_n(y)$ is only admissible if none of the x_i is a global minimiser of s_n, i.e. if $f_n^* < s_n(x_i)$, $i = 1, \dots, n$.

For admissible choices satisfying (3.32), the minimisation of g_n on $\mathcal{X}\backslash\{x_1, \dots, x_n\}$ is essentially guided by the dual goals of globally exploring fairly unvisited regions of the parameter space and locally exploiting promising regions of attraction. While for a low target value f_n^* the method performs global search in which the new point x_{n+1} is sampled away from already evaluated points, a high target value close or equal to $\min_{y\in\mathcal{X}} s_n(y)$ is supposed to sample x_{n+1} either in the vicinity of a global minimiser of s_n if $f_n^* < \min_{y\in\mathcal{X}} s_n(y)$ or as a global minimiser of s_n if $f_n^* = \min_{y\in\mathcal{X}} s_n(y)$, cf. Regis and Shoemaker [2007b]. Moreover, for $f_n^* \to \infty$, one may derive the following special case.

Remark 3.13. *For $f_n^* < \min_{y\in\mathcal{X}} s_n(y)$, the next point x_{n+1} satisfies as a global minimiser of g_n on $\mathcal{X}\backslash\{x_1,\dots,x_n\}$ the inequality*

$$\mu_n(x_{n+1}) \leq \mu_n(y)\left[1+\frac{s_n(x_{n+1})-s_n(y)}{f_n^*-s_n(x_{n+1})}\right]^2, \qquad y\in\mathcal{X}\backslash\{x_1,\dots,x_n\}.$$

Hence, as $f_n^* \to -\infty$, the boundedness of s_n on $\mathcal{X}$ implies

$$\mu_n(x_{n+1}) \leq \mu_n(y), \qquad y\in\mathcal{X}\backslash\{x_1,\dots,x_n\}.$$

Consequently, the choice $f_n^* = -\infty$ reduces the minimisation of g_n on $\mathcal{X}\backslash\{x_1,\dots,x_n\}$ to

$$\min_{y\in\mathcal{X}\backslash\{x_1,\dots,x_n\}} \mu_n(y),$$

which samples x_{n+1} as far away as possible from the points $x_1,\dots,x_n$ and thus leads to global search of the RBF method in the most unexplored regions of the parameter space.

In practice, f_n^* is typically set in cycles by some predefined strategy, where each cycle consists of a range of target values starting with a very low value to invoke global search and ending with a value equal to $\min_{y\in\mathcal{X}} s_n(y)$ to perform a purely local search, see Subsection 3.3.3 for a more detailed specification.

3.3.1.4 Algorithm

Altogether, Gutmann's original RBF method for minimising a deterministic and continuous function $f:\mathcal{X}\to\mathbb{R}$ on a compact set $\mathcal{X}$ can be summarised by the following basic algorithm.

Algorithm 3.14. *(RBF Method).*

```
0. Initial step:
    - Choose a conditionally positive definite radial basis function $\phi$ of order $m$.
    - Generate a $\mathcal{P}_m^d$-unisolvent set of points $\{x_1,\dots,x_{n_0}\}\subset\mathcal{X}$.
    - Evaluate $f=\sum_{k=1}^{l} g(f_k)$ at the points $x_1,\dots,x_{n_0}$, and set $n=n_0$.
1. Iteration step:
    while $n \leq n^{\max}$ do
```

```
        - Construct the optimal interpolant $s_n \in \mathcal{A}_\phi(\mathcal{X})$ solving
              $$\min_{s \in \mathcal{A}_\phi(\mathcal{X})} \|s\|_\phi \qquad s.t. \quad s(x_i) = f(x_i), \quad i = 1, \dots, n.$$
        - Choose an admissible target value $f_n^* \in \big[-\infty, \min_{y \in \mathcal{X}} s_n(y)\big]$.
        - Determine $x_{n+1}$, which is the value of $y$ that solves
              $$\min_{y \in \mathcal{X} \setminus \{x_1, \dots, x_n\}} \mu_n(y) \big[f_n^* - s_n(y)\big]^2.$$
        - Evaluate $f$ at $x_{n+1}$, and set $n = n + 1$.
    end while
```

3.3.2 Convergence of Method

Under particular assumptions on the objective function, the choice of radial basis function ϕ and the sequence of employed target values $\{f_n^*\}$, it is possible to establish convergence of Gutmann's RBF method for any continuous function f. Since the method is a deterministic global optimisation algorithm which sequentially samples the objective function by only using the information collected during its execution, i.e. a sequential sampling algorithm, convergence to the global minimum may be defined for these kind of algorithms according to Törn and Žilinskas [1989], Section 1.2.1.2.

Definition 3.15. *A sequential sampling algorithm for $f : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R}$ with sequence of iterates $\{x_n\}_{n \in \mathbb{N}}$ is said to converge to the global minimum of f (assuming it exists) if*

$$\min_{1 \le i \le n} f(x_i) \to \min_{x \in \mathcal{X}} f(x), \qquad \textit{as } n \to \infty.$$

A sufficient and necessary condition for the convergence of a deterministic sequential sampling algorithm to the global minimum of any continuous function is given by Törn and Žilinskas [1989], Theorem 1.3, stating that the generated sequence of iterates should be dense in a compact set. Applied to Gutmann's RBF method, this theorem can be reformulated as follows, cf. Gutmann [2001b], Theorem 4.4.

Theorem 3.16. *Algorithm 3.14 converges for every continuous function f if and only if it generates a sequence of points $\{x_n\}$ that is dense in $\mathcal{X}$.*

To facilitate the presentation of the convergence of the method, we first state the main theorem along with two particular convergence results, and give the related proof thereafter.

3.3.2.1 Convergence Results

Unfortunately, it has only been possible so far to show convergence of the method in the case of spline type radial basis functions, cf. Table 3.1, due to the dependence of one of the required lemmas on Theorem 3.11. It is not known whether a similar result can also be derived for other radial basis functions.

Moreover, the convergence of the RBF method does not allow a free choice of the target values f_n^*, as they have to be set sufficiently low in enough iterations to enforce a global search and reach the global minimum of f. To be more specific, convergence of the method is achieved if, for infinitely many $n \in \mathbb{N}$, the condition

$$f_n^* < \min_{y \in \mathcal{X}} \left[s_n(y) - \tau \|s_n\|_\infty \Delta_n^{\rho/2}(y) \right] \tag{3.33}$$

is satisfied, where $\tau > 0$ is a constant, $\|\cdot\|_\infty$ denotes the supremum norm of a function on $\mathcal{X}$ and the minimum distance function Δ_n is given by

$$\Delta_n(y) := \min_{1 \leq i \leq n} \|y - x_i\|_2, \qquad y \in \mathcal{X}. \tag{3.34}$$

The choice of the remaining constant $\rho \geq 0$ depends on the specification of the radial basis function ϕ, where $\rho < 1$, for $\phi(r) = r$, and $\rho < 2$, otherwise.

Given these preliminary insights, the main theorem for establishing the convergence of the RBF method is formulated as follows, cf. Gutmann [2001b], Theorem 4.5.

Theorem 3.17. *Let ϕ be a conditionally positive definite surface spline of order m from Table 3.1, and let $\{x_n\}$ be the sequence of iterates generated by Algorithm 3.14. Further, let s_n be the optimal interpolant from $\mathcal{A}_\phi(\mathcal{X})$ to the data $(x_i, f(x_i))$, $i = 1, \ldots, n$. Assume that, for infinitely many $n \in \mathbb{N}$, the choice of f_n^* satisfies* (3.33)*, where τ, Δ_n and ρ are given as above. Then, the sequence $\{x_n\}$ is dense in $\mathcal{X}$.*

In view of assumption (3.33), two particular convergence results can be obtained immediately from Theorems 3.16 and 3.17. The first one follows from observing that the right-hand side of (3.33) is finite for each n, cf. Gutmann [2001b], Corollary 4.6.

Corollary 3.18. *Let ϕ and m be as in Theorem 3.17. Further, let f be continuous, and assumes that, for infinitely many $n \in \mathbb{N}$, it holds $f_n^* = -\infty$. Then, Algorithm 3.14 converges.*

A further convergence result applies if $\{\|s_n\|_\infty\}_{n \in \mathbb{N}}$ is uniformly bounded, in which case the right-hand side of condition (3.33) can be replaced by $\min_{y \in \mathcal{X}} [s_n(y) - \tau \Delta_n^{\rho/2}(y)]$. This constraint is much easier to check in an implementation than (3.33). The uniform boundedness of $\{\|s_n\|_\infty\}$, in turn, can be guaranteed if the function f belongs to the native space of the radial basis function ϕ, as shown by the next lemma, cf. Gutmann [2001b], Lemma 4.7.

Lemma 3.19. *Let $\{x_n\}$ be a sequence in $\mathcal{X}$ with pairwise different points such that $\{x_1, \ldots, x_{n_0}\}$ is $\mathcal{P}_m^d$-unisolvent. For $n \geq n_0$, let s_n denote the optimal interpolant to f at $x_1, \ldots, x_n$, and let $f \in \mathcal{N}_\phi(\mathcal{X})$. Then, $\|s_n\|_\infty$ is bounded above by a number that depends only on $x_1, \ldots, x_{n_0}$ and on f. More specifically, it holds*

$$|s_n(y)| \leq \frac{1}{\sqrt{\alpha_1}} \|f\|_{\mathcal{N}_\phi} + \|f\|_\infty, \qquad y \in \mathcal{X} \backslash \{x_1, \ldots, x_n\},$$

where $\alpha_1 > 0$ is a constant depending on $x_1, \ldots, x_{n_0}$, and $|s_n(y)| \leq \|f\|_\infty$, $y \in \{x_1, \ldots, x_n\}$.

In the special case of surface spline functions ϕ, Theorem 3.11 gives a sufficient condition on the smoothness of a function f such that it lies in the native space $\mathcal{N}_\phi(\mathcal{X})$. Hence, the preceding lemma together with Theorems 3.16 and 3.17 allow to conclude the following result, cf. Gutmann [2001b], Corollary 4.8.

Corollary 3.20. *Let ϕ and m be as in Theorem 3.17. Further, let ν_d be as in Theorem 3.11, and let $f \in C^{\nu_d}(\mathcal{X})$. Assume that, for infinitely many $n \in \mathbb{N}$, it holds*

$$f_n^* < \min_{y \in \mathcal{X}} \left[s_n(y) - \tau \Delta_n^{\rho/2}(y) \right],$$

where τ, Δ_n and ρ are given as above. Then, Algorithm 3.14 converges.

3.3.2.2 Proof of Convergence

The proof of Theorem 3.17 is rather technical and requires the use of the following three lemmas on the behaviour of the functions μ_n, $n \in \mathbb{N}$. In particular, the first two lemmas, corresponding to Lemmas 4.9 and 4.10 in Gutmann [2001b], respectively, address the limit behaviour of $\{\mu_n(x_n)\}$ for the generated sequence $\{x_n\}$ and apply to any radial basis function of Table 3.1, where Lemma 3.21 is used to show the statement in Lemma 3.22. In contrast, the last Lemma 3.23 deals with $\{\mu_n(y)\}$ for some point $y \in \mathcal{X}$ not belonging to $\{x_n\}$ and only applies to surface spline type radial basis functions.

Lemma 3.21. *Let ϕ be a conditionally positive definite radial basis function of order m from Table 3.1, and let $\{z_1, \ldots, z_k\}$ be a $\mathcal{P}_m^d$-unisolvent set in a compact set $\mathcal{X} \subset \mathbb{R}^d$. Let $\{x_n\}$ and $\{y_n\}$ be convergent sequences in $\mathcal{X}$ that have the same limit $x^* \notin \{z_1, \ldots, z_k\}$ and satisfy $x_n \neq y_n$, $n \in \mathbb{N}$. Further, let $\tilde{l}_n(x_n, \cdot)$ be the optimal interpolant to the data $(z_1, 0), \ldots, (z_k, 0), (y_n, 0)$ and $(x_n, 1)$. Then,*

$$\lim_{n \to \infty} \|y_n - x_n\|_2^\rho \, \tilde{\mu}_n(x_n) = \infty,$$

where $\tilde{\mu}_n$ is the function given by (3.29) for the interpolant $\tilde{l}_n(x_n, \cdot)$, and where $0 \leq \rho < 1$, for $\phi(r) = r$, and $0 \leq \rho < 2$, otherwise.

Lemma 3.22. *Let ϕ and m be chosen as in Lemma 3.21, where ρ takes a value as indicated. Let $\{x_n\}$ be a sequence in $\mathcal{X}$ with pairwise different points such that $\{x_1, \ldots, x_{n_0}\}$ is $\mathcal{P}_m^d$-unisolvent. Then, for every convergent subsequence $\{x_{n_k}\}_{k \in \mathbb{N}}$ of $\{x_n\}$, it holds*

$$\lim_{k \to \infty} \Delta_{n_k - 1}^\rho(x_{n_k}) \, \mu_{n_k - 1}(x_{n_k}) = \infty,$$

where μ_{n_k-1} and Δ_{n_k-1} are the functions given by (3.29) and (3.34), respectively, for $n = n_k - 1$.

By using Theorem 3.11 on the smoothness of surface splines, the next lemma shows that $\{\mu_n(y)\}$ is uniformly bounded for any point y bounded away from the sequence $\{x_n\}$, see Lemma 4.11 in Gutmann [2001b]. As already remarked, there has not been any extension of this result to other radial basis function since then, to the best of our knowledge.

Lemma 3.23. *Let ϕ be a conditionally positive definite surface spline of order m from Table 3.1, and let $\{x_n\}$ be a sequence in $\mathbb{R}^d$ with pairwise different points such that $\{x_1, \ldots, x_{n_0}\}$ is $\mathcal{P}_m^d$-unisolvent. Further, let $y_0 \in \mathbb{R}^d$ satisfy $\|y_0 - x_n\|_2 \geq \delta$, $n \in \mathbb{N}$, for some $\delta > 0$. Then, there exists $\widetilde{C} > 0$, depending only on y_0 and δ, such that*

$$\mu_n(y_0) \leq \widetilde{C}, \qquad \forall\, n \geq n_0,$$

where μ_n is the function given by (3.29).

Eventually, using Lemmas 3.21 – 3.23, the proof of Theorem 3.17, establishing that Algorithm 3.14 generates a sequence $\{x_n\}$ which is dense in $\mathcal{X}$, is given as follows.

Proof of Theorem 3.17. We argue by contradiction, and assume that there exist $y_0 \in \mathcal{X}$ and $\delta > 0$ such that $B_\delta(y_0) = \{x \in \mathbb{R}^d : \|x - y_0\|_2 < \delta\}$ does not contain any x_n, $n \in \mathbb{N}$. The iteration step of Algorithm 3.14 then yields

$$g_n(x_{n+1}) \leq g_n(y_0), \qquad n \geq n_0,$$

where n_0 is the number of points chosen in the initial step of the algorithm. Moreover, since assumption (3.33) holds for infinitely many $n \in \mathbb{N}$, there exists a subsequence $\{n_k\}_{k\in\mathbb{N}}$, $n_k \in \mathbb{N}$, such that

$$f^*_{n_k-1} < \min_{y\in\mathcal{X}} \left[s_{n_k-1}(y) - \tau\|s_{n_k-1}\|_\infty \Delta^{\rho/2}_{n_k-1}(y)\right], \qquad k \in \mathbb{N}, \tag{3.35}$$

where $\tau > 0$, Δ_{n_k-1} is given by (3.34) for $n = n_k-1$, and ρ is a constant satisfying $0 \leq \rho < 1$, for $\phi(r) = r$, and $0 \leq \rho < 2$, for the other surface splines. Applied to the sequence $\{x_n\}$, condition (3.35) yields

$$s_{n_k-1}(x_{n_k}) - f^*_{n_k-1} > \tau\|s_{n_k-1}\|_\infty \Delta^{\rho/2}_{n_k-1}(x_{n_k}), \qquad k \in \mathbb{N}. \tag{3.36}$$

Also, since $\{x_{n_k}\}$ is a sequence in a compact set, it contains a convergent subsequence, where we assume that $\{x_{n_k}\}$ itself converges, without loss of generality.

Now, for all $k \in \mathbb{N}$, x_{n_k} is a minimiser of $g_{n_k-1}(y)$, $y \in \mathcal{X}\backslash\{x_1, \ldots, x_{n_k-1}\}$. Hence, if $f^*_{n_k-1} > -\infty$, we have

$$\mu_{n_k-1}(x_{n_k})\left[f^*_{n_k-1} - s_{n_k-1}(x_{n_k})\right]^2 \leq \mu_{n_k-1}(y_0)\left[f^*_{n_k-1} - s_{n_k-1}(y_0)\right]^2. \tag{3.37}$$

Given $\|s_{n_k-1}\|_\infty > 0$, the latter inequality (3.37) implies

$$\begin{aligned}
\mu_{n_k-1}(x_{n_k}) &\leq \mu_{n_k-1}(y_0)\left[\frac{s_{n_k-1}(y_0) - f^*_{n_k-1}}{s_{n_k-1}(x_{n_k}) - f^*_{n_k-1}}\right]^2 \\
&\leq \mu_{n_k-1}(y_0)\left[1 + \frac{|s_{n_k-1}(y_0) - s_{n_k-1}(x_{n_k})|}{s_{n_k-1}(x_{n_k}) - f^*_{n_k-1}}\right]^2 \\
&\leq \mu_{n_k-1}(y_0)\left[1 + \frac{|s_{n_k-1}(y_0) - s_{n_k-1}(x_{n_k})|}{\tau\Delta^{\rho/2}_{n_k-1}(x_{n_k})\|s_{n_k-1}\|_\infty}\right]^2 \\
&\leq \mu_{n_k-1}(y_0)\left[1 + \frac{2}{\tau\Delta^{\rho/2}_{n_k-1}(x_{n_k})}\right]^2, \qquad (3.38)
\end{aligned}$$

where condition (3.36) and the definition of $\|\cdot\|_\infty$ are used in the third and last line, respectively. In case $\|s_{n_k-1}\|_\infty = 0$, the term $[f^*_{n_k-1} - s_{n_k-1}(y)]$ becomes a constant such that inequality (3.38) can be derived straightforwardly from (3.37). This inequality also holds for $f^*_{n_k-1} = -\infty$, due to Remark 3.13.

Eventually, multiplying both sides of inequality (3.38) by $\Delta^\rho_{n_k-1}(x_{n_k})$ yields

$$\mu_{n_k-1}(x_{n_k})\Delta^\rho_{n_k-1}(x_{n_k}) \leq \mu_{n_k-1}(y_0)\left[\Delta^{\rho/2}_{n_k-1}(x_{n_k}) + \frac{2}{\tau}\right]^2, \qquad (3.39)$$

for which contradiction is shown by means of Lemmas 3.21 – 3.23. Specifically, by using Lemma 3.21, Lemma 3.22 shows that the left-hand side of inequality (3.39) tends to infinity for $k \to \infty$. However, since Lemma 3.23 states that $\mu_n(y_0)$ is bounded above by some constant independent of n and $\{\Delta_{n_k-1}(x_{n_k})\}$ is uniformly bounded on the compact set $\mathcal{X}$, it follows that the right-hand side is bounded above by some constant independent of k, contradicting (3.39). Consequently, $B_\delta(y_0)$ contains a point x_n, $n \in \mathbb{N}$, implying that $\{x_n\}$ is dense in $\mathcal{X}$. □

3.3.3 Practical Issues

In this subsection, we summarise relevant practical issues regarding the implementation of the RBF method in Algorithm 3.14, as suggested by Gutmann [2001a,b] and other related authors. This is absolutely indispensable to any user since the method provides plenty of flexibility, starting from the choice of a suitable radial basis function to dealing with numerical issues, where the performance of a generic implementation is typically not competitive enough for specific applications.

3.3.3.1 Initialisation

Initialising Algorithm 3.14, a conditionally positive definite radial basis function ϕ needs to be selected, which is usually one of the globally supported choices from Table 3.1. This has the advantage that the global behaviour of the objective function is better captured over the entire parameter space, leading to a more stable construction of response surface models along the iterations. In theory, a proper choice of ϕ depends on the smoothness of the objective function f to be approximated, as indicated in Subsection 3.2.2. Each ϕ is associated with a native space $\mathcal{N}_\phi(\mathcal{X})$ and ϕ should therefore be chosen such that $f \in \mathcal{N}_\phi(\mathcal{X})$. However, since typically little or no information about the smoothness of f is known, it is reasonable to employ rather smooth basis functions that prevent the surface models from becoming too oscillating. Choices that have practically turned out to perform well for most well-behaved objective functions are surface splines, especially $\phi(r) = r^2 \log r$ and $\phi(r) = r^3$, as remarked by Björkman and Holmström [2000] and Gutmann [2001b]. In particular, their native space contains a very large class of functions. Some more guidelines for the selection of appropriate radial basis functions are given in Schaback and Wendland [2006], albeit in a general context. For use within the RBF method, Costa and Nannicini [2014] present an automatic selection technique which assesses the model quality of different radial basis functions to pick the most appropriate one based on a cross-validation procedure.

Further flexibility can be added to the choice of radial basis function by adjusting the positive shape parameter ζ if available, e.g. Fasshauer [2007], Chapter 17, and the references therein, or by considering ϕ with a modified 2-norm $\|\cdot\|_{W_x} := \|W_x \cdot\|_2$, where W_x is a diagonal weight matrix, leading to the notion of so-called anisotropic radial basis functions, e.g. Fowkes [2011]. Yet, it is difficult to apply these approaches with a credible estimation technique to a constantly enlarging set of sample points when no additional knowledge on the objective function is given.

To initialise the RBF method, any strategy for generating a $\mathcal{P}_m^d$-unisolvent set of sample points may be used, see Section 3.1.1. In addition, Iske [2000] and De Marchi et al. [2005] propose techniques for constructing optimal point sets for interpolation by radial basis functions where the points are chosen to be uniformly well-distributed. In any case, since the RBF method is commonly used with box constraints, the initial points are usually chosen as the corners of the box, to which the midpoint may be added for further stability. Also, as suggested by Gutmann [2001b], including ‘good’ points may prove advantageous in that the method is directed towards promising regions of the parameter space. By using box constraints, it is highly recommended to transform the parameter space to the unit hypercube $[0, 1]^d$, giving all dimensions equal weight, see, for instance, Björkman and Holmström [2000]. The main optimisation procedure is then carried out on the normalised space and only refers to the original space for evaluation of the objective function.

3.3.3.2 Optimisation of Subproblems

Gutmann’s RBF method, as well as any other response surface method, has the unavoidable disadvantage that several subproblems have to be solved in each iteration. In particular, Algorithm 3.14 requires solving a linear system to build s_n, followed by either a minimisation of s_n or a minimisation of s_n and μ_n or g_n to find x_{n+1}, depending on the choice of f_n^*. Since considerable time is spent on these subproblems, it is well worth having procedures in place that are able to solve them efficiently.

Construction of Response Surface

To solve the interpolation system (3.11) for constructing the model s_n, Powell [1996] presents a factorisation procedure that makes use of the conditional positive definiteness of ϕ, i.e. the fact that $\lambda^\top \Phi \lambda > 0$ for $\lambda \in \mathbb{R}^n \backslash \{0\}$ with $P^\top \lambda = 0$. Considering the QR decomposition of the polynomial basis matrix P by

$$P = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix} \begin{pmatrix} R \\ 0 \end{pmatrix}, \tag{3.40}$$

it follows that the columns of Q_2 span the null space of $P^\top$, and any $\lambda \in \mathbb{R}^n \backslash \{0\}$ with $P^\top \lambda = 0$ can thus be expressed as $\lambda = Q_2 z$ for some vector z. This, in turn, implies $z^\top Q_2^\top \Phi Q_2 z = \lambda^\top \Phi \lambda > 0$, $z \in \mathbb{R}^{n-\tilde{m}} \backslash \{0\}$, such that the matrix $Q_2^\top \Phi Q_2$ is positive definite and admits a Cholesky factorisation $Q_2^\top \Phi Q_2 = LL^\top$. The factorisation, requiring $\mathcal{O}(n^3)$ operations, can then be used to solve the linear system for the coefficients λ and c by means of backward and forward substitution in at most $\mathcal{O}(n^2)$ operations.

Upon observing that the interpolation matrix in (3.11) is updated in each iteration by addition of a single row and column for a new point, Björkman and Holmström [2000] show that the computation of the Cholesky factorisation can be further improved. In particular, in their implementation of the method an iterative strategy to update the triangular matrix L is applied, which is based on an additional QR decomposition with a matrix of Givens rotations. In this way, the cost of the factorisation can be reduced to $\mathcal{O}(n^2)$, such that the total expense for solving the linear system (3.11) amounts to $\mathcal{O}(n^2)$ operations.

Determination of Next Evaluation Point

To determine x_{n+1}, a target value $f_n^* \in [-\infty, \min_{y \in \mathcal{X}} s_n(y)]$ needs to be first chosen. Therefore, an estimate of the global minimum of s_n is required, which may be obtained by minimising s_n on $\mathcal{X}$. Moreover, if $f_n^* = \min_{y \in \mathcal{X}} s_n(y)$ and this choice is admissible, then x_{n+1} is the global minimiser of s_n on $\mathcal{X}$. Otherwise, if $f_n^* < \min_{y \in \mathcal{X}} s_n(y)$, then a global minimiser of either μ_n or g_n on $\mathcal{X} \backslash \{x_1, \ldots, x_n\}$ provides the new point x_{n+1} in the cases $f_n^* = -\infty$ and $f_n^* > -\infty$, respectively. However, unlike s_n whose minimisation can be carried out on the entire domain $\mathcal{X}$, both functions μ_n and g_n are only well-defined on $\mathcal{X} \backslash \{x_1, \ldots, x_n\}$ as their values $\mu_n(y)$ and $g_n(y)$ tend to infinity as y approaches x_i, $i = 1, \ldots, n$. This may cause problems in the implementation of the method, to which end the following equivalence is suggested, cf. Gutmann [2001b], Proposition 4.12.

Proposition 3.24. *The function v_n defined by*

$$v_n(y) := \left[\phi(0) - \begin{pmatrix} m_n(y) \\ \pi(y) \end{pmatrix}^\top \begin{pmatrix} \Phi & P \\ P^\top & 0 \end{pmatrix}^{-1} \begin{pmatrix} m_n(y) \\ \pi(y) \end{pmatrix} \right], \qquad y \in \mathbb{R}^d, \tag{3.41}$$

is identical to $1/\mu_n$ *on* $\mathbb{R}^d \backslash \{x_1, \ldots, x_n\}$.

Proof. For any $y \in \mathbb{R}^d \backslash \{x_1, \ldots, x_n\}$, the proof follows in a straightforward manner by using the definition of $\mu_n(y)$ in (3.29) and solving the equations in the linear system (3.26) for the coefficient $\beta(y)$ by rearranging and applying the Schur complement of the invertible block A_n in the matrix $A_n(y)$. □

Definition (3.41) implies $v_n(x_i) = 0$, $i = 1, \ldots, n$, since the i-th column of the matrix $\left(\begin{smallmatrix} \Phi & P \\ P^\top & 0 \end{smallmatrix}\right)$ is $(m_n(x_i)^\top, \pi(x_i)^\top)^\top$ and the i-th element of the vector $m_n(x_i)$ is $\phi(0)$. Moreover, $v_n(y) = 1/\mu_n(y)$ is positive and finite for $y \in \mathcal{X} \backslash \{x_1, \ldots, x_n\}$, and the further away y is from the sample points x_i, the higher becomes $v_n(y)$. Hence, v_n can be interpreted as a measure of the approximation error of s_n at any $y \in \mathcal{X}$, where no errors occur at the interpolation points but anywhere else depending on the distance to these points.

In consequence to definition (3.41), the minimisation of μ_n on $\mathcal{X} \backslash \{x_1, \ldots, x_n\}$ is thus equivalent to maximising the utility function v_n on $\mathcal{X}$. This applies in case $f_n^* = -\infty$. If $-\infty < f_n^* < \min_{y \in \mathcal{X}} s_n(y)$, then the minimisation of g_n on $\mathcal{X} \backslash \{x_1, \ldots, x_n\}$ may be replaced with the maximisation of the utility function

$$h_n(y) := \frac{v_n(y)}{\left[f_n^* - s_n(y)\right]^2}, \qquad y \in \mathcal{X}, \tag{3.42}$$

which sets the error made by the surface s_n in relation to its distance to the employed target value f_n^*.

By minimising the functions s_n, $-v_n$ and $-h_n$ on $\mathcal{X}$, the respective formulas (3.22), (3.41) and (3.42) provide that their smoothness essentially depends on the smoothness of the term $\phi(\|\cdot\|_2)$. It can therefore be concluded for the radial basis functions in Table 3.1 that

$$s_n, v_n, h_n \in \begin{cases} C^{\nu-1}(\mathbb{R}^d), & \text{for surface splines,} \\ C^\infty(\mathbb{R}^d), & \text{for (inverse) multiquadrics and Gaussians.} \end{cases}$$

Moreover, due to the form of v_n and h_n, the factorisation of the interpolation matrix A_n in (3.30), which is required to construct the surface s_n, can be conveniently reused for evaluating v_n and h_n. Hence, despite constituting again global optimisation problems like the original problem (3.1), the respective subproblems have objective functions that can be evaluated quickly, allow for analytical gradients and carry a particular structure which can be exploited by adequate methods.

To minimise s_n, $-v_n$ and $-h_n$ on $\mathcal{X}$, Gutmann [2001b], Section 5.1, suggests to use either a limited memory branch-and-bound algorithm or a heuristic approach for problems with small dimension. Underlying idea of both methods is to locate promising points by a branch-and-bound technique or by maximising the distance to the nearest sample point within a reasonable number of iterations, respectively, from which local searches are then run to find the local optima. Björkman and Holmström [2000] consider that it is sufficient to approximatively solve the minimisation of s_n by starting a local search from the sample point with the lowest function value $f_n^{\min}$, where a version of the BFGS algorithm is employed in their method. For solving the second subproblem, the authors suggest minimising the transformed utility functions $-\log v_n$ and $-\log h_n$ instead of $-v_n$ and $-h_n$, respectively, to avoid a flat minimum and numerical issues when the latter functions are very small. The respective problems are then solved by an implemented variant of the DIRECT algorithm (Jones et al. [1993]), which is also part of their TOMLAB optimisation environment.

3.3.3.3 Choice of Target Value

A further important issue in the implementation of the method is the choice of suitable target values f_n^*. To effectively balance between a global and a local search, it has proven most useful to set the f_n^* by repeatedly employing short cycles, where each cycle starts with a low value to enforce the global aspect and successively gets closer to $\min_{y\in\mathcal{X}} s_n(y)$ until the target value is eventually set to $f_n^* = \min_{y\in\mathcal{X}} s_n(y)$, corresponding to a pure local search. In particular, the following two strategies, as suggested by Gutmann [2001b], Section 5.2, have turned out to be perform well in the majority of applications:

I. For each cycle of length 3 and starting at $n = n^{\mathrm{cy}}$, set

$$\begin{aligned} f^*_{n^{\mathrm{cy}}} &= -\infty, \\ f^*_{n^{\mathrm{cy}}+1} &= \begin{cases} \min\limits_{y\in\mathcal{X}} s_{n^{\mathrm{cy}}+1}(y) - 0.1|\min\limits_{y\in\mathcal{X}} s_{n^{\mathrm{cy}}+1}(y)|, & |\min\limits_{y\in\mathcal{X}} s_{n^{\mathrm{cy}}+1}(y)| > 0, \\ -0.1, & |\min\limits_{y\in\mathcal{X}} s_{n^{\mathrm{cy}}+1}(y)| = 0, \end{cases} \\ f^*_{n^{\mathrm{cy}}+2} &= \min_{y\in\mathcal{X}} s_{n^{\mathrm{cy}}+2}(y). \end{aligned}$$

II. For each cycle of length $l^{\mathrm{cy}} + 1$, where typically $l^{\mathrm{cy}} \leq 5$, and starting at $n = n^{\mathrm{cy}}$, set

$$f_n^* = \min_{y\in\mathcal{X}} s_n(y) - \left(\frac{l^{\mathrm{cy}} - (n - n^{\mathrm{cy}})}{l^{\mathrm{cy}}}\right)^2 \tilde{\Delta}_n, \qquad n^{\mathrm{cy}} \leq n \leq n^{\mathrm{cy}} + l^{\mathrm{cy}},$$

where $\tilde{\Delta}_n$ depends on the computed function values and is set according to the range of function values either to

$$\tilde{\Delta}_n^{(a)} = \max_{1\leq i\leq n} f(x_i) - \min_{y\in\mathcal{X}} s_n(y),$$

or to

$$\tilde{\Delta}_n^{(b)} = \max_{1\leq i\leq n_\iota} f(x_{\iota(i)}) - \min_{y\in\mathcal{X}} s_n(y),$$

where $n_\iota = \left\lfloor \frac{l^{\mathrm{cy}} - (n - n^{\mathrm{cy}})}{l^{\mathrm{cy}}} n \right\rfloor$ and ι is a permutation of $\{1, \ldots, n\}$ such that $f(x_{\iota(1)}) \leq \ldots \leq f(x_{\iota(n)})$.

Note that in both strategies the target value is set to $f_n^* = \min_{y\in\mathcal{X}} s_n(y)$ at the end of each cycle. This choice, however, is only admissible if none of the sample points $x_1, \ldots, x_n$ is a global minimiser of s_n, i.e. if $\min_{y\in\mathcal{X}} s_n(y) < f_n^{\min} = \min\{f(x_1), \ldots, f(x_n)\}$. Hence, to ensure admissibility in an implementation of the method, the difference between $\min_{y\in\mathcal{X}} s_n(y)$ and $f_n^{\min}$ is checked to be sufficiently large and if this is not the case the choice of f_n^* is then reset to a marginally lower value. More specifically, if

$$\min_{y\in\mathcal{X}} s_n(y) \geq f_n^{\min} - 10^{-4}|f_n^{\min}|, \qquad f_n^{\min} \neq 0,$$

or

$$\min_{y\in\mathcal{X}} s_n(y) \geq -10^{-4}\min\{1, f_n^{\max}\}, \qquad f_n^{\min} = 0,$$

where $f_n^{\max} := \max\{f(x_1), \ldots, f(x_n)\}$, then the target value f_n^* is reset for $f_n^{\min} \neq 0$ to

$$f_n^* = f_n^{\min} - 10^{-2}|f_n^{\min}|,$$

and for $f_n^{\min} = 0$ to

$$f_n^* = \begin{cases} -10^{-2}\min\{1, f_n^{\max}\}, & f_n^{\max} > 0, \\ -10^{-2}, & f_n^{\max} = 0, \end{cases}$$

thus providing a value slightly below $\min_{y\in\mathcal{X}} s_n(y)$ and avoiding that the global minimiser of s_n is not too close to a previously evaluated point.

3.3.3.4 Further Issues

As remarked by Gutmann [2001b], Section 5.3, numerical problems may arise when there are large differences between function values, in which cases the built surfaces tend to oscillate strongly. It may even lead to situations where the minimum of s_n is much lower than the best known function value $f_n^{\min}$, implying a choice of f_n^* that overemphasises global search, as observed by Björkman and Holmström [2000]. To circumvent these issues, Gutmann suggests to replace large function values in each iteration by the median of all available function values if they exceed this median. Holmström [2008] argues that if the surface is oscillating wildly then setting f_n^* is needless as it produces unrealistic new points. Instead, he advises to add the minimum of the surface s_n repeatedly until the interpolation stabilises. Eventually, Regis and Shoemaker [2013] point out that response surface models are often affected by extreme (i.e. very high or very low) function values and to this end propose applying a function-stabilising transformation on the objective function if extreme values are detected. As an effective choice to reduce the influence of extreme values, they propose the transformation

$$\operatorname{plog}(x) = \begin{cases} \log(1+x), & x \geq 0, \\ -\log(1-x), & x < 0. \end{cases}$$

Chapter 4

A Modified Radial Basis Function Method with Extended Local Search

Gutmann's original RBF method for the global minimisation of an expensive objective function as described in Subsection 3.3 is theoretically well-founded and has turned out to work reliably well on a number of well-behaved optimisation problems with reasonably smooth objective functions. This includes most of the available test problems consulted in the literature on response surface methods as well as some real-life problems of moderate nature, where in both situations a reasonably small number of function evaluations is sufficient to find a global optimum. Yet, when applied to solve optimisation problems with a more complex and potentially highly nonlinear type of objective function, it has been reported by several authors that the method lacks efficiency and tends to converge only very slowly to a global minimum, see, e.g., Gutmann [2001b], Chapter 5,[18] but above all Regis and Shoemaker [2007b] and Holmström [2008].

As outlined most comprehensively by Regis and Shoemaker [2007b], the poor practical performance of the RBF method on these kind of problems can essentially be attributed to a highly sensitive choice of target values, which is due to the intricate structure of the underlying objective function and easily leads to a malfunctioning local search. To be more explicit, recall that for a low target value f_n^* the method performs global search such that the next iterate x_{n+1} is away from the previously evaluated points, while for a choice of f_n^* close or equal to $\min_{y\in\mathcal{X}} s_n(y)$ it is expected to perform local search. The point x_{n+1} is then sampled either in the vicinity of a global minimiser of the surface s_n if $f_n^* < \min_{y\in\mathcal{X}} s_n(y)$, or as a global minimiser of s_n if $f_n^* = \min_{y\in\mathcal{X}} s_n(y)$ and this choice is admissible, cf. Regis and Shoemaker [2007b], Theorem 1. However, even though f_n^* is relatively close to $\min_{y\in\mathcal{X}} s_n(y)$, it is not guaranteed that the method actually performs local search as this also depends on the approximant s_n (via the utility function g_n) and thus on the structure of the objective function f. In particular, whereas for rather smooth objective functions with a well-proportioned

[18] In particular, the author remarks that this behaviour is not unique to the RBF method but may also be encountered for the early Bayesian and regression-based methods, see Subsection 3.1.2.

structure the local search mechanism of the RBF method can be controlled quite effectively by a suitable choice of target values, it is frequently observed that the method has considerable difficulties in minimising objective functions whose local minima are situated at the bottom of steep and narrow valleys but are otherwise largely flat, cf., e.g., the well-known Shekel test functions from the Dixon-Szegö test set (see Dixon and Szegö [1978]). For these functions, the disproportional structure with extremely 'bumpy' regions of attraction though little curvature and very small differences in function values in the remaining regions makes it very difficult to capture the intended impact of a locally set f_n^* on the next iterate x_{n+1}. As a result, the method is then indeed able to detect points close to a local minima, but in turn often fails to make significant progress towards the local minimiser within an acceptable number of function evaluations, or to find new local minima with lower function values. Instead, supposed local search points with f_n^* relatively close or equal to $\min_{y\in\mathcal{X}} s_n(y)$ are either sampled away from the global minimisers of the response surfaces in the flatter regions of the objective function, if these are yet not covered sufficiently well by global search points, or are sampled in the vicinity but in an ineffective manner. This, though, leads to a slow convergence of the RBF method to a global minima.

To overcome the main drawback of poor practical performance of the RBF method mainly noticed on relevant test problems, some suggestions have been made in the literature, see also Subsection 3.1.2 for a brief outline. In the first place, Gutmann [2001b], Chapter 5, himself already notes that the RBF method performs disappointingly on some test functions with above described characteristics. On this account, he conjectures that it may be of advantage to employ a specific local search technique at some stages of the method, see his concluding Chapter 7, without following up on it though. Based on their conclusive insights, Regis and Shoemaker [2007b] suggest to prevent a malfunctioning local search by restricting the global minimisation of g_n to a small hyperrectangle centred around a global minimiser of the current surface s_n, whenever f_n^* is reasonably close to $\min_{y\in\mathcal{X}} s_n(y)$ and the method is supposed to perform a local search. Depending on the expected influence of a local search step, the size of the hyperrectangle is then adjusted, thus promoting a better balance between local and global search. Rather than adjusting the parameter space, Holmström [2008] directly addresses the choice of target values to achieve an improved practical performance of the method. Substantiated by practical experience, he argues that the choice of target values is too static and very dependent on the scaling of the problem. To make it more flexible, he suggests an adaptive approach where in each iteration a set of potential evaluation points is pre-sampled by minimising g_n for a range of target values, from which then suitable new evaluation points are selected by a clustering algorithm. Eventually, Cassioli and Schoen [2013] propose an algorithm to improve the accuracy of response surfaces in case a lower bound of the objective function is known. Specifically, by solving a sequence of convex quadratic programmes with equality and inequality constraints, refined lower-bounded response surfaces may be constructed, which then prevent the RBF method from selecting rather unrealistic target values. This, in turn, reinforces the local search mechanism and thus improves the practical convergence of the method.

Unfortunately, a slow convergence is typically also observed when the RBF method is applied to solve calibration problems, or data-fitting problems in general, due to their inherent highly nonlinear structure. Most commonly, these problems are characterised by an objective function that is bounded from below by zero and exhibits different types of behaviour: while in regions with lower function values the objective function appears rather insensitive to any changes in the parameter, it reacts very strongly to parameter changes in other regions of the parameter space, leading to considerable differences in function values. Moreover, the local minima are frequently situated inside elongated, wide and shallow regions of attraction, such that it is fairly straightforward to detect these regions but more difficult to converge to a local minimum, cf., for instance, the Rosenbrock test function (e.g., Dixon and Szegö [1978]) as a standard example illustrating this issue.

If, on the one hand, this type of objective function is minimised in its original form by the RBF method, the structure involves that the response surfaces tend to oscillate strongly in the lower range of function values, with $\min_{y \in \mathcal{X}} s_n(y)$ being much lower than the best known function value $f_n^{\min}$. Since f_n^* is set even lower than $\min_{y \in \mathcal{X}} s_n(y)$, however, this typically leads to an unrealistic choice of target values that overemphasises global search and generally produces irregular search points, see also Holmström [2008]. In particular, supposed local search points are then indeed sampled in the flatter regions of attraction, but in an overly global and thus effectless manner away from previously evaluated points until $\min_{y \in \mathcal{X}} s_n(y)$ is sufficiently close to $f_n^{\min}$ and local search may actually take effect. Depending on the dimension of the objective function and its scale, though, this may require quite a considerable amount of function evaluations in the beginning of an optimisation which prevents the method from a reasonable convergence, notwithstanding that the wide and flat form of the valleys seem to favour local search at later stages.

On the other hand, if such an objective function is minimised by applying a suitable transformation (e.g., a log-transformation) or some other modification that scales down large function values and enlarges small ones (e.g., by replacing large function values by the median of all function values if they exceed this value), the resulting function typically exhibits relatively small differences in function values but has elongated, narrow and steep valleys, similar to the critical problems mentioned in literature. Accordingly, even though a better approximation is facilitated for lower function values, as $\min_{y \in \mathcal{X}} s_n(y)$ tends to be comparably close to $f_n^{\min}$, the distinctive form of the valleys relative to the overall objective function leads to above situations causing a malfunctioning local search and thus a slow convergence of the method, cf. Regis and Shoemaker [2007b].

In either case, it has to be stressed that the elongated valleys, in which the local minimisers of calibration problems frequently lie, additionally hamper the convergence of the RBF method if they cannot be modelled adequately well. Specifically, in such cases, the particular form of the valleys involves that global minimisers of the response surfaces are very likely to be found too close to a previously evaluated point along a valley, even though this point is yet not in the immediate vicinity of the respective local minimiser. Due to the inverse arc-shaped form of the involved function μ_n, however, this implies that intrinsic local search points are actually only sampled for very large target values sufficiently close or equal to $\min_{y \in \mathcal{X}} s_n(y)$, and therefore must lie again very close to a previously evaluated point, cf. Regis and Shoemaker [2007b]. Local search will thus become highly ineffective, which significantly slows down the convergence of the method towards a local minimiser. In particular, the elongated form of the valleys are also a main reason why the suggested improvements by Regis and Shoemaker [2007b], Holmström [2008] and Cassioli and Schoen [2013] have only a limited effect on the performance of the method for our application, if any at all, see Subsection 4.3.1.

In consideration of above drawbacks and the suggested improvements thus far, we propose the following two enhancements to the original RBF method when applied to calibration problems and other data-fitting problems. Both add to the existing literature on the RBF method and essentially contribute to a significant improvement in the performance of the method. Nevertheless, either enhancement may just as well be used by itself, or in combination with some compatible modification as suggested in the literature, to solve similar global optimisation problems if rendered possible by their structure.

The first enhancement, henceforth termed *modified RBF method*, concerns the construction of response surface models and takes advantage of the particular nonlinear structure inherent to calibration problems. Specifically, given that the objective function in these kind of optimisation problems is computed through

$$f(x) = \sum_{k=1}^{l} g\big(f_k(x)\big), \qquad x \in \mathbb{R}^d, \tag{4.1}$$

where the $f_k : \mathbb{R}^d \to \mathbb{R}$ are residual functions and $g : \mathbb{R} \to \mathbb{R}_{\geq 0}$ a nonnegative continuous function measuring the discrepancy between model and observed data, we approximate f by first interpolating each f_k with an individual residual surface s_k and then forming a universal surface $s = \sum_{k=1}^{l} g(s_k)$, rather than directly approximating f by a single surface.

Despite requiring very minor additional computational effort for a manageable number of residuals, the modified construction has several important advantages over the usual approximation. Firstly, by assessing the contribution of each residual to the objective function instead of the objective function itself, the approach is able to approximate the objective function more accurately and to capture its global behaviour better, especially in those regions of the parameter space that are most problematic. This involves that the resulting universal surface implicitly assumes a lower bound greater than zero (as g is nonnegative) and is less oscillating than in an ordinary construction while still preserving the original difference in function values. In particular, it thus avoids the numerical issues as reported by Gutmann [2001b], Björkman and Holmström [2000] and Holmström [2008], cf. Subsection 3.3.3.

Secondly, the construction allows to remain within Gutmann's well-established framework in which new evaluation points are selected by means of a target value. As shall be seen in

Section 4.1, the concept of minimising a semi-norm can be formulated in a multivariate setup on the residual surfaces, such that a new evaluation point corresponds to a Pareto optimal solution of a multi-objective optimisation problem under a particular parameterisation. This derivation then facilitates on the theoretical side that already available convergence results by Gutmann can be adopted subject to minor modifications, while on the practical side target values can be selected more effectively, thus enhancing the local search mechanism, cf. Holmström [2008] and Cassioli and Schoen [2013].

Eventually, by interpolating the residual functions instead of the objective, we are able to gain useful information on the underlying parameterised models, which in turn can be used to advantage for modelling related matters. For example, in the calibration of derivative pricing models, the approximation of the expensive pricing vector can be reused cheaply for the valuation of the involved financial derivatives.

To further improve the poor practical performance on more involved but not prohibitively expensive global optimisation problems, our second enhancement directly addresses the sampling stage of Gutmann's RBF method and may be used with the first modification of constructing improved response surfaces. It essentially consists of a simple yet effective *extended local search technique* that complements the local search mechanism of both methods to promote convergence to a global minimum, thus pursuing the initial idea of Gutmann [2001b], Chapter 7. To be more specific, the technique makes use of the ability of the (modified) RBF method to provide a reasonable global model of the objective function by which regions of attraction for a global minimiser can be identified. However, once such a region has been detected and no substantial progress is made by the inherent local search of the method, e.g., due to ineffectively sampled local search points, a conventional local search method is enabled by means of a clustering technique. This method then further explores the identified region for a local minimiser, before continuing upon an updated surface with the ordinary search mechanism of the (modified) RBF method to detect new regions of attraction.

The use of external local search information when most required thus facilitates the method to better advance into more complex neighbourhoods of local minima, without relying on a strategy of target values that usually causes an insufficiently working local search mechanism for more intricate problems, cf. Regis and Shoemaker [2007b] and Holmström [2008]. In particular, as the choice of local search method is intentionally left unspecified, the extended technique allows to exploit valuable gradient information if available, either analytically or as obtained by some practicable numerical approximation[19], while it may be equally applied in a derivative-free context. Related ideas of using local search techniques are also employed in the general response surface methods by Regis and Shoemaker [2013] and Ji et al. [2013], for instance, and in other global optimisation methods to improve their performance, such as in the MCS algorithm by Huyer and Neumaier [1999]. After all, multi-start

[19] The direct use of analytical or numerical gradients in the construction of interpolating response surfaces is not advisable since they easily lead to overly oscillating and inexact surfaces, see, e.g., Fowkes [2011], Section 3.4. Also, they require expensive function evaluations in less relevant parts of the parameter space.

strategies, which are frequently used to calibrate financial models, also work by carrying out a number of local searches, albeit in a less sophisticated manner.

Similar to the outline of Gutmann's RBF method in Section 3.3, we proceed by first giving a theoretical description of the modified RBF method with extended local search in Section 4.1 and then state the related convergence of the method in Section 4.2. Eventually, Section 4.3 is devoted to the numerical analysis of the method, where we illustrate its practical applicability on some relevant test problems, as well as by fitting the Nelson-Siegel and Svensson models and calibrating the Hull-White model under the SAA strategy.

4.1 Description of Method

To describe the modified RBF method with extended local search for minimising the objective function (4.1) on a compact set $\mathcal{X}$, we place ourselves in the same setup as the original RBF method, cf. Subsection 3.3.1. Accordingly, assume that a conditionally positive definite radial basis function ϕ of order m is given, along with a polynomial basis $\{p_j\}_{j=1}^{\widetilde{m}}$, and let $\{x_1, \ldots, x_n\} \subset \mathcal{X}$ be a $\mathcal{P}_m^d$-unisolvent set of previously generated points at which the corresponding residual function values are known. The respective components of the method are then as follows, summarised by a compact overview at the end of this section.

4.1.1 Construction of Response Surface

To better capture the nonlinear structure of the objective function in (4.1), we build a universal surface s_n by

$$s_n(x) = \sum_{k=1}^{l} g\big(s_{n,k}(x)\big), \quad x \in \mathbb{R}^d, \tag{4.2}$$

where $g : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is a nonnegative continuous function and each residual surface $s_{n,k}$, relying on the same radial basis function ϕ and polynomial basis $\{p_j\}_{j=1}^{\widetilde{m}}$, is of the form

$$s_{n,k}(x) = \sum_{i=1}^{n} \lambda_{ik}\phi(\|x - x_i\|) + \sum_{j=1}^{\widetilde{m}} c_{jk}p_j(x), \quad x \in \mathbb{R}^d,$$

for real coefficients $\{\lambda_{ik}\}_{i=1}^{n}$ and $\{c_{jk}\}_{j=1}^{\widetilde{m}}$. The latter coefficients are then determined by requiring that each residual surface $s_{n,k}$ satisfies the interpolation and side constraints

$$\begin{aligned} s_{n,k}(x_i) &= f_k(x_i), && i = 1, \ldots, n, \\ \sum_{i=1}^{n} \lambda_{ik}p_j(x_i) &= 0, && j = 1, \ldots, \widetilde{m}, \end{aligned}$$

respectively. This ensures that s_n in (4.2) in itself meets the interpolation conditions $s_n(x_i) = f(x_i)$, $i = 1, \dots, n$, and eventually amounts to solving the linear system with multiple right-hand side

$$\begin{pmatrix} \Phi & P \\ P^\top & 0 \end{pmatrix} \begin{pmatrix} \Lambda \\ C \end{pmatrix} = \begin{pmatrix} F_X \\ 0 \end{pmatrix}, \tag{4.3}$$

where $\Phi \in \mathbb{R}^{n \times n}$ and $P \in \mathbb{R}^{n \times \widetilde{m}}$ assume their usual form given by (3.9) and (3.12), respectively, $\Lambda \in \mathbb{R}^{n \times l}$ and $C \in \mathbb{R}^{\widetilde{m} \times l}$ denote the coefficient matrices, and $F_X \in \mathbb{R}^{n \times l}$ is the matrix of residual function values.

In particular, note that the interpolation matrix on the left-hand side of (4.3) is the same as in the ordinary interpolation system (3.11). Hence, by Theorem 3.7, the system (4.3) is uniquely solvable for the $\mathcal{P}_m^d$-unisolvent set of points $\{x_1, \dots, x_n\}$. Moreover, it follows from the variational theory on radial basis functions that each $s_{n,k} \in \mathcal{A}_\phi(\mathcal{X})$ minimises the semi-norm among all interpolants $s \in \mathcal{A}_\phi(\mathcal{X})$ subject to the same conditions $s(x_i) = f_k(x_i)$, $i = 1, \dots, n$, cf. Theorem 3.9. Eventually, it is worth emphasising that, even though each $s_{n,k}$ belongs to the linear function space $\mathcal{A}_\phi(\mathcal{X})$, the resulting universal response surface s_n is not in $\mathcal{A}_\phi(\mathcal{X})$, unless g is the identity function.

As to the additional computational effort involved in solving the system (4.3), recall that the factorisation of the interpolation matrix in the ordinary linear system (3.11) as well as the ensuing solve step to obtain the coefficients λ and c from the factorised matrices both require $\mathcal{O}(n^2)$ operations, see Subsection 3.3.3. Hence, since both systems have the same interpolation matrix, the cost of factorisation is the same, while it requires the solution of l linear systems with the same system matrix and different right-hand sides to obtain the coefficient matrices Λ and C from $F_X \in \mathbb{R}^{n \times l}$. Consequently, the total effort of solving system (4.3) amounts to $\mathcal{O}(ln^2)$ operations, such that the additional cost of interpolating residual functions instead of the objective becomes negligible for a moderate size l. In contrast, the approach pursued by Cassioli and Schoen [2013] is computationally much more involved. Here, the response surface s_n in an iteration of the method is gradually refined by solving a series of convex quadratic programmes with equality and inequality constraints, until the minimum of the surface is larger than the known lower bound of the objective function f or a specified number of iterations is exceeded. In particular, this requires to build and minimise a new response surface in each single step of the refinement process, having added the global minimiser of the previous surface under the constraint that its new surface value is bound from below by the known value. Of course, one has to bear in mind that the approach by Cassioli and Schoen is more general and applies to a wider class of optimisation problems than data-fitting problems. Nevertheless, the latter represent a very important subclass.

Once constructed, note that the evaluation of $s_n(x)$ at any $x \in \mathbb{R}^d$ requires $2l(n + \widetilde{m})$ operations for obtaining the residual surface values $s_{n,k}(x)$, $k = 1, \dots, l$, plus the cost of forming the $g(s_{n,k})$'s and taking the subsequent sum over these, instead of the original $2(n + \widetilde{m})$ for a single surface s_n.

4.1.2 Determination of Next Evaluation Point

Given above construction, we can proceed in a similar way to Gutmann for finding a new evaluation point x_{n+1}. Specifically, for an admissible target value $f_n^* \in [-\infty, \min_{y\in\mathcal{X}} s_n(y)]$, with $f_n^* \neq s_n(x_i)$, $i = 1, \ldots, n$, we let x_{n+1} be the value of $y \in \mathcal{X}\backslash\{x_1, \ldots, x_n\}$ such that it solves

$$\min_{y\in\mathcal{X}\backslash\{x_1,\ldots,x_n\}} \mu_n(y)\big[f_n^* - s_n(y)\big]^2, \tag{4.4}$$

where μ_n is the nonnegative function given by definition (3.29) for the Lagrange basis function $l_n(y, \cdot\,)$ of (3.25) satisfying the interpolation conditions (3.24), i.e.

$$\begin{aligned} l_n(y, x_i) &= 0, \qquad i = 1, \ldots, n, \\ l_n(y, y) &= 1. \end{aligned}$$

As the solution of (4.4), it can be shown that y corresponds to a Pareto optimal solution[20] of the multi-objective optimisation problem in which the semi-norms of the augmented residual interpolants $s_{y,k} \in \mathcal{A}_\phi(\mathcal{X})$ through $(x_1, f_k(x_1)), \ldots, (x_n, f_k(x_n))$ and $(y, f_{n,k}^*)$ are simultaneously minimised for some artificially given but yet not further specified residual target values $f_{n,k}^*$, $k = 1, \ldots, l$. Put another way, we thus apply Jones's response surface technique with radial basis functions in a multi-objective setup.

To see the equivalence to a Pareto optimal solution, we operate on the residual interpolants and again first rewrite each optimal interpolant $s_{y,k} \in \mathcal{A}_\phi(\mathcal{X})$ satisfying the interpolation constraints

$$\begin{aligned} s_{y,k}(x_i) &= f_k(x_i), \qquad i = 1, \ldots, n, \\ s_{y,k}(y) &= f_{n,k}^*, \end{aligned}$$

as

$$s_{y,k}(x) = s_{n,k}(x) + \big[f_{n,k}^* - s_{n,k}(y)\big]l_n(y, x), \quad x \in \mathbb{R}^d,$$

where $l_n(y, \cdot\,) \in \mathcal{A}_\phi(\mathcal{X})$ is the optimal Lagrange basis function (3.25), common to all $s_{y,k}$ and meeting condition (3.24). By the latter representation, each squared semi-norm $s_{y,k}$ can then be simplified as in Gutmann's RBF method, such that the problem of minimising all semi-norms at once leads to the multi-objective optimisation problem

$$\min_{y\in\mathcal{X}\backslash\{x_1,\ldots,x_n\}} \mu_n(y)\Big(\big[f_{n,1}^* - s_{n,1}(y)\big]^2, \ldots, \big[f_{n,l}^* - s_{n,l}(y)\big]^2\Big). \tag{4.5}$$

Now, as stated, for instance, in Ehrgott [2005], Proposition 3.9, any optimal solution of a weighted sum scalarisation of (4.5) with positive weights is a Pareto optimal solution of (4.5). In particular, the weights are free parameters and therefore may be set a posteriori and in dependence of a solution y of (4.4), such that y corresponds to a Pareto optimal solution of (4.5) under a suitably chosen parameterisation implicitly defining the target values $f_{n,k}^*$.

[20] A feasible solution x^* is called Pareto optimal for the problem $\min_{x\in\mathcal{X}} f(x)$ with $f = (f_1, \ldots, f_l)$ if there is no other $x \in \mathcal{X}$ such that $f_k(x) \leq f_k(x^*)$, $k = 1, \ldots, l$, and $f(x) \neq f(x^*)$.

Theorem 4.1. *Let $y \in \mathcal{X} \backslash \{x_1, \ldots, x_n\}$ be an optimal solution of problem* (4.4) *with admissible target value f_n^*. Then, y is a Pareto optimal solution of the multi-objective optimisation problem* (4.5) *for implicitly defined residual target values $f_{n,k}^*$, $k = 1, \ldots, l$.*

Proof. Let $y \in \mathcal{X} \backslash \{x_1, \ldots, x_n\}$ be a global minimiser of (4.4). This problem can be rewritten as

$$\min_{y \in \mathcal{X} \backslash \{x_1, \ldots, x_n\}} \sum_{k=1}^{l} w_k^{\text{sc}}(y) g_{n,k}(y), \tag{4.6}$$

where

$$g_{n,k}(y) := \mu_n(y) \big[f_{n,k}^* - s_{n,k}(y)\big]^2, \qquad y \in \mathcal{X} \backslash \{x_1, \ldots, x_n\},$$

and the positive weights $w_k^{\text{sc}}(y)$ are to be set accordingly in terms of y.

If $f_n^* = \min_{y \in \mathcal{X}} s_n(y)$ and this choice is admissible, then we implicitly define the residual target values as $f_{n,k}^* = s_{n,k}(y)$, $k = 1, \ldots, l$, such that the objective functions of both problems (4.4) and (4.6) yield zero for any choice of positive weights $w_k^{\text{sc}}(y)$. If $f_n^* < \min_{y \in \mathcal{X}} s_n(y)$, then we set for implicitly defined $f_{n,k}^* \neq s_{n,k}(y)$ the positive weights to

$$w_k^{\text{sc}}(y) = \begin{cases} \dfrac{1}{l} \dfrac{\big[f_n^* - s_n(y)\big]^2}{\big[f_{n,k}^* - s_{n,k}(y)\big]^2}, & f_n^* \neq -\infty, \\ \dfrac{1}{l} \dfrac{1}{\big[f_{n,k}^* - s_{n,k}(y)\big]^2}, & f_n^* = -\infty. \end{cases}$$

In any of these cases, problem (4.6) constitutes a weighted sum scalarisation of (4.5) with positive weights, and hence y is a Pareto optimal solution of the latter. □

In particular, thus note that the functionality of the target value f_n^* remains the same as in Gutmann's RBF method since it is set in relation to the universal response surface s_n.

4.1.3 Extended Local Search

To equip the (modified) RBF method with an extended local search technique, we initially recall that for determining new evaluations points the critical target values are commonly set in cycles by some predefined strategy, cf. Subsection 3.3.3. Each cycle has the purpose of balancing between global and local search such that the generated sample points may approximately be divided into global and local search points, according to the supposed functionality of the target value within the chosen strategy. Once an unexplored region of attraction is detected by the method for local exploitation, it is thus natural that local search points accumulate along the convergence of $\operatorname{argmin}_{y \in \mathcal{X}} s_n(y)$ to the local minimiser of the objective function within this region, be it due to a well-performing local search mechanism where most points fall into the supposed vicinity or a malfunctioning one where only a few intrinsic local search points are actually sampled nearby. Yet, if inherent local search steps prove to be increasingly ineffective such that no substantial progress is made by the method in a number of iterations, the accumulation becomes notably distinctive, resulting in a concentrated clustering of local search points around $\operatorname{argmin}_{y\in\mathcal{X}} s_n(y)$. This may then be taken as indication to further explore the associated region by an external local search method, provided that function evaluations are not too expensive, before returning to the ordinary search procedure of the (modified) RBF method to uncover new promising regions.

Describing the extended local search technique in step n more formally, we assume that the new iterate x_{n+1} has been sampled in a supposed local search of the method and that f has been evaluated at that point. We let the set of interpolation points $\mathcal{X}_n = \{x_1, \ldots, x_n\}$ combined with x_{n+1} be decomposed into the disjoint union $\mathcal{X}_n = \mathcal{X}_n^{\mathrm{glob}} \cup \mathcal{X}_n^{\mathrm{loc}} \cup \mathcal{X}_n^{\mathrm{ls}}$, where $\mathcal{X}_n^{\mathrm{glob}}$ and $\mathcal{X}_n^{\mathrm{loc}}$ denote the sets of points generated by global and local searches of the (modified) RBF method, respectively, and $\mathcal{X}_n^{\mathrm{ls}}$ is the set of interpolation points provided by the external local search technique. Moreover, we let $\mathcal{X}_n^{\mathrm{opt}}$ be the set of optimal solutions obtained from already executed extended local searches.

In its most basic form, the extended local search technique then proceeds as depicted in Algorithm 4.2, following a standard local search step of the (modified) RBF method, see Algorithm 4.3 for its integration. It requires as input the parameters $\epsilon^{\mathrm{cl}} > 0$ and $n^{\mathrm{cl}} \in \mathbb{N}$ to discern any clusters among local search points, a maximum relative error tolerance $\epsilon^{\mathrm{rel}} > 0$ to set the critical minimum function value $\bar{f}_n^{\min}$ for initialising an extended local search, and a minimum distance value $\epsilon^{\mathrm{sel}} > 0$ to extract interpolation points from the set of points evaluated during a local search run. As default values, we preset the respective parameters to $\epsilon^{\mathrm{cl}} = 0.05$ and $n^{\mathrm{cl}} = 5$, as well as to $\epsilon^{\mathrm{rel}} = 0.2$ and $\epsilon^{\mathrm{sel}} = 0.05$, provided that the technique is applied on a unit hypercube with the Euclidean distance.

Algorithm 4.2. *(Extended Local Search).*

```
Check whether points from the set $\mathcal{X}_n^{loc} \cup \mathcal{X}_n^{opt}$ form a density-based cluster $\mathcal{C}(x_n^*; \epsilon^{cl}, n^{cl})$
around an $x_n^* = \operatorname{argmin}_{y\in\mathcal{X}} s_n(y)$ with respect to $\epsilon^{cl}$ and $n^{cl}$.

if $\mathcal{C}(x_n^*; \epsilon^{cl}, n^{cl}) \neq \emptyset$

    Set $\bar{f}_n^{\min} = f_n^{\min} + \epsilon^{rel} |f_n^{\min}|$, if $f_n^{\min} \neq 0$, and $\bar{f}_n^{\min} = \epsilon^{rel}$, if $f_n^{\min} = 0$.

    if $\mathcal{C}(x_n^*; \epsilon^{cl}, n^{cl}) \cap \mathcal{X}_n^{opt} = \emptyset$ and $\min_{x_i \in \mathcal{C}(x_n^*; \epsilon^{cl}, n^{cl})} f(x_i) < \bar{f}_n^{\min}$
        - Run a local search method, starting from $\bar{x}_0 = \operatorname{argmin}_{x_i \in \mathcal{C}(x_n^*; \epsilon^{cl}, n^{cl})} f(x_i)$ and
          returning a set of evaluated points $\bar{\mathcal{L}}$ with local minimiser $\bar{x}^*$.

        - Select a subset $\bar{\mathcal{L}}^{sel} \subseteq \bar{\mathcal{L}}$ by:

            Choosing $\bar{x}^*$ if $\bar{x}^* \notin \mathcal{X}_n$, and $\operatorname{argmin}_{\bar{x} \in \bar{\mathcal{L}} \setminus \{\bar{x}^*\}} f(\bar{x})$ otherwise.
            Sorting the remaining points in $\bar{\mathcal{L}}$ in ascending order of their function
            values to successively pick the point with the next lowest function value,
            which is a minimum distance $\epsilon^{sel}$ away from any point in $\mathcal{X}_n$ and in $\bar{\mathcal{L}}^{sel}$.

        - Add $\bar{\mathcal{L}}^{sel}$ to $\mathcal{X}_n^{ls}$, and $\bar{x}^*$ to $\mathcal{X}_n^{opt}$. Set $n = n + \bar{n}^{sel}$ and $n^{funEvals} = n^{funEvals} + \bar{n}$,
          where $\bar{n}^{sel} = |\bar{\mathcal{L}}^{sel}|$ and $\bar{n} = |\bar{\mathcal{L}}|$.
        end if
    end if
```

Given that points generated during local search steps of the (modified) RBF method accumulate upon locating a region of attraction, Algorithm 4.2 first checks whether these points form in union with $\mathcal{X}_n^{\text{opt}}$ a density-based cluster $\mathcal{C}(x_n^*; \epsilon^{\text{cl}}, n^{\text{cl}})$ around a global minimiser x_n^* of the surface s_n, with respect to the parameters ϵ^{cl} and n^{cl}. By a cluster $\mathcal{C}(x_n^*; \epsilon^{\text{cl}}, n^{\text{cl}})$, we understand a nonempty subset of points from $\mathcal{X}_n^{\text{loc}} \cup \mathcal{X}_n^{\text{opt}}$, where the ϵ^{cl}-neighbourhood of x_n^*, denoted by $B_{\epsilon^{\text{cl}}}(x_n^*)$ for some implicitly specified distance function, contains at least n^{cl} points and for every point x_i in the cluster there is a point $x_j \in \mathcal{C}(x_n^*; \epsilon^{\text{cl}}, n^{\text{cl}})$, $i \neq j$, such that $x_i \in B_{\epsilon^{\text{cl}}}(x_j)$ and $B_{\epsilon^{\text{cl}}}(x_j)$ contains at least n^{cl} points. Note that x_n^* thus only belongs to $\mathcal{C}(x_n^*; \epsilon^{\text{cl}}, n^{\text{cl}})$ itself if it coincides with a local search point or an optimal solution of an extended local search. For a detailed presentation on this notion of clustering, we refer to Ester et al. [1996], for instance.

A cluster hence represents a sufficiently dense accumulation of mostly local search points in arbitrary shape, where ϵ^{cl} and n^{cl} both determine its range and level of density and thus also influence the incentive to initialise an extended local search in the associated region of attraction. While for a relatively small value of ϵ^{cl}, a cluster most likely emerges around a not/slowly progressing x_n^* by local search points for which f_n^* is sufficiently close or equal to $\min_{y \in \mathcal{X}} s_n(y)$, a comparably larger choice also includes supposed local search points that are sampled further away. Based on ϵ^{cl}, the parameter n^{cl} then captures the respective local search points in which the method does not make any substantial progress, such that a smaller choice triggers an earlier start of a potential extended local search, and vice versa. In any case, note that once a cluster has been formed in a region of attraction, it persists over the remaining optimisation, irrespective of the location of future x_n^*'s, but gradually increases in size according to its level of function values as more local search points are sampled along the iterations. Each cluster thus also indicates to which extent a particular region of attraction has been explored by local search points of the (modified) RBF method and by optimal solutions of extended local searches (hence the union of $\mathcal{X}_n^{\text{loc}}$ with $\mathcal{X}_n^{\text{opt}}$), provided that ϵ^{cl} is not set disproportionally large.

If, for reasonably chosen ϵ^{cl} and n^{cl}, a cluster $\mathcal{C}(x_n^*; \epsilon^{\text{cl}}, n^{\text{cl}})$ is identified around an x_n^*, it is further examined for initialising a local search method in the associated region of attraction. Specifically, to prevent that an extended local search is executed unnecessarily in the designated region, we require that none of the optimal solutions of prior extended local searches is an element of $\mathcal{C}(x_n^*; \epsilon^{\text{cl}}, n^{\text{cl}})$ and that the minimum function value of the identified cluster is within a specified relative error tolerance ϵ^{rel} from $f_n^{\min}$. Hence, while the first condition essentially rules out that an extended local search is initialised from a neighbourhood of a local minimiser that has already been thoroughly explored by a local search method, the second condition excludes premature initialisations from clusters in which the minimum function value is not sufficiently low compared to the current minimum function value $f_n^{\min}$. Given that both conditions are satisfied, a local search method is then started from the minimiser in $\mathcal{C}(x_n^*; \epsilon^{\mathrm{cl}}, n^{\mathrm{cl}})$, which returns a set $\bar{\mathcal{L}}$ of all points evaluated during the search. Note that the choice of local search algorithm is basically left to the user, but obviously depends on the nature of the objective function. Yet, if gradient information is available in any acceptable form, the technique works most effectively.

After a successful completion of an external local search run, a subset of points $\bar{\mathcal{L}}^{\mathrm{sel}} \subseteq \bar{\mathcal{L}}$ is selected for subsequent integration into the set of interpolation points. If the local minimiser $\bar{x}^*$ is yet not an interpolation point, it is included as first element in $\bar{\mathcal{L}}^{\mathrm{sel}}$; otherwise, the point in $\bar{\mathcal{L}}$ with the next lowest function value is chosen. Any further elements of $\bar{\mathcal{L}}^{\mathrm{sel}}$ are then determined in ascending order of their function values, given that they are a predefined distance ϵ^{sel} away from any point in $\mathcal{X}_n$ and any other point already included in $\bar{\mathcal{L}}^{\mathrm{sel}}$. This has the purpose of extracting the main path of generated local search points to improve the interpolating response surface, while avoiding that numerical issues from interpolation points lying too close together may arise.

Eventually, before resuming the main search procedure of the (modified) RBF method with an improved surface, the set of interpolation points needs to be updated by adding the selection $\bar{\mathcal{L}}^{\mathrm{sel}}$ to the set $\mathcal{X}_n^{\mathrm{ls}}$, along with increasing the numbers of interpolation points and of total function evaluations by the cardinality of the sets $\bar{\mathcal{L}}^{\mathrm{sel}}$ and $\bar{\mathcal{L}}$, respectively. Also, we add $\bar{x}^*$ to the set of optimal solutions $\mathcal{X}_n^{\mathrm{opt}}$, keeping track of the performed local search.

4.1.4 Algorithm

To sum up, the following algorithm employs the suggested improvements on the standard RBF method for minimising an objective function $f = \sum_{k=1}^{l} g(f_k)$ on a compact set $\mathcal{X}$, where g is a nonnegative continuous function measuring the discrepancy between the model and the observations and the f_k's are deterministic and continuous residual functions.

Algorithm 4.3. *(Modified RBF Method with Extended Local Search).*

0. ***Initial step:***

- *Choose a conditionally positive definite radial basis function ϕ of order m.*
- *Generate a $\mathcal{P}_m^d$-unisolvent set of points $\{x_1, \ldots, x_{n_0}\} \subset \mathcal{X}$.*
- *Evaluate $f = \sum_{k=1}^{l} g(f_k)$ at the points $x_1, \ldots, x_{n_0}$, and set $n = n_0$.*

1. ***Iteration step:***

while $n \leq n^{\max}$ *or* $n^{\mathit{funEvals}} \leq n^{\mathit{maxFunEvals}}$ ***do***

- *Construct the interpolant $s_n = \sum_{k=1}^{l} g(s_{n,k})$, where each $s_{n,k} \in \mathcal{A}_\phi(\mathcal{X})$ solves*

$$\min_{s \in \mathcal{A}_\phi(\mathcal{X})} \|s\|_\phi \qquad s.t. \quad s(x_i) = f_k(x_i), \quad i = 1, \ldots, n.$$

```
    - Choose an admissible target value $f_n^* \in \big[-\infty, \min_{y\in\mathcal{X}} s_n(y)\big]$.
    - Determine $x_{n+1}$, which is the value of $y$ that solves
          $\min_{y\in\mathcal{X}\setminus\{x_1,\dots,x_n\}} \mu_n(y)\big[f_n^* - s_n(y)\big]^2$.
    - Evaluate $f = \sum_{k=1}^{l} g(f_k)$ at $x_{n+1}$, and set $n = n+1$.
    - if $x_{n+1} \in \mathcal{X}_n^{loc}$
          Perform an extended local search step according to Algorithm 4.2.
      end if
end while
```

4.2 Convergence of Method

Due to the way in which the modified RBF method with extended local search is constructed, its convergence to the global minimum of any continuous function can be shown in a straightforward manner. In fact, since Algorithm 4.3 employs a similar technique as Gutmann's RBF method for selecting new points and the extended local search technique merely improves its local performance, we may infer its convergence from Gutmann [2001b], see Section 3.3.2, subject to minor modifications.

First recall that the main task of establishing convergence of a deterministic sequential sampling method for any continuous function is to show the density of the sequence of generated points. Applied to the current method, this can be formulated as follows.

Theorem 4.4. *Algorithm 4.3 converges for every continuous function f if and only if it generates a sequence of points $\{x_n\}$ that is dense in $\mathcal{X}$.*

In line with Theorem 4.4, the density of the sequence of generated iterates is then shown in the next theorem, cf. the original version by Gutmann in Theorem 3.17. Again, note that the theorem only applies to surface splines and under a particular assumption on the sequence of target values. Moreover, as the density of the generated sequence essentially relies on the inherent search mechanism of the modified RBF method, it is required that the extended local search step of Algorithm 4.2 is not initialised infinitely many times. This may be ensured by a realistic choice of the initialising parameters ϵ^{cl}, n^{cl} and ϵ^{rel} (i.e. ϵ^{cl} and n^{cl} should not be set too small and ϵ^{rel} not too large).

Theorem 4.5. *Let ϕ be a conditionally positive definite surface spline of order m from Table 3.1, and let $\{x_n\}$ be the sequence generated by Algorithm 4.3. Further, let $s_n = \sum_{k=1}^{l} g(s_{n,k})$, where g is a nonnegative continuous function and each $s_{n,k}$ is the optimal interpolant from $\mathcal{A}_\phi(\mathcal{X})$ to the data $(x_i, f_k(x_i))$, $i = 1, \dots, n$, such that $f(x_i) = \sum_{k=1}^{l} g(f_k(x_i))$. Assume that, for infinitely many $n \in \mathbb{N}$, it holds*

$$f_n^* < \min_{y\in\mathcal{X}} \big[s_n(y) - \tau \|s_n\|_\infty \Delta_n^{\rho/2}(y)\big], \tag{4.7}$$

where τ, Δ_n and ρ are given as in Section 3.3.2, and that Algorithm 4.2 is not initialised infinitely many times. Then, the sequence $\{x_n\}$ is dense in $\mathcal{X}$.

Proof. The proof follows in the same way as the proof of Theorem 3.17, since the function μ_n in the iteration step of the algorithm is the same as in Gutmann's RBF method and the target values f_n^* are set according to the universal surface $s_n = \sum_{k=1}^{l} g(s_{n,k})$, with nonnegative function g. Thus, inequality (3.39) can be derived by assumption that $\{x_n\}$ is not dense in $\mathcal{X}$, for which contradiction is then shown by means of Lemmas 3.21 – 3.23. □

From Theorems 4.4 and 4.5, the first particular convergence result for the present RBF method is immediately obtained by setting f_n^* to the lowest value possible in an infinite number of iterations.

Corollary 4.6. *Let ϕ and m be as in Theorem 4.5, and let $f = \sum_{k=1}^{l} g(f_k)$, where g is a nonnegative continuous function and each f_k is continuous. Assume that, for infinitely many $n \in \mathbb{N}$, it holds $f_n^* = -\infty$, and that Algorithm 4.2 is not initialised infinitely many times. Then, Algorithm 4.3 converges.*

The second particular convergence result applies to the case in which the residual functions are in the associated native space of the radial basis function. In consequence, $\{\|s_n\|_\infty\}$ is uniformly bounded such that $\|s_n\|_\infty$ may be dropped from the right-hand side of (4.7).

Lemma 4.7. *Let $\{x_n\}$ be a sequence in $\mathcal{X}$ with pairwise different points such that $\{x_1, \ldots, x_{n_0}\}$ is $\mathcal{P}_m^d$-unisolvent. For $n \geq n_0$, let $s_n = \sum_{k=1}^{l} g(s_{n,k})$, where g is a nonnegative continuous function and each $s_{n,k}$ denotes the optimal interpolant to f_k at $x_1, \ldots, x_n$. Further, let $f_k \in \mathcal{N}_\phi(\mathcal{X})$, $k = 1, \ldots, l$. Then, $\|s_n\|_\infty$ is bounded above by a number that depends only on $x_1, \ldots, x_{n_0}$ and the functions g and f_k, $k = 1, \ldots, l$.*

Proof. According to Lemma 3.19, each $\|s_{n,k}\|_\infty$, $k = 1, \ldots, l$, is bounded above by a number depending on $x_1, \ldots, x_{n_0}$ and on f_k, such that the range of $s_{n,k}$ is a bounded closed set. Since g is a continuous function on that set, it follows that $g(s_{n,k})$ is bounded by a number that only depends on $x_1, \ldots, x_{n_0}$, on g and on the residual function f_k. Hence, s_n as the finite sum over all $g(s_{n,k})$ is also bounded above. □

Given Theorem 3.11 along with Lemma 4.7, Theorems 4.4 and 4.5 then provide the following corollary.

Corollary 4.8. *Let ϕ and m be as in Theorem 4.5. Further, let ν_d be as in Theorem 3.11, and let $f = \sum_{k=1}^{l} g(f_k)$ where g is a nonnegative continuous function and each $f_k \in C^{\nu_d}(\mathcal{X})$. Assume that, for infinitely many $n \in \mathbb{N}$, it holds*

$$f_n^* < \min_{y \in \mathcal{X}} \left[s_n(y) - \tau \Delta_n^{\rho/2}(y) \right],$$

where τ, Δ_n and ρ are given as in Section 3.3.2, and that Algorithm 4.2 is not initialised infinitely many times. Then, Algorithm 4.3 converges.

4.3 Numerical Analysis

In this section, we will assess the numerical aspects of the modified RBF method with extended local search on different data-fitting problems. We begin by comparing the method on relevant test problems with Gutmann's RBF method as well as with the extensions suggested to overcome the slow convergence of the original method. We then consider the calibration of the Hull-White model, where model prices are computed either (semi-)analytically or by the Monte Carlo SAA strategy, and the enhanced fitting of the Nelson-Siegel and the Svensson models, as outlined in Sections A.2 and A.3, respectively. In particular, for each of the considered cases, we describe the underlying data and present the numerical results in comparison with competing approaches or specifically selected methods.

Note that all numerical computations were carried out in MATLAB, Version 8.2 (R2013b), on the IRIDIS Compute Cluster provided by the University of Southampton.

Implementation of Method

To solve all optimisation problems by means of the modified RBF method with extended local search, we choose the thin plate spline radial basis function $\phi(r) = r^2 \log r$. This choice, however, is of less importance since the method directly interpolates the residual functions to build the response surfaces, thus already capturing the main structure of the objective functions. For each problem and if not further specified, we initialise the method at the corner points and the midpoint of the respective box-constrained parameter spaces $\mathcal{X}$, which is then transformed to the unit hypercube $[0,1]^d$ for the optimisation procedure.

Throughout our entire analysis, we minimise the response surfaces s_n and either $-\log v_n$ or $-\log h_n$ on $[0,1]^d$ by the DIRECT algorithm[21] of Jones et al. [1993], and additionally run the local solver 'fmincon' as provided by the optimisation toolbox in MATLAB from the best point found by DIRECT and the sample point with the lowest function value $f_n^{\min}$ obtained thus far. If not stated otherwise, we use DIRECT with a maximum number of function evaluations and iterations of 300 and 30, respectively, and the solver 'fmincon' with its default settings.

As for the choice of suitable target values f_n^*, we adopt the technique advocated by Gutmann [2001a] and set them in cycles according to some strategy, cf. Subsection 3.3.3.3. Specifically, we will adopt strategy II with cycle length 4 (i.e. $l^{\text{cy}} = 3$) and $\tilde{\Delta}_n^{(b)}$ for all considered problems, but replace the prefactor of $\tilde{\Delta}_n^{(b)}$ by the repeated sequence $(1, 0.25, 0.06, 0)$. This strategy is chosen as to account for the fact that the interpolation of residuals typically yields improved response surfaces with an implicit lower bound, which then facilitates to set the target values closer to the surface minima to enhance[22] the inherent local search mechanism.

[21] A publicly available source code of the algorithm for MATLAB is given, for instance, by Finkel [2004].

[22] If a strategy was chosen that is too global, then the advantage of an improved approximation could not be fully exploited and would thus result in a slower convergence.

Eventually, considering the extended local search described in Subsection 4.1.3, we classify inherently sampled points by the modified RBF method as local search points if they are generated within our adopted target value strategy by a prefactor of $\tilde{\Delta}_n^{(b)}$ that is less than or equal to 0.25, cf. Regis and Shoemaker [2007b]. These points then constitute the set $\mathcal{X}_n^{\text{loc}}$. Moreover, since we work on a unit hypercube, we initialise the extended local search in Algorithm 4.2 by its default values and use either 'fmincon' or 'pattern search' as local search method, depending on whether the objective function allows for the explicit use of gradients or not.

4.3.1 Relevant Test Problems

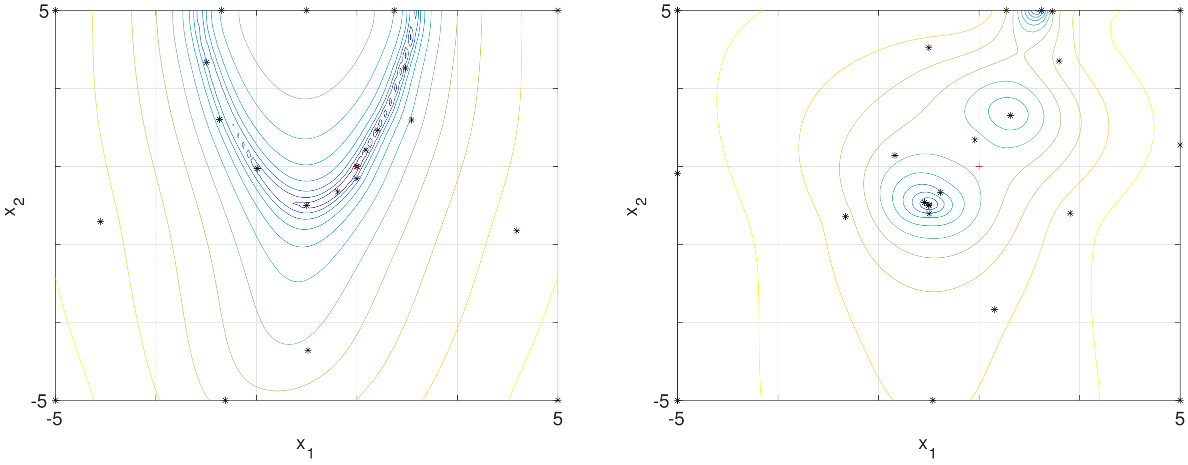
In order to assess the performance of the modified RBF method with extended local search and to point out its main advantages over the RBF method and its improvements on data-fitting problems, we first consider some relevant test problems. Specifically, we use a set of low-dimensional nonlinear least-squares problems from the collection proposed by Moré et al. [1981] for unconstrained optimisation, with commonly adopted bounds. The name of each function, its dimension, the domain as well as the number of known local minima and the global minimum on the domain are given in Table 4.1. Note that some of these problems may have a unique minimiser only, a situation which cannot be ruled out for most nonlinear least-squares problems from the outset unless a particular structure allows to show so. However, since our main concern is analysing the slow convergence of the RBF method caused by a malfunctioning local search, these problems still serve our purpose. In particular, any reasonable global optimisation algorithm is expected to cope with a unique minimiser, too, without a huge overhead over a purely local technique.

Function	Dimension	Domain	No. of local minima	f^*
Rosenbrock	2	$[-5,5]^2$	1	0
Freudenstein and Roth	2	$[-10,10]^2$	2	0
Beale	2	$[-4.5,4.5]^2$	2	0
Helical valley	3	$[-10,10]^3$	1	0
Bard	3	$[-0.25,0.25]\times[0.01,2.5]^2$	2	8.21×10^{-3}
Gulf	3	$[0.1,100]\times[0,25.6]\times[0,5]$	≥ 1	0
Box (3D)	3	$[0,2]\times[5,9.5]\times[0,20]$	≥ 2	0
Wood	4	$[-10,10]^4$	≥ 1	0
Brown and Dennis	4	$[-25,25]^4$	≥ 1	8.58×10^{5}
Kowalik and Osborn	4	$[-5,5]^4$	≥ 2	3.08×10^{-4}

Table 4.1: Test functions, their dimension, the domain, the number of local minima that are known at least, and the global minimum.

Illustrative Example

To illustrate the main effect of an improved interpolation via the residuals, Figure 4.1 shows the contour plots of the response surfaces that result from applying the (modified) RBF method on the Rosenbrock test function for 20 iterations, i.e. 25 function evaluations. Accordingly, it is clearly recognisable that the modified RBF method is able to capture the main structure of the objective function after only a few function evaluations and to already sample points in the banana-shaped region of attraction. By comparison, the RBF method exhibits considerable difficulties in assuming the mere form of the objective function and in finding new points along the promising region of the parameter space. Similar effects may also be observed for other test functions listed in Table 4.1, as well as for the objective functions arising from the calibration of the Hull-White model and the fitting of the Nelson-Siegel and Svensson models.



(a) Modified RBF method

(b) RBF method

Figure 4.1: Contour plots of the response surfaces (in plog-scale) and corresponding interpolation points after 20 iterations of the (modified) RBF method on the Rosenbrock function. The methods employ the same settings, where the RBF method additionally uses the plog-transformation on the objective function. The global minimum is indicated by a red plus.

The effect of additionally using an extended local search becomes especially apparent from Figure 4.1(b) for the RBF method. Here, the points sampled in a local search step of the method accumulate around minimisers of the response surfaces, but do not make any direct progress towards the global minimum situated in the elongated valley. This may thus be enforced by incorporating a local search method that is adequately initialised from a cluster of local search points and then fully explores the valley. Note that for the modified RBF method employing a local search technique may become negligible on problems with a very low dimension if the target values are well-adjusted, such as for the Rosenbrock test function. However, for problems of higher dimension, the technique may become indispensable.

Numerical Results

On each test problem of Table 4.1, we now compare the modified RBF method with extended local search (using 'fmincon') to the RBF method by Gutmann [2001a], as well as the suggested improvements by Regis and Shoemaker [2007b] in form of the improved strategy, by Holmström [2008] as the ARBF method, and by Locatelli and Schoen [2013] using knowledge of the lower bound zero. To be consistent among most of the methods and enable a fair comparison, we use our own implementation of the RBF method in MATLAB, also extended by the improved strategy and the suggested modification to include a known lower bound. We initialise these methods with the same setting as described above for the modified RBF method with extended local search, except that we use the target value strategy II with $l^{\text{cy}} = 4$, $\tilde{\Delta}_n^{(b)}$ and the implied prefactors. In particular, for the latter two extensions we use the default settings as given in the respective papers. As for the ARBF method, we use the implementation 'ARBFMIP' provided in the TOMLAB optimisation environment[23] with the default settings employed by the author to carry out his numerical tests. Eventually, to individually assess the impact of each of the two components of the modified RBF method with extended local search, we also consider the modified RBF method by itself and, similarly, the RBF method with the extended local search component. To avoid numerical difficulties arising from large differences in function values, the plog-transformation is used on all objective test functions and for all methods, except when residuals are interpolated and when knowledge of a lower bound is exploited (as otherwise the impact is very marginal).

We terminate each algorithm if the modified[24] relative error $|f_n^{\min} - f^*|/(1 + f^*)$ is smaller than 1%, where $f_n^{\min}$ denotes the current minimum function value and f^* is the known global minimum of the test function on its domain. The numbers of function evaluations required to reach this stopping criterion are reported in Table 4.2 for the respective methods. As a maximum number of function evaluations for all algorithms, we set $n^{\max} = 300$. If an algorithm exceeds this limit, then we indicate the respective case in the table by an asterisk.

	RB	FR	BE	HV	BA	GU	B3	WO	BD	KO
Modified RBF method	**16**	**51**	44	**36**	164	**24**	**12**	**40**	160	*
Modified RBF method with ext. LS	**16**	103	**43**	**36**	**91**	**24**	**12**	**40**	192	**190**
RBF method	229	*	47	*	191	93	212	*	**95**	*
RBF method with ext. LS	118	71	185	116	106	165	105	*	184	*
RBF method with impr. strategies	*	*	67	*	201	82	*	*	**95**	*
ARBF	184	165	45	*	174	84	187	*	112	*
RBF method with lower bound	109	226	113	*	151	87	236	*	*	*

Table 4.2: Number of function evaluations required by different methods on the test problems listed in Table 4.1.

[23] See `http://www.tomopt.com/tomlab/`.

[24] The modification is due to the fact that $f^* = 0$ for almost all test problems.

Table 4.2 shows that the modified RBF method (with extended local search) clearly outperforms the compared methods in terms of function evaluations to reach the stopping criterion, except for a single test instance. In particular, most instances can even be solved with a quite low number of evaluations by the method, whereas other methods require considerable more evaluations, not uncommonly exceeding the maximal number of evaluations (in almost all of these cases, the modified relative error is not even below 10%). This indicates that the method is indeed able to overcome the slow convergence caused by a malfunctioning local search, as observed for the original RBF method and its modifications on most data-fitting problems.

From the bottom three rows one may also observe that the suggested improvements for the RBF method have only a limited effect (if any at all) on more involved data-fitting problems, presumably due to elongated form of the valleys as exemplarily shown in Figure 4.1(a) for the Rosenbrock function. This may be explained as follows. Since the approach by Regis and Shoemaker [2007b] ensures a local search by restricting the minimisation of g_n to a small hyperrectangle around $\operatorname{argmin}_{y\in\mathcal{X}} s_n(y)$, it hardly affects the sampling of intrinsic yet ineffective local search points in the close vicinity of $\operatorname{argmin}_{y\in\mathcal{X}} s_n(y)$ but merely confines those sample points for which the restriction is still active even though the choice of f_n^* does not entail a local search anymore, i.e. where f_n^* is not close enough to $\min_{y\in\mathcal{X}} s_n(y)$. Latter points, however, are naturally sampled in non-descent direction and therefore do not contribute to a direct convergence of the method to local minimisers. Similarly, the approach by Holmström [2008] does not necessarily improve the slow convergence in elongated valleys either, as pre-sampled local search points produced by a range of sufficiently high target values in an iteration of the method tend to cluster around $\operatorname{argmin}_{y\in\mathcal{X}} s_n(y)$. Any selection of local search points for evaluation will thus only result in little progress compared to the effort invested, whereas points produced by lower target values sample in less explored regions of the parameter space. Finally, even though the approach by Cassioli and Schoen [2013] may improve the approximation and thus the selection of target values by knowledge of a lower bound, latter is not guaranteed to be sufficiently tight to avoid an overemphasised global search. Moreover, despite being adjusted to a lower bound, the constructed response surfaces are generally not capable of adequately capturing the inherent fitting structure to yield a significant improvement for our purposes.

4.3.2 Calibration of the Hull-White Model

In addition to the considered test problems, we will now investigate the applicability of the modified RBF method with extended local search on the calibration of the Hull-White model, as described in Section A.2. Specifically, we will first calibrate the Hull-White model under (semi-)analytical model prices to a series of market prices, where we compare the modified RBF with extended local search with the DIRECT algorithm by Jones et al. [1993] and the MCS algorithm by Huyer and Neumaier [1999], even if both are not designed to solved expensive problems. For a selected calibration date, we will then further assess the use of the SAA strategy for approximately solving the original calibration problem and show the effect of different sample sizes.

4.3.2.1 Data

To calibrate the Hull-White model in a traditional single-curve framework, we consider the time period from 1 January 2004 to 31 December 2014 and use payer swaptions of different contractual features as calibrations instruments. As additional tests have shown, the calibration via caps/floors essentially provides the same insights on our applied methods as the use of swaptions, such that we opt to omit this case. In order to compute swaption market prices, we take a cube of at-the-money Libor implied volatilities, spanned by the maturities 1m, 3m, 6m, 1y, 2y, 3y, 5y, 7y, 10y, 15y, 20y, 25y and 30y, and the tenors 1y–10y, 15y, 20y, 25y and 30y. The required data is retrieved from the market via Bloomberg L.P.[25] and then transformed to the respective vector of market prices C^{mkt} by Black's formula, see Section A.2.3 for further details. For the sake of simplicity, we take the 6-month Euribor curve from Bloomberg L.P.[26] as single reference rate for the floating rate in the underlying swaps and use a day-count convention of actual/actual for both swap legs. Eventually, note that, since we work in a single-curve framework, the Euribor curve is thus also used as term structure of interest rates, where any intermediate dates are inferred by linear interpolation. Overall, the entire data set for the calibration of the Hull-White model thus amounts to 2816 dates, each with 182 swaption market prices of different contractual features.

4.3.2.2 Numerical Results

Given above data set, we calibrate the Hull-White model in either form to the available market prices in the least-squares sense (without any transformation) and solve the resulting optimisation problems by the indicated method on the box constraints of Table 4.3. Note that because of the differentiability of the objective function on the parameter space, we use 'fmincon' as local solver for the modified RBF method with extended local search.

	κ	σ
LB	10^{-8}	10^{-8}
UB	5	0.5

Table 4.3: Lower (LB) and upper (UB) bounds for calibrating the Hull-White model to the given data set.

Due to the expensiveness of most problems, we terminate each optimisation after three minutes of computation time and visually assess the obtained solutions. In particular, we

[25] The respective Bloomberg tickers begin with 'EUSV' and terminate with the acronym indicating maturity and tenor, where we use the last quote 'Px_Last' of each day.

[26] The respective Bloomberg ticker is 'EUR006M Index'.

illustrate the optimal value $\bar{f}^*$ of a considered objective function $\bar{f}$ on a logarithmic scale in terms of the root-mean-square error (RMSE) measure $\sqrt{\frac{1}{l}\bar{f}^*}$. This may be interpreted as the average error in terms of basis points (bps).

Calibration using (Semi-)analytical Model Prices

To first compare the performance of the modified RBF method with extended local search to the DIRECT and MCS algorithms, we calibrate the Hull-White model under (semi-)analytical swaption prices to the given market data. For both comparing methods, we use the readily available MATLAB codes in their default settings, as provided by Finkel [2004] and Huyer and Neumaier [2000], respectively, but modify them in such a way that the methods are able to run for the prespecified time budget. The RMSEs and corresponding optimal parameters resulting from the calibrations of the Hull-White model are then depicted in Figure 4.2 for all three methods.

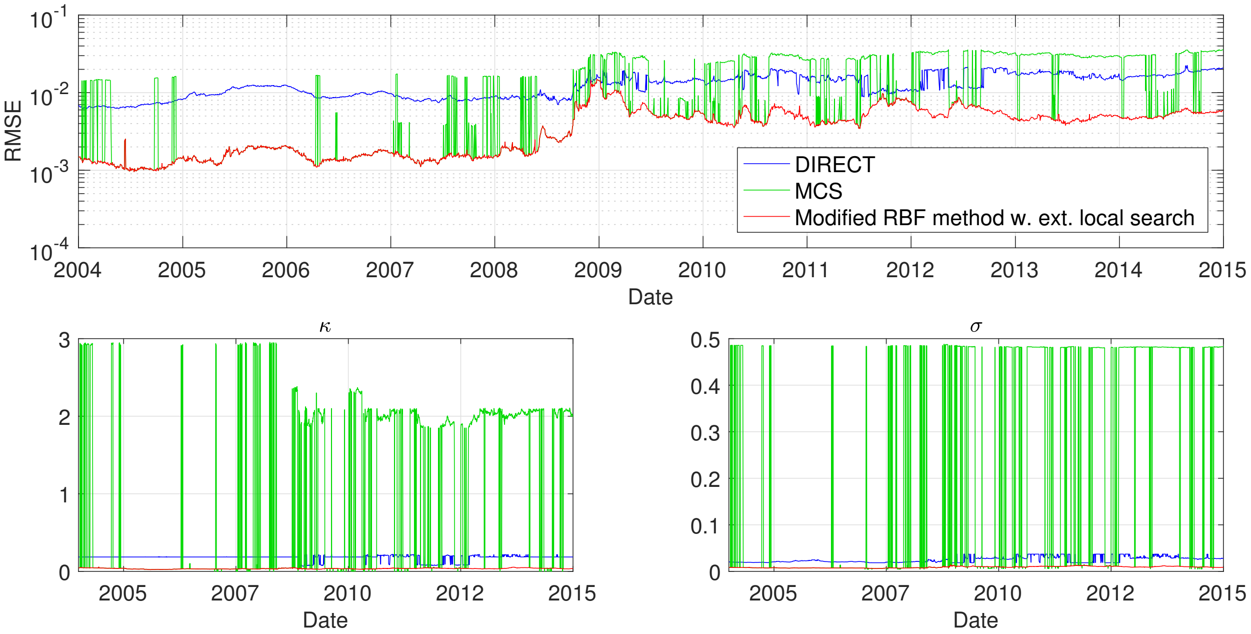


Figure 4.2: RMSEs and parameters obtained by calibrating the Hull-White model to the given data, using DIRECT, MCS, and the modified RBF method with extended local search.

Similar to the previous analysis on relevant test problems, it is observable that the modified RBF method with extended local search also outperforms the other methods on the calibration of the Hull-White model. In particular, one may recognise that DIRECT does not seem to be able to find the global minimum of any problem instance for the specified settings within the given time budget of three minutes (and in almost all of these cases not even if the budget is doubled, as further internal tests have shown), whereas MCS is at least able to detect the global minimum of about half of the problem instances. This may be attributed to the highly nonlinear data-fitting problem structure both methods are not able to cope with for a reasonable number of function evaluations. As a consequence, the RMSEs obtained by these methods are higher than those of the modified RBF method with extended local search, with the evolutions of calibrated parameters appearing irregularly oscillating.

Calibration using Monte Carlo Model Prices

In order to further compare the (semi-)analytical calibration of the Hull-White model with the calibration where model prices are computed by the SAA Monte Carlo strategy, we consider the calibration date as per 28 May 2009 and take the minimiser x^* and corresponding value f^* found by the modified RBF method with extended local search as reference. By fixing the sample sizes N at the values 10, 100, 1000 and 10000, respectively, and drawing for each sample size 100 random samples $Z_1, \ldots, Z_N$ with a different seed, we then construct by the SAA strategy a number of deterministic calibration problems that approximate the original problem with differing quality. Each problem is solved within the given time by the modified RBF method with extended local search, where we collect the best found point $\hat{x}_N^*$ and its corresponding function value $\hat{f}_N^*$ to compute the respective differences to the original counterparts by $\|\hat{x}_N^* - x^*\|_2$ and $|\sqrt{\frac{1}{l}\hat{f}_N^*} - \sqrt{\frac{1}{l}f^*}|$. The results obtained in this way are eventually presented in Figure 4.3 in form of box plots.

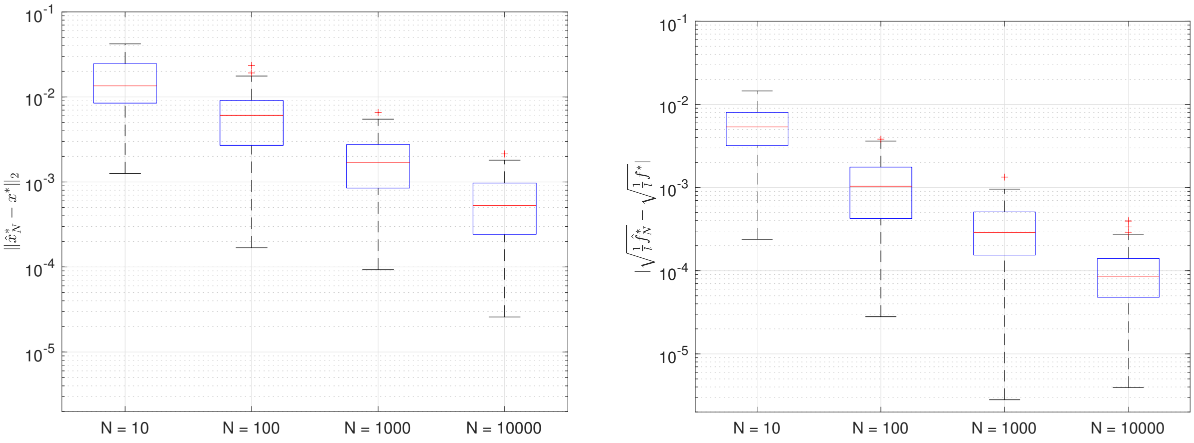


Figure 4.3: Box plots obtained by calibrating the Hull-White model under the SAA strategy with different sample sizes N by the modified RBF method with extended local search. For each sample size, the calibration problem is solved 100 times with a different seed, where a single optimisation procedure is stopped after a run time of three minutes.

Both subfigures underpin that the SAA strategy is able to yield fairly accurate approximating solutions to the original calibration of the Hull-White model, where the quality of approximation naturally increases with the sample size N. In particular, one may recognise that the rate of accuracy approximately follows $\sqrt{\mathrm{LLog}(N)}/\sqrt{N}$, as established in Section 2.2. Yet, as the time budget for carrying out an optimisation is limited, one should be aware that for considerably large sample sizes the modified RBF method with extended local search is not always able to perform the required number of function evaluations to locate a global minimum. Unlike as it is illustrated in Figure 4.3, one may then observe that the accuracy of the approximating optimal solutions deteriorates.

4.3.3 Fitting of the Nelson-Siegel and Svensson Models

Eventually, we also assess the performance of the modified RBF method with extended local search on the inexpensive fitting of the Nelson-Siegel and Svensson models. To this end, we will conduct a comprehensive computational study in which we show that the fitting of both models via the penalty approach, as described in Section A.3, and by use of our proposed method is able to compete with existing methods in terms of solution quality.

4.3.3.1 Data

To fit the models to market data, zero rates $y_1^{\text{mkt}}, \ldots, y_l^{\text{mkt}}$ are required which, since not directly observable in the market, are constructed from instruments that are actively traded in the market. For the current analysis, we therefore retrieve daily Euro par swap rates with maturities from one to 15 years for the time period from 1 January 2004 to 31 December 2014 from Bloomberg L.P.[27] and convert them into the corresponding zero rates by the usual bootstrapping technique via formula (A.32), see also Hirsa [2012], Section 7.7, for instance. The resulting data set consists of 2769 daily zero rate curves with 15 maturities each, to which the models are fitted, see Figure 4.4 for an illustration of the zero rate curves.

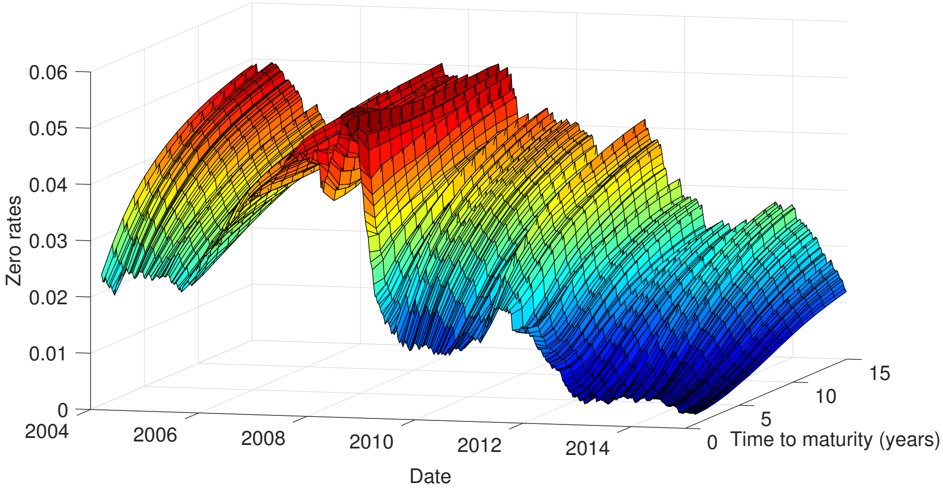


Figure 4.4: Market zero rates constructed from swap par rates of every fifth business day for the time period from 1 January 2004 to 31 December 2014.

[27] The respective Bloomberg tickers are 'EUSA1 CMPN Curncy', 'EUSA2 CMPN Curncy',..., 'EUSA15 CMPN Curncy', where we use the last quote 'Px_Last' of each day.

4.3.3.2 Numerical Results

To solve the series of fitting problems with different input data, we minimise f^{pen} (with $\eta^{\text{pen}} = 10^{-6}$ and $\kappa^{\max} = 100$ and $\kappa^{\max} = 180$ for the Nelson-Siegel and the Svensson model, respectively) by the modified RBF method with extended local search on the box constraints given in Table 4.4. However, in addition to the set of corner points and the midpoint of the box-constrained parameter space, we construct the initial response surface of the method on the best point of the previous fit, and, when indicated, on further points that may be supplied through knowledge on the problem structure. This may then instantly direct the method towards promising regions of the parameter space. Due to the nondifferentiability of the objective function, we use 'pattern search' as local solver and stop the method if the number of function evaluations (not including evaluating initial and local search points) exceeds the predefined values $n^{\max} = 50$ and $n^{\max} = 150$ in the case of the Nelson-Siegel and the Svensson model, respectively.

	λ_1
LB	10^{-3}
UB	5

(a) Nelson-Siegel model

	λ_1	λ_2
LB	10^{-4}	10^{-8}
UB	4	15

(b) Svensson model

Table 4.4: Lower (LB) and upper (UB) bounds for fitting the Nelson-Siegel and the Svensson models to the given data set using f^{pen}.

In order to compare our enhanced approach with already existing methods discussed in Subsection A.3.1.1, we divide the latter into two main classes:

(a) methods minimising the objective function f,

(b) methods minimising the objective function f^{sep} and f^{pen},

and equally fit them to the above described data set. All obtained results are then analysed in terms of model fit and solution quality, where we visualise the time series of fitting errors on a logarithmic scale by using the monotone RMSE measure $\sqrt{\frac{1}{l}\bar{f}(\lambda^*, \beta^*)}$, with $\bar{f}(\lambda^*, \beta^*)$ denoting the minimum objective function value of the considered approach.

Comparison with Other Methods Minimising f

We first compare our approach with the minimisation of the objective function f, as suggested by De Pooter [2007], Gilli et al. [2010] and Gauthier and Simonato [2012] using different methodologies. Specifically, we consider differential evolution and a multi-start strategy as competing approaches, since De Pooter's suggestion describes a purely local procedure which is outperformed by a multi-start procedure with suitably chosen initial points.

For comparison with the differential evolution heuristic proposed by Gilli et al. [2010], we adopt the relevant code from the appendix of their paper, with minor adjustments to allow for negative interest rates. Similar to Gauthier and Simonato [2012], we further consider a multi-start strategy in which f is first evaluated at an equidistant grid of points, from which then a subset of points with small function values is chosen to initialise the local solver. In this way, the strategy aims at covering all regions of the parameter space as best as possible while at the same time taking into account the higher dimension. To have each dimension of the constrained parameter space covered by a sufficient number of points without exceeding the computation times of our proposed approach, we evaluate f at 5^4 resp. 4^6 equidistantly distributed grid points, from which the best 10 resp. 50 are selected in the case of the Nelson-Siegel and Svensson models, respectively. To avoid parameter regions leading to severe multicollinearity in both approaches and enable a fair comparison with the penalty method, we have taken the original constraints of each of the alternative approaches, adjusted them to the present parameterisation and restricted the nonlinear parameters such that they approximately correspond to the penalisation level implied by the penalty approach and no optima obtained by the penalty approach are cut off by these constraints. The resulting constraints are reported in Table 4.5, including the special case of the Svensson model where $\lambda_1 \approx \lambda_2$.

	β_1	β_2	β_3	λ_1
LB	-0.15	-0.3	-0.3	0.1
UB	0.15	0.3	0.3	5

(a) Nelson-Siegel model

	β_1	β_2	β_3	β_4	λ_1	λ_2
LB	-0.15	-0.3	-0.3	-0.3	0.01	0.01
UB	0.15	0.3	0.3	0.3	4	15

(b) Svensson model

Table 4.5: Lower (LB) and upper (UB) bounds for fitting the Nelson-Siegel and the Svensson models to the given data set using f. For the Svensson model, the nonlinear constraint $|\lambda_1 - \lambda_2| \geq 0.2$ is additionally imposed to prevent $\lambda_1 \approx \lambda_2$.

The root-mean-square errors and condition numbers resulting from the fit of the Nelson-Siegel and the Svensson models are shown for all three approaches in Figure 4.5 and Figure 4.6, respectively.

From Figure 4.5, it can be seen that the approaches attain the same local minimum, most likely the unique global minimum, at almost all fitting dates, with exception of a few dates where either the multi-start strategy or differential evolution, or both, seem to find different local minima with larger objective function values (see 06/2004 – 06/2005, 06/2007 and 06/2008). In particular, the larger objective function values in the latter period are accompanied by extreme condition numbers underpinning the bad quality of the obtained optimal solutions.

Figure 4.6 visualises the fitting errors and condition numbers of the Svensson model. It shows that the discrepancy in model fit and solution optimality among the approaches is

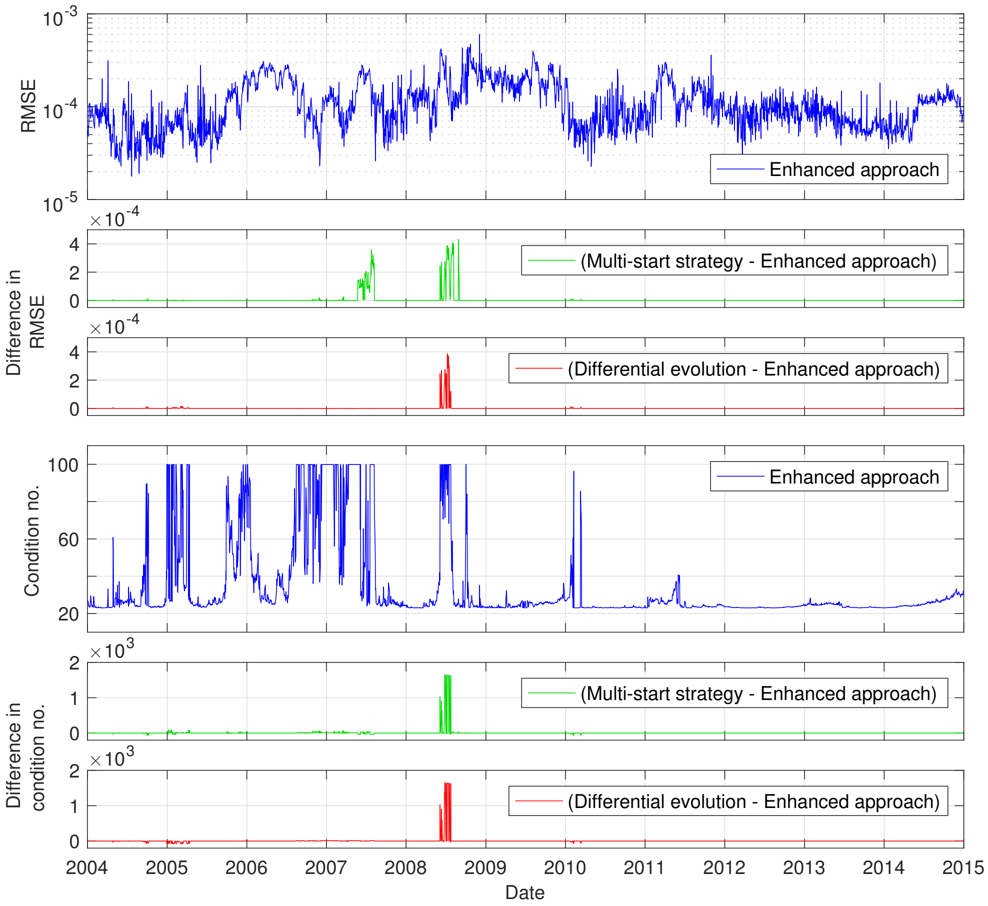


Figure 4.5: RMSEs and condition numbers resulting from fitting the Nelson-Siegel model to the given data, using a multi-start strategy, a differential evolution to minimise f, and our proposed approach.

noticeably larger than for the Nelson-Siegel model, due to the increased dimension of the problem and the considerable difference in handling multicollinearity. Again, the proposed method outperforms the multi-start strategy and differential evolution throughout the considered time period, provided that the conditioning of the optimal solution is reasonable. In particular, as plots of the condition numbers show, the optimal solutions of the proposed method appear to be more stable than their counterparts while at the same time a better fit can be obtained in most cases. Note that one has to bear in mind that for reasons of comparison, the constraints for minimising f have been adjusted using our knowledge of the stability analysis in Subsection A.3.2. A separate analysis has shown that the original approaches as mentioned in the literature with different model parameterisations and constraints either yield a worse fit or a more sensitive optimal solution than the ones obtained in the comparison.

Plots of the fitting errors also indicate that, for the given data, the objective function values only differ marginally, with differences below 1 bp, even though different optimal

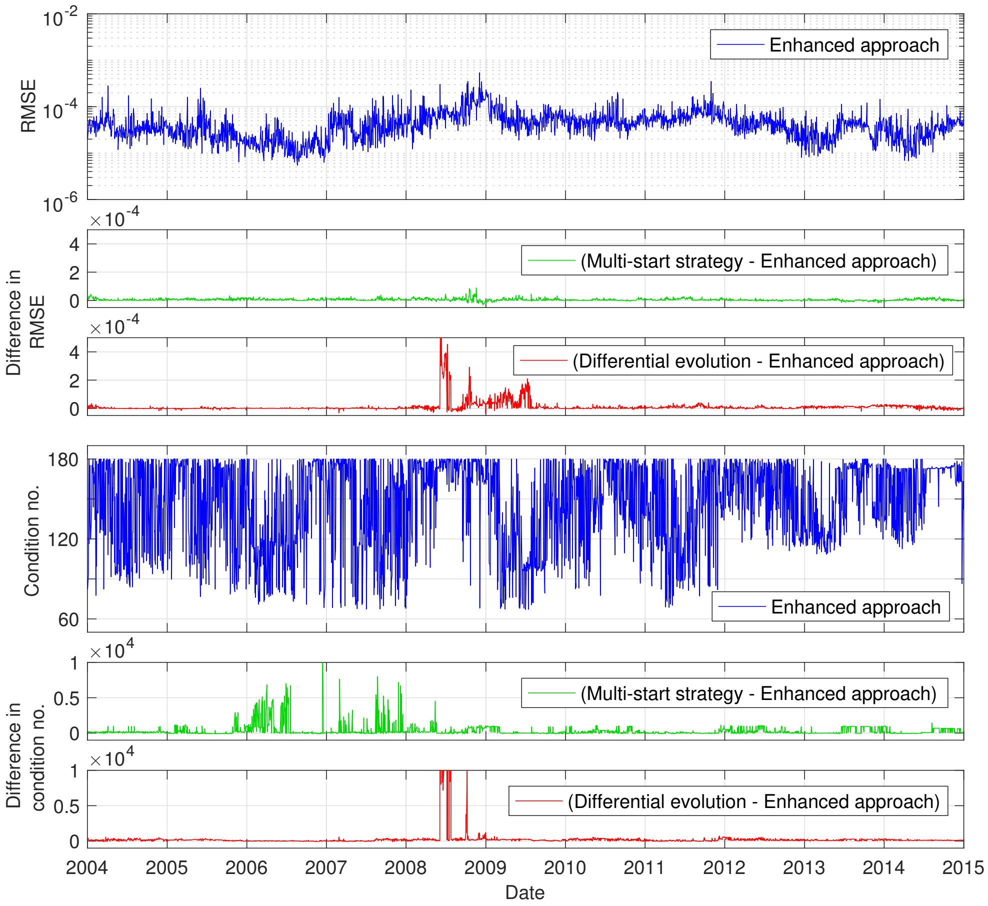


Figure 4.6: RMSEs and condition numbers resulting from fitting the Svensson model to the given data, using a multi-start strategy, a differential evolution to minimise f, and our proposed approach.

solutions are obtained.

Comparison with Methods Minimising f^{sep} and f^{pen}

We will now also consider methods that take into account separability in the objective function and compare them with our proposed technique. To this category belong grid search methods, as examined by Nelson and Siegel [1987] and Annaert et al. [2013], strategies in which the nonlinear parameter is fixed, see Fabozzi et al. [2005], Diebold and Li [2006], De Pooter [2007] and De Rezende [2011], as well as the multi-start strategy described by Gauthier and Simonato [2012], which all aim at minimising the objective f^{sep}. Since, by construction, any grid with corresponding function values may be incorporated into the initial surface of an RBF method, we can omit grid search methods as well as the strategy of fixing nonlinear parameters in the sequel. Instead, we apply a multi-start strategy to minimise the function f^{sep} as well as its penalised variant f^{pen}. This approach then also assesses the performance of our modified RBF method with extended local search more directly.

Because of the reduced dimension of the parameter space, we follow a multi-start strategy that employs a local solver from an equidistant grid of initial points, without any preselection by function evaluations. Given the nondifferentiability of the objective function f^{pen}, the local solver 'patternsearch' is used in the multi-start framework for minimising f^{pen}, whereas 'fmincon' is applied to f^{sep}.[28] To better compare computation times between these multi-start approaches and our approach, we set the grid sizes to 15^1 resp. 10^2 points for minimising f^{sep} and to 15^1 resp. 6^2 for minimising f^{pen} in the fitting of the Nelson-Siegel and the Svensson models, respectively. Exploiting the structure of the RBF method, we also use the initial grid of the multi-start strategy for minimising f^{pen} in the construction of the initial response surface. Numerical issues in the minimisation of f^{sep} arising from multicollinearity are dealt with by choosing the constraints on the nonlinear parameters in the same way as specified in Table 4.5, while again Table 4.4 applies to the minimisation of f^{pen} with multi-starts.

Using f^{sep} and f^{pen}, the fit of the Nelson-Siegel model is carried out in one-dimensional parameter spaces, which are covered well by the specified grid of initial points. Moreover, in the case of minimising f^{sep}, constraints have been determined so that they correspond closely to the level of allowed condition number of the penalisation. It is thus not surprising that the multi-start strategies yield about the same fitting errors and stability of optimal solutions as our proposed approach, cf. Figure 4.5.

To examine the fitting results of the Svensson model for the various approaches, we refer to the errors and the corresponding condition numbers of the matrix $\Psi(\lambda^*)$ as illustrated in Figure 4.7.

Accordingly, applying a multi-start strategy to minimise f^{sep} yields similar fitting errors as the technique proposed here, except in the time periods 2008 and 2013 – 2015, where the RMSE errors of multi-start strategy on f^{sep} are noticeably lower. However, plotting the relevant condition numbers reveals that the corresponding minimisers are significantly less stable for the multi-start strategy than for the penalised approach. It appears that suitable constraints guaranteeing a predetermined level of stability are difficult to obtain for the Svensson model. Hence, we conclude that on the given data set our approach is able to improve both the model fit and the stability of optimal parameters over a comparable multi-start strategy for minimising f^{sep}.

To assess the effectiveness of our chosen global solver, we compare the fitting errors produced by a multi-start strategy minimising f^{pen} and errors produced by our proposed method. As it can be seen from the corresponding graph, multi-start does not necessarily yield model fits as good as those obtained by the modified RBF method with local search, even when both methods use the same initial grid.

[28] As the rank of $\Psi(\lambda)$ is constant within the constraints set on the nonlinear parameters, f^{sep} is differentiable on the resulting parameter space, see Subsection A.3.1.2. Thus, 'fmincon' may be applied as well.

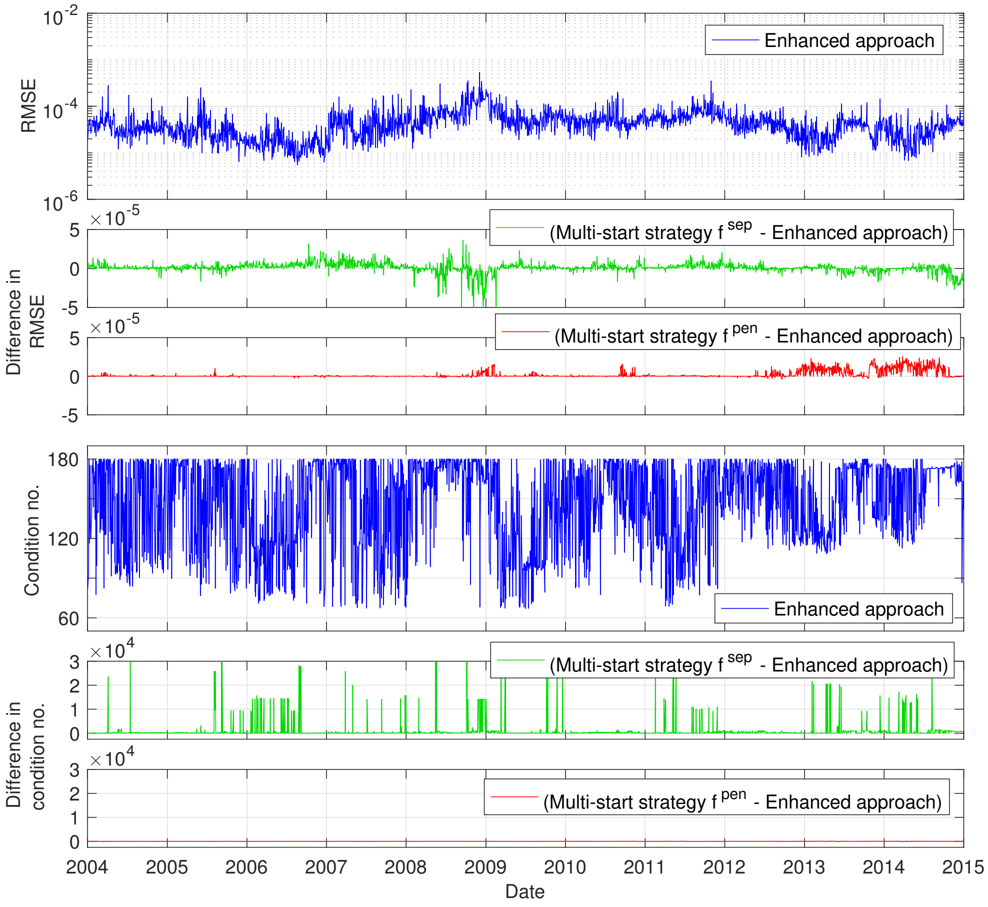


Figure 4.7: RMSEs and condition numbers resulting from fitting the Svensson model to the given data, using a multi-start strategy to minimise f^{sep} and f^{pen}, and our proposed approach.

Chapter 5

Global Optimisation of Expensive, Noisy Objective Functions

Thus far, we have considered in this thesis the global optimisation of an expensive objective function f on a compact set $\mathcal{X} \subset \mathbb{R}^d$, where function evaluations have been exact. This included the case where a general deterministic objective function f is minimised itself, but also (by abuse of notation) where a deterministic approximation $\hat{f}_N$ to an underlying objective function f is minimised within the SAA strategy for some fixed number of Monte Carlo simulations N. However, if we now seek to minimise a noisy objective function $\hat{f}$, such as in the case of the VSAA strategy where the underlying problem of minimising an unknown objective function f is approximately solved through a sequence of random objective functions $\{\hat{f}_{N_k}\}$ with schedule $\{N_k\}$, then the problem at hand can no longer be tackled in the same manner. In particular, the fact that the observed function values are perturbed by noise renders conventional response surface methods for deterministic objectives unreliable or even unworkable, and thus requires more suitable methods that are able to deal with the presence of noise.

Yet, even though the need for appropriate response surface methods in the presence of noise arises from many practical situations, the available literature on these kind of methods is scarce compared to their exact counterparts, cf. the below overview in Section 5.1. In particular, almost all of the available methods fall into the category of Bayesian and regression-based approaches which, due to their underlying probabilistic concept, offer an appealing framework for an extension to noisy functions, see, e.g., Huang et al. [2006] and Villemonteix et al. [2009]. As for response surface methods that explicitly rely on radial basis functions, we are only aware of the method proposed by Jakobsson et al. [2010]. It constructs suitable approximants by minimising a tradeoff between the semi-norm of the surface and its residual sum of squares to the noisy observations, while new evaluation points are determined by a quality function that depends on the distance to already evaluated points and corresponding response surface values. Further related contributions exist, but are merely concerned with establishing the similarity between the P-algorithm and Gutmann's RBF method in the presence of noise (Žilinskas [2010]), or with enhancing the practical convergence of the RBF method if noisy but less expensive function evaluations are additionally available (Costa and Nannicini [2014]). Eventually, it has to be noted that no proof of convergence for any of the available response surface methods dealing with noisy objective functions seems to exist in the literature, to the best of our knowledge.

Given these contributions and motivated by our financial application for which no readily available method seems to be known, we present in this chapter a *RBF method for noisy objective functions* in which the level of noise may be controlled by means of pointwise error bounds. The method is essentially based on Gutmann's original RBF method for deterministic objective functions, providing a mathematically solid and numerically robust framework, and takes up some of the initial ideas used by Žilinskas [2010] for extending Gutmann's method in order to prove its similarity to the P-algorithm in a noisy setup. Specifically, in establishing the method, we mainly addresses the two main components of response surface methods that require modification to be able to handle noisy function values: the construction of adequate response surfaces and the determination of new evaluation points once a surface has been constructed. Since radial basis function interpolation is no longer feasible in the present situation, we first consider common possibilities for the approximation of a noisy function by means of radial basis function and discuss their suitability for integration into a response surface methods. As regularised least-squares approximants explicitly seek to balance between the smoothness of the surface and the closeness to the data, where the additional regularisation parameter may be set in accordance with the available error bounds, they turn out to be particularly suited for our purposes. Moreover, the least-squares criterion allows for a convenient adaption of Jones's general technique to determine new evaluation points through target values, by analogy with Gutmann's original proceeding. In particular, this functionality then also facilitates to establish convergence of method, where we show that the convergence properties of Gutmann's deterministic method are kept when the exact function values are replaced by corresponding noisy values, albeit under some simplified assumption on the error bounds.

In continuation to Section 3.1 on response surface methods, we begin in Section 5.1 by giving a more detailed overview of response surface methods in the presence of noise. Using previous results on radial basis function interpolation, Section 5.2 discusses the construction of different radial basis function approximants for integration into a response surface method, given the availability of error bounds. By means of regularised least-squares approximants, we then present in Section 5.3 a RBF method that is able to deal with noisy objective functions. We give a detailed description of the method, state and prove several convergence results that hold under a simplistic assumption on the error bounds, and discuss practical aspects for the implementation of the method. Eventually, we illustrate the practical applicability of the method on relevant test problems and by calibrating the Hull-White model under the VSAA strategy, where we also assess the use of different sample size schedules in comparison to the SAA strategy.

5.1 Overview of Related Literature

Despite its importance in application, the global optimisation of expensive objective functions in the presence of noise has attracted considerably less attention than the equivalent optimisation without noise, see Subsection 3.1.2. Nevertheless, a few contributions have been made in the literature, which are mentioned below according to their main characteristics.

Because of the close relation between *Bayesian methods* and *regression-based methods*, also leading to the same results for the canonical choice of Gaussian processes in a noisy setup (e.g., by transferring the technique for exact function values as suggested by Fowkes [2011]), we will now treat both approaches jointly. Specifically, since the methods from these classes are typically formulated as Gaussian process regressions, we will start by outlining the regression-based approach. Subsequently, as no applicable noisy method using Jones's general technique is known, we will then describe response surface methods and techniques that explicitly work with radial basis functions but apply different strategies for determining new evaluation points.

Regression-Based Methods

Due to the underlying probabilistic framework that is provided by response surface methods based on regression, these type of methods may be straightforwardly extended to noisy function values, see, e.g., Schonlau [1997]. Specifically, according to the noise-free setup in Section 3.1.2, the true objective function f is still assumed to be a realisation of a stochastic process $\{F(x)\}$ in form of

$$F(x) = \sum_{j=1}^{\widetilde{m}} c_j p_j(x) + Z(x), \quad x \in \mathcal{X},$$

with regression functions p_j, unknown parameters c_j, and a stochastic process $\{Z(x)\}$ whose covariance structure is given by $\mathbb{C}\mathrm{ov}^{\mathbb{P}}(Z(x), Z(y)) = \sigma_Z^2 \Gamma_Z(x,y)$, $x, y \in \mathcal{X}$, for a process variance σ_Z^2 and some correlation function $\Gamma_Z(\,\cdot\,,\cdot\,)$. However, by observing the noisy objective function values $\hat{f}(x_1), \ldots, \hat{f}(x_n)$, each contaminated by some additive and i.i.d. normally distributed random error ϵ_i with mean zero and unknown variance σ_ϵ^2, the resulting linear regression model now becomes

$$\hat{f}_X = Pc + z + \epsilon,$$

where $\hat{f}_X = (\hat{f}(x_1, \ldots, \hat{f}(x_n))^\top$, $P \in \mathbb{R}^{n \times \widetilde{m}}$ is the design matrix, $c = (c_1, \ldots, c_{\widetilde{m}})^\top$ the parameter vector, $z = (Z(x_1), \ldots, Z(x_n))^\top$ and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^\top$.

With this modification, calculating the best linear unbiased predictor $s_n(x) = a(x)^\top \hat{f}_X$ for f based on the observations $\hat{f}(x_1), \ldots, \hat{f}(x_n)$ then proceeds on the same lines as for noise-free observations. In fact, minimising the MSE (3.5) subject to the same unbiased constraint, the predictor can be shown to have the expression

$$s_n(x) = \pi(x)^\top \hat{c} + r(x)^\top \big(\widetilde{\Gamma} + \sigma_\epsilon^2/\sigma_Z^2 I\big)^{-1} (\hat{f}_X - P\hat{c}), \tag{5.1}$$

where $\pi(x) = (p_1(x), \ldots, p_{\widetilde{m}}(x))^\top \in \mathbb{R}^{\widetilde{m}}$, $\hat{c} = (P^\top(\widetilde{\Gamma} + \sigma_\epsilon^2/\sigma_Z^2 I)^{-1}P)^{-1}P^\top(\widetilde{\Gamma} + \sigma_\epsilon^2/\sigma_Z^2 I)^{-1}\hat{f}_X$ is the generalised least-squares estimate of c, $\widetilde{\Gamma} \in \mathbb{R}^{n \times n}$ denotes the correlation matrix with $\widetilde{\Gamma}_{ij} = \Gamma_Z(x_i, x_j)$, $I \in \mathbb{R}^{n \times n}$ the identity matrix, and $r(x) = (\Gamma_Z(x, x_1), \ldots, \Gamma_Z(x, x_n))^\top \in \mathbb{R}^n$. The MSE of the prediction can be obtained as

$$v_n(x) = \sigma_Y^2 \left[1 - \begin{pmatrix} r(x) \\ \pi(x) \end{pmatrix}^\top \begin{pmatrix} \widetilde{\Gamma} + \sigma_\epsilon^2/\sigma_Z^2 I & P \\ P^\top & 0 \end{pmatrix}^{-1} \begin{pmatrix} r(x) \\ \pi(x) \end{pmatrix} \right], \tag{5.2}$$

noting that the additional parameters σ_Z^2 and σ_ϵ^2 are typically estimated iteratively by cross validation or maximum likelihood to the given data. In particular, due to the use of noisy observed function values, the resulting prediction s_n is no longer interpolative at the data points x_i, $i = 1, \ldots, n$. Also, the value of the MSE v_n at the x_i is not zero anymore.

Using above probabilistic framework, Huang et al. [2006] propose to extend the EGO algorithm by Jones et al. [1998] to noisy objective function values where errors in the evaluations are assumed to be additive and i.i.d. normally distributed. The suggested method, called *Sequential Kriging Optimisation (SKO)* method, further considers $\{Z(x)\}$ as a Gaussian process, whose correlation function $\Gamma_Z(\,\cdot\,,\cdot\,)$ is specified by (3.6), and then constructs response surfaces and MSEs in the form of (5.1) and (5.2), respectively. To select a new evaluation point x_{n+1}, an *augmented expected improvement* criterion is derived that takes into account the additional random noise in the observed function values by maximising the function

$$u_n(x) = \mathbb{E}^{\mathbb{P}}\big[\max\{s_n(x_n^{\min}) - F(x), 0\}\big]\left(1 - \frac{\sigma_\epsilon}{\sqrt{v_n(x) + \sigma_\epsilon^2}}\right), \tag{5.3}$$

where $x_n^{\min} := \operatorname{argmin}_{1 \le i \le n}\{s_n(x_i) + \tau\sqrt{v_n(x_i)}\}$ stands for the current 'effective best solution', and τ denotes a constant reflecting the degree of risk aversion. In particular, the expectation in (5.3) thus reduces in the same way as the ordinary expected improvement to expression (3.3), where $f_n^{\min}$ is replaced by $s_n(x_n^{\min})$.

A further extension to the expected improvement criterion that may be used in above setup for noisy objective function values is suggested by Gramacy and Lee [2011]. Their idea is to consider the expected improvement at a reference point $y \in \mathcal{X}$ under the model s_n, given that the candidate point $x \in \mathcal{X}$ is added to the set of data points, i.e. the conditional expected improvement $\mathbb{E}^{\mathbb{P}}[\max\{f^{\min} - F(y \,|\, x), 0\}]$, where $f^{\min} = \min_{x \in \mathcal{X}} \mathbb{E}^{\mathbb{P}}[F(x)]$. By choosing a density function $\rho_{\mathcal{X}}(y)$ over $y \in \mathcal{X}$, e.g., a uniform density over a bounded domain, a new point x_{n+1} is then found by maximising the *integrated expected conditional improvement*

$$u_n(x) = -\int_{\mathcal{X}} \mathbb{E}^{\mathbb{P}}[\max\{f^{\min} - F(y \,|\, x), 0\}]\, \rho_{\mathcal{X}}(y)\, \mathrm{d}y, \quad x \in \mathcal{X}.$$

Eventually, in Villemonteix et al. [2009], the authors also show that their IAOG method for exact function values, based on the conditional entropy of a minimiser for determining new evaluation points, can be extended to handle noisy observations. Specifically, they assume in their Gaussian process model that errors in the observed function values are additive and i.i.d. normally distributed with known mean and variance, such that the resulting predictor can be built in known fashion as above. Moreover, to find new evaluation the conditional entropy criterion of a minimiser essentially remains the same except that its estimation requires the conditional simulations to be carried out on the noisy observed function functions.

Response Surface Methods/Techniques Based on Radial Basis Functions

In his theoretical paper, Žilinskas [2010] addresses the similarity between the P-algorithm and Gutmann's RBF method, as pointed out in Gutmann [2001a], and shows that it also extends to the presence of noise if appropriate modifications are made in both algorithms. In particular, for the RBF method, he suggests to construct response surfaces by means of radial basis functions through minimising the 'bumpiness', subject to the condition that the residual sum of squares of the surface to the noisy observations is proportional to the variance of the involved additive noise, which is assumed to be constant and known. New evaluation points may then be determined similar to the deterministic case by means of target values, i.e. by minimising the 'bumpiness' of an augmented surface such that it interpolates a chosen target value and such that the residual sum of squares of the augmented surface to the noisy observations is proportional to the known variance. However, even though Žilinskas focuses on establishing the theoretical similarity between the P-algorithm and the RBF method in a noisy setup, no explicit algorithm making use of this result is proposed.

Based on radial basis functions, Jakobsson et al. [2010] present an algorithm, called *qualSolve*, for the global optimisation of expensive black-box functions subject to noise. Here, suitable response surfaces are constructed by minimising the convex sum of the 'bumpiness' of a surface and its sum of squares to the noisy observations, where the additionally introduced parameter to balance between both measures is estimated via a leave-one-out cross validation procedure. Moreover, to select new evaluation points a quality function is maximised, which is calculated at each point in the domain by the minimum distance to previously evaluated points and weighted by the value of the response surface at that point. In particular, the weights are adjusted periodically in order to alternate between local and global search, thus reflecting the mechanism of conventional response surface methods. The authors also extend the algorithm to optimise multi-objective functions.

Eventually, even though not a method for noisy functions but merely a technique to speed up the practical convergence of Gutmann's RBF method in case noisy and less expensive function values are additionally available is the approach by Costa and Nannicini [2014]. In particular, given that the additional noisy function values lie in a reasonable error bound around their unknown counterparts, a response surface is constructed by minimising the 'bumpiness' subject to the constraints that the exact values are interpolated and the difference to noisy values is within the corresponding error bounds. New evaluation points are then found by Gutmann's conventional strategy in which target values are set, albeit without any modification to account for the added noisy values.

5.2 Radial Basis Function Approximation

The investigated response surface methods thus far were based on the assumption of observing exact objective function values, such that interpolation naturally provided the method of choice for approximation, see Section 3.2. If the observed objective function values are contaminated by noise or other forms of inaccuracies, however, then other approximation techniques are needed in order to adequately reconstruct the underlying objective function. Specifically, in these cases, an interpolation is not suitable anymore as it would result in a 'wrong model' to the available function values in the first place; since the values are noisy, they are almost surely not the correct values we are looking for. Moreover, by interpolating noisy function values, too much weight would be given to the involved noise, which may easily lead to a model overfitting the data and becoming unnecessarily oscillating, thus corresponding poorly to the underlying true objective function.

Unlike in the case of interpolation, there exist various possibilities to approximate a set of noisy function values, where a proper choice depends mostly on the nature of the available data and on the intended use of the resulting approximant. Since our purpose is to integrate the approximation technique into a response surface method, we are therefore not only interested in a technique that approximates the noisy function values adequately, but which is also able to reliably build a sequence of response surfaces along the iterations of the method and which allows for convenient selection of new evaluation points. In any case, though, we will continue to use approximants of the generic form (3.7), along with their associated radial basis function spaces, as these as such also have proven to be highly useful for approximation purposes, see, e.g., Buhmann [2003], Iske [2004] or Wendland [2005a].

To address the topic of radial basis function approximation in the remainder of this section, we begin in Subsection 5.2.1 by briefly describing the prevailing approximation problem motivated by our main application. Subsection 5.2.2 then discusses several possibilities to construct radial basis function approximants within this setup and their suitability for integration into a response surface method.

5.2.1 Approximation Problem

For deriving a suitable approximant, let $\hat{f}(x_1), \ldots, \hat{f}(x_n)$ be some noisy function values of some unknown underlying deterministic function $f : \mathbb{R}^d \to \mathbb{R}$ at the pairwise distinct data points $x_1, \ldots, x_n$. By means of a closer specification of the approximate relation

$$s(x_i) \approx \hat{f}(x_i), \qquad i = 1, \ldots, n, \tag{5.4}$$

our main objective in the sequel is then to find an approximant s to the data $(x_1, \hat{f}(x_1)), \ldots, (x_n, \hat{f}(x_n))$ such that the unknown function f is recovered most suitably. In particular, thus note that while in an interpolation the error between model and function values vanishes at the sample points x_i, cf. condition (3.8), an approximation scheme typically induces a small error at the data sites.

To further specify the approximate relation (5.4), we assume that some positive error bounds $\epsilon_i = \epsilon(x_i)$ are available along with the noisy function values $\hat{f}(x_1), \ldots, \hat{f}(x_n)$, such that the latter differ by at most ϵ_i from the true but unknown function f at the evaluated points $x_1, \ldots, x_n$, respectively. Put differently, we thus have

$$\left|f(x_i) - \hat{f}(x_i)\right| \leq \epsilon_i, \qquad i = 1, \ldots, n, \tag{5.5}$$

for some error bounds $\epsilon_i > 0$. In particular, assumption (5.5) may then provide useful guidance in constructing a suitable approximant s from the noisy observations $\hat{f}(x_1), \ldots, \hat{f}(x_n)$: since s aims at recovering the unknown function f by approximating $\hat{f}(x_1), \ldots, \hat{f}(x_n)$, it is reasonable to also require that

$$\left|s(x_i) - \hat{f}(x_i)\right| \leq \epsilon_i, \qquad i = 1, \ldots, n,$$

is satisfied in some way that has to be defined more closely.

5.2.2 Construction of Approximants

Given the prevailing approximation problem, we now review the most common approaches for radial basis function approximation, with the main objective of finding an approach which best suits for an incorporation into a response surface method for noisy objective functions. A useful overview of different approximation techniques can be found in Fasshauer [2007], for instance, whereas Locatelli and Schoen [2013], Section 3.2.4, very briefly discuss this issue within the context of response surface optimisation. Moreover, since all of the approaches presented below appear in one form or another in some application, we additionally refer the interested reader to the literature on learning theory (e.g., Girosi [1992]), regularisation networks (e.g., Evgeniou et al. [2000]), smoothing splines (e.g., Wahba [1990]), and on support vector machines (e.g., Schölkopf and Smola [2002]) for further practical information.

For the explicit construction of radial basis function approximants, we resume working in the setup established in Section 3.2.2, letting ϕ be a prescribed conditionally positive definite radial basis function with polynomial degree m, and considering some space $\mathcal{D} \subset \mathbb{R}^d$ which is supposed to be large enough to contain at least one $\mathcal{P}_m^d$-unisolvent subset.

5.2.2.1 Least-squares Approximation

A technique frequently employed to recover an unknown function f from a set of noisy function values $\hat{f}(x_1), \ldots, \hat{f}(x_n)$ through radial basis functions is a *least-squares approximation*, see, e.g., Buhmann [2003], Chapter 8, or Iske [2004], Section 3.10. To this effect, approximants s of the generic form (3.7)[29] are considered for a reduced numbered of pairwise distinct centres $\{\tilde{x}_j\}_{j=1}^{\tilde{n}} \subset \mathcal{D}$, $\tilde{n} + \widetilde{m} < n$, which usually coincide with some of the sample

[29] The use of polynomials is not explicitly required in the least-squares setup to ensure the existence of optimal approximants for all choices of ϕ, in contrast to interpolation. Yet, polynomials may still be added to the linear combination of radial basis functions to increase the flexibility of the resulting approximation.

points $x_1, \dots, x_n$, but may also be different. This form is then used to obtain an optimal approximant $s^* \in \tilde{\mathcal{A}}_\phi(\mathcal{D})$ by solving the linear least-squares problem

$$\min_{s \in \tilde{\mathcal{A}}_\phi(\mathcal{D})} \sum_{i=1}^{n} w_i \big(s(x_i) - \hat{f}(x_i)\big)^2, \tag{5.6}$$

where the linear function space $\tilde{\mathcal{A}}_\phi(\mathcal{D})$ is given, in line with definition (3.16), by

$$\tilde{\mathcal{A}}_\phi(\mathcal{D}) := \widetilde{\mathcal{F}}_\phi(\mathcal{D}) + \mathcal{P}_m^d,$$

with

$$\widetilde{\mathcal{F}}_\phi(\mathcal{D}) := \left\{ \sum_{i=1}^{\tilde{n}} \lambda_i \phi(\|\cdot - \tilde{x}_i\|_2) : \tilde{n} \in \mathbb{N}, \lambda \in \mathbb{R}^{\tilde{n}}, \{\tilde{x}_i\}_{i=1}^{\tilde{n}} \subset \mathcal{D}, \sum_{i=1}^{\tilde{n}} \lambda_i p(x_i) = 0, p \in \mathcal{P}_m^d \right\},$$

and where the positive weights $w_1, \dots, w_n$ are additionally included to take care of potential heteroscedasticity in the data.

Due to the side conditions in $\tilde{\mathcal{A}}_\phi(\mathcal{D})$, problem (5.6) actually constitutes a linear least-squares problem with equality constraints. These are given by $\widetilde{P}^\top \lambda = 0$ for the modified polynomial basis matrix $\widetilde{P} \in \mathbb{R}^{\tilde{n} \times \widetilde{m}}$, $\widetilde{P}_{ij} = p_j(\tilde{x}_i)$, $i = 1, \dots, \tilde{n}$, $j = 1, \dots, \widetilde{m}$. Hence, by substituting s with centres $\{\tilde{x}_i\}_{i=1}^{\tilde{n}}$ into problem (5.6), the latter can be rewritten in matrix form as

$$\begin{aligned} &\min_{(\lambda, c)^\top \in \mathbb{R}^{\tilde{n} + \widetilde{m}}} && \big\| W^{1/2} \big(\widetilde{\Phi} \lambda + P c - \hat{f}_X \big) \big\|_2^2 \\ &\text{s.t.} && \widetilde{P}^\top \lambda = 0, \end{aligned} \tag{5.7}$$

where $\widetilde{\Phi} \in \mathbb{R}^{n \times \tilde{n}}$ is the modified interpolation matrix with entries $\widetilde{\Phi}_{ij} = \phi(\|x_i - \tilde{x}_j\|_2)$, $P \in \mathbb{R}^{n \times \widetilde{m}}$ the usual polynomial basis matrix with $P_{ij} = p_j(x_i)$, $W := \operatorname{diag}(w_1, \dots, w_n) \in \mathbb{R}^{n \times n}$ the positive weight matrix, and $\hat{f}_X = (\hat{f}(x_1), \dots, \hat{f}(x_n))^\top \in \mathbb{R}^n$ the vector of observed function values. According to Björck [1996], Section 5.1, problem (5.7) admits a solution if $\operatorname{rank}(\widetilde{P}) = \widetilde{m}$, which is equivalent to requiring that the set of centres $\{\tilde{x}_1, \dots, \tilde{x}_{\tilde{n}}\}$ is $\mathcal{P}_m^d$-unisolvent, see Section 3.2.1. Moreover, given its existence, a solution to (5.7) is unique if and only if $\operatorname{rank}\big(\begin{smallmatrix} \widetilde{\Phi} & P \\ \widetilde{P}^\top & 0 \end{smallmatrix} \big) = \tilde{n} + \widetilde{m}$, which can be shown to hold in case the $\mathcal{P}_m^d$-unisolvent set of centres forms a subset of the sample points $\{x_1, \dots, x_n\}$, see, e.g., Iske [2004], Theorem 17. By a straightforward application of the Karush-Kuhn-Tucker (KKT) conditions for necessary first-order optimality (e.g., Nocedal and Wright [2006], Chapter 12), the optimal solution $(\lambda^*, c^*)^\top$ to (5.7) is then determined by the linear system

$$\begin{pmatrix} \widetilde{\Phi}^\top W \widetilde{\Phi} & \widetilde{\Phi}^\top W P & \widetilde{P} \\ P^\top W \widetilde{\Phi} & P^\top W P & 0 \\ \widetilde{P}^\top & 0 & 0 \end{pmatrix} \begin{pmatrix} \lambda^* \\ c^* \\ v^* \end{pmatrix} = \begin{pmatrix} \widetilde{\Phi}^\top W \hat{f}_X \\ P^\top W \hat{f}_X \\ 0 \end{pmatrix},$$

for optimal Lagrange multipliers $v^* \in \mathbb{R}^{\widetilde{m}}$.

A least-squares ansatz may notably reduce the complexity of constructing an approximant if $\tilde{n} \ll n$. Yet, its main drawback is the choice of a suitable set of centres which defines both the smoothness of an approximant and its closeness to the data. Intuitively, selecting a smaller number of centres tends to increase the smoothness of the approximant, which, however, is additionally affected by the locations of the centres in a rather indefinite way. This ambiguity, though, makes it difficult to incorporate the technique into an optimisation algorithm, as argued, for instance, by Žilinskas [2010]. By all means, strategies for choosing a set of centres are available, such as scattered data filtering techniques (e.g., Iske [2004], Section 4.4) or adaptive algorithms (e.g., Wendland [2005b], Section 15.6, and Fasshauer [2007], Chapter 21), which then might even be combined with available error bounds. But these strategies are generally used on a stand-alone basis as they lack theoretical foundation. In particular, if any of them is used to sequentially build response surfaces, it may be observed that removing any 'global' data point from a set of centres is likely to result in an immediate resampling in the very same region where essential information has been withdrawn. Eventually, Locatelli and Schoen [2013], Section 3.2, claim that a least-squares ansatz may be particularly useful in later stages of an algorithm, when a good global picture of the objective function is available and sample points accumulate exceedingly around some local minima.

Refinements of the least-squares ansatz may be found in relation with radial basis function approximation as well, such as *moving least-squares* (e.g., Levin [1998] and Wendland [2005b], Chapter 4). These variants, though, merely focus on improving the local approximation quality, rather than resolving the issue of how to select suitable sets of centres when used in a response surface method. Therefore, they also remain fairly unsuitable for our purpose.

In least-squares approximation, the smoothness of an approximant is predominantly determined by the number and locations of the centres, whose influence, though, is rather difficult to capture. As seen in Subsection 3.2.2 on the variational theory of radial basis functions, however, the smoothness of an approximant in the form (3.7) may as well be measured by its semi-norm on the function space $\mathcal{A}_\phi$. It thus seems meaningful to explicitly include this measure into the construction of a suitable approximant, which is done by the following two approaches.

5.2.2.2 Relaxed Interpolation

Given the interpolation problem (3.19), an obvious way to construct a radial basis function approximant under the assumption of error bounds is to relax the interpolation constraints and allow the approximant to deviate at the sampled points x_i from the observed function values $\hat{f}(x_i)$ by at most ϵ_i. Accordingly, requiring approximants to be as smooth as possible and letting the centres of their form (3.7) coincide again with all sample points, an optimal approximant $s^* \in \mathcal{A}_\phi(\mathcal{D})$ is found upon solving

$$\begin{aligned} \min_{s \in \mathcal{A}_\phi(\mathcal{D})} \quad & \|s\|_\phi^2 \\ \text{s.t.} \quad & \big|w_i\big(s(x_i) - \hat{f}(x_i)\big)\big| \le \epsilon_i, \quad i = 1, \dots, n, \end{aligned} \tag{5.8}$$

where the inequality constraints confine the maximal deviation from the noisy function values and the positive weights $w_1, \dots, w_n$ are taken into account for generality. Clearly, by imposing $\epsilon_i = 0$ on all $i = 1, \dots, n$, in this setup, the exact case of interpolation may be recovered.

By definition of the semi-norm (3.18) and the side conditions in $\mathcal{A}_\phi(\mathcal{D})$, problem (5.8) presents a quadratic programme with both equality and inequality constraints. Since ϕ is assumed to be conditionally positive definite, the problem is strictly convex such that it admits a unique solution for a $\mathcal{P}_m^d$-unisolvent set of points $\{x_1, \dots, x_n\}$. However, as a consequence of the involved inequality constraints, the optimal solution s^* to (5.8) can no longer be determined by solving a linear system of equations, as compared to an interpolation or a classical least-squares ansatz. Instead, standard solvers for convex quadratic programmes are required in order to solve the problem efficiently, such as active set methods or interior points methods, see, e.g., Nocedal and Wright [2006], Chapter 16, for further details. Additionally, a gradient projection method may be useful if the problem is reformulated in terms of its simpler bound-constrained dual, which in matrix notation assumes the form

$$\begin{aligned} \min_{(\upsilon,\tilde{\upsilon})^\top \in \mathbb{R}^{2n}} \quad & -\frac{1}{4}\big(\upsilon - \tilde{\upsilon}\big)^\top W \Phi W \big(\upsilon - \tilde{\upsilon}\big) - \upsilon^\top \big(\epsilon + W \hat{f}_X\big) - \tilde{\upsilon}^\top \big(\epsilon - W \hat{f}_X\big) \\ \text{s.t.} \quad & \upsilon, \tilde{\upsilon} \geq 0, \end{aligned}$$

for inequality constraint Lagrange multipliers $\upsilon, \tilde{\upsilon} \in \mathbb{R}^n$, corresponding equality constraint multipliers being zero on $\mathcal{A}_\phi(\mathcal{D})$, and where W and $\hat{f}_X$ are given as before, and $\epsilon = (\epsilon_1, \dots, \epsilon_n)^\top$ denotes the vector of available error bounds.

A distinctive feature of problem (5.8) follows by closer examination of the first-order necessary conditions, which may be simplified to

$$\lambda_i^* = -\frac{1}{2} w_i (\upsilon_i^* - \tilde{\upsilon}_i^*), \qquad i = 1, \dots, n, \tag{5.9}$$

along with

$$\upsilon_i^* \geq 0, \quad \Big(w_i\big(s^*(x_i) - \hat{f}(x_i)\big) - \epsilon_i\Big) \leq 0, \quad \upsilon_i^* \Big(w_i\big(s^*(x_i) - \hat{f}(x_i)\big) - \epsilon_i\Big) = 0, \tag{5.10}$$

and

$$\tilde{\upsilon}_i^* \geq 0, \quad \Big(w_i\big(\hat{f}(x_i) - s^*(x_i)\big) - \epsilon_i\Big) \leq 0, \quad \tilde{\upsilon}_i^* \Big(w_i\big(\hat{f}(x_i) - s^*(x_i)\big) - \epsilon_i\Big) = 0. \tag{5.11}$$

where $\upsilon^*, \tilde{\upsilon}^* \in \mathbb{R}^n$ denote the optimal Lagrange multipliers of the inequality constraints. The complementary conditions in (5.10) and (5.11) then imply that either the optimal approximant s^* interpolates $\hat{f}(x_i) \pm \epsilon_i / w_i$ at a sampled point x_i or $\upsilon_i^* = \tilde{\upsilon}_i^* = 0$, or possibly both. In case $\upsilon_i^* = \tilde{\upsilon}_i^* = 0$, however, equation (5.9) yields $\lambda_i^* = 0$, such that the i-th radial basis function is omitted in the construction of s^*. In particular, it is thus natural to expect that the set of active indices i with $|w_i(s(x_i) - \hat{f}(x_i))| = \epsilon_i$ is rather large if the approximating surface is fitted to a small or moderate data set. In fact, many function values are approximated by interpolating the endpoints of their (potentially scaled) error bounds, which may provide an unrealistic property for some applications. On the contrary, if the surface is constructed from a large number of data function values, then the set of active indices will rather be a small subset of the complete data, resulting in a reduction of the computational complexity, as pointed out by Schaback and Wendland [2006], Section 3.

Because of the above characteristic, the approach is most frequently known in the literature on machine learning, where the active indices correspond to the so-called 'support vectors' of the approximating surface, see, e.g., Schölkopf and Smola [2002]. Within the same context, other variations may also be encountered, above all Vapnik's ϵ-insensitive loss function, which allows to violate the 'hard' ϵ-constraints by absorbing the incurred losses linearly into the objective function. While this may overcome that approximants tend to interpolate the endpoints of the error bounds, it still has the drawback of posing an inequality constrained problem. Yet, if different loss functions are employed, then a solution by means of a linear system is possible, e.g., Wahba [1999]. The most popular choice in this regard is given by the next approach for radial basis functions.

5.2.2.3 Regularised Least-squares Approximation

The *regularised least-squares approximation*, as described, for instance, in Wendland and Rieger [2005] or Wendland [2005a], is another approach that explicitly incorporates the semi-norm into the construction. However, instead of imposing inequality constraints to regulate the discrepancy to noisy function values, the closeness to the values $\hat{f}(x_1), \ldots, \hat{f}(x_n)$ is assessed by the residual sum of squares $1/n \sum_{i=1}^{n} w_i(s(x_i) - \hat{f}(x_i))^2$. Consequently, an optimal approximant $s^* \in \mathcal{A}_\phi(\mathcal{D})$ is sought as the solution of

$$\min_{s \in \mathcal{A}_\phi(\mathcal{D})} \gamma \|s\|_\phi^2 + \frac{1}{n} \sum_{i=1}^{n} w_i \big(s(x_i) - \hat{f}(x_i)\big)^2, \tag{5.12}$$

where the additional parameter $\gamma > 0$ is introduced to control the trade-off between the smoothness of the approximant and its closeness to the noisy function values in form of the residual sum of squares. Specifically, for large γ, the smoothness of the approximant is emphasised, where in the limit, i.e. for $\gamma \to \infty$, a linear regression through the data is obtained, while for small γ the closeness to the data is enforced, yielding an interpolation of the values $\hat{f}(x_1), \ldots, \hat{f}(x_n)$ in case $\gamma = 0$.

Taking into account the constraints on the coefficients λ in the space $\mathcal{A}_\phi(\mathcal{D})$, problem (5.12) comprises an equality constrained quadratic programme. Thus, similar to pure interpolation and a plain least-squares approximation but unlike the previous relaxed approach, it can be reduced to solving a linear system of equations in order to find an optimal solution $s^* \in \mathcal{A}_\phi(\mathcal{D})$. Since this result will play an important role in the sequel of this chapter, we state it in form of a theorem and give the corresponding proof. In similar form, the result may be found in Wendland [2005a], Theorem 4, and the subsequent discussion therein.

Theorem 5.1. *Let ϕ be a conditionally positive definite radial basis function of order m, and assume that a $\mathcal{P}_m^d$-unisolvent set of points $\{x_1, \ldots, x_n\} \subset \mathcal{D}$ with corresponding function* *values $\hat{f}(x_1), \ldots, \hat{f}(x_n)$ is given. Then, for any $\gamma > 0$, the approximant $s^* \in \mathcal{A}_\phi(\mathcal{D})$ whose coefficients are determined by the linear system*

$$\begin{pmatrix} \Phi + n\gamma W^{-1} & P \\ P^\top & 0 \end{pmatrix} \begin{pmatrix} \lambda \\ c \end{pmatrix} = \begin{pmatrix} \hat{f}_X \\ 0 \end{pmatrix}, \tag{5.13}$$

where $W = \operatorname{diag}(w_1, \ldots, w_n)$ and $\hat{f}_X = (\hat{f}(x_1), \ldots, \hat{f}(x_n))^\top$, is the unique element of $\mathcal{A}_\phi(\mathcal{D})$ that solves the regularised least-squares approximation problem (5.12).

Proof. Let $\gamma > 0$ be fix, and observe that problem (5.12) can be rewritten as

$$\begin{aligned} \min_{(\lambda,c)^\top \in \mathbb{R}^{n+\tilde{m}}} \quad & n\gamma\lambda^\top \Phi \lambda + \left\| W^{1/2}(\Phi\lambda + Pc - \hat{f}_X) \right\|_2^2 \\ \text{s.t.} \quad & P^\top \lambda = 0. \end{aligned} \tag{5.14}$$

By the conditional positive definiteness of ϕ, problem (5.14) is strictly convex. Hence, a unique solution exists if the set of sample points $\{x_1, \ldots, x_n\}$ is $\mathcal{P}_m^d$-unisolvent, guaranteeing that the matrix $P^\top$ has full row rank. Applying the KKT conditions to (5.14) further provides the linear equations

$$(n\gamma\Phi + \Phi^\top W \Phi)\lambda + \Phi^\top W P c + P\upsilon = \Phi^\top W \hat{f}_X \tag{5.15}$$

$$P^\top W(\Phi\lambda + Pc) = P^\top W \hat{f}_X \tag{5.16}$$

$$P^\top \lambda = 0, \tag{5.17}$$

where $\upsilon \in \mathbb{R}^{\tilde{m}}$ denotes the Lagrange multiplier for the constraint $P^\top\lambda = 0$. Since ϕ is conditionally positive definite, the matrix Φ is invertible for any $\lambda \in \mathbb{R}^n \backslash \{0\}$ satisfying (5.17). Hence, multiplying equation (5.15) by $(\Phi W)^{-1}$ simplifies to

$$(\Phi + n\gamma W^{-1})\lambda + Pc + (\Phi^\top W)^{-1} P \upsilon = \hat{f}_X,$$

which, by substituting into (5.16), yields $P^\top \Phi^{-1} P \upsilon = 0$. However, since $\{x_1, \ldots, x_n\}$ is $\mathcal{P}_m^d$-unisolvent, the latter implies $\upsilon = 0$, such that we arrive at the stated linear system (5.13). □

Theorem 5.1 reveals that an optimal regularised least-squares approximant is computed by solving a modified linear system, where the modification consists in adding a scaled inverted weight matrix W^{-1} to the interpolation matrix Φ. In particular, this implies that errors in function values are taken into account by interpolating some perturbed noisy function values. The magnitude of the perturbation essentially depends on the regularisation parameter γ, whose determination thus plays a decisive role. Contrary to the case of a plain least-squares approximation, though, the parameter γ has a clear and intuitive interpretation, which facilitates its determination and allows for a better control over the involved inaccuracies. In approximation theory, most of the strategies applied to determine a suitable choice of γ are based on statistical criteria, such as cross-validation procedures (e.g., Wahba [1990], Chapter 4) or some discrepancy principle from regularisation theory (e.g., Hansen

[2010], Chapter 5). Their incorporation into response surface tools, however, may be subject to some modification as they casually produce unrealistic extreme values for small to moderate sample sizes, as remarked by Wahba [1990], Chapter 4. In any case, this is required when including assumption (5.5) on the pointwise error bounds, as nearly all of the strategies adjust γ by assessing the overall impact of the noise.

The regularised least-squares approach is sometimes further enhanced by reducing the centres of the radial basis functions to a strict subset of the sample points $\{x_1, \ldots, x_n\}$, e.g., **?**, Section 6. Since the smoothing effect of the parameter γ is different from the one implied by the number and locations of the centres, e.g., Girosi [1992], this approach may most suitably be incorporated in a response surface method towards the end of the optimisation, as already argued by Locatelli and Schoen [2013] in the least-squares ansatz. However, one should be aware that in this case three additional parameters are to be determined.

Summing up, the latter regularised least-squares approximation seems to provide the most suitable approach for integration in a response surface method. A plain least-squares approximation has the main drawback that an appropriate set of centres is difficult to determine for a sequence of response surface models along the iterations of a method. In turn, relaxing the interpolation to allow for pointwise deviations as in the second approach enables a convenient handling over the admissible errors in the first place. Yet, internal tests on Monte Carlo based applications have shown that the approximants commonly pass through the endpoints of the error bounds and thus constitute poor approximations to the true function. More importantly, the approach requires to solve computationally more involved inequality constrained optimisation problems instead of linear systems. It therefore lacks efficiency, especially having in mind that the problems need not only be solved for constructing approximants but also for determining new evaluation points in subsequent subproblems. Finally, using a regularised least-squares approach allows to control the balance between smoothness and closeness to the data by means of a regularisation parameter, which is easy to interpret and only requires minor adjustments for incorporation into a response surface method. It further retains the advantageous feature of having to solve 'only' a (modified) linear system in order to construct an approximant and to select a new evaluation point.

5.3 A Radial Basis Function Method for Noisy Objective Functions

In this section, we present a RBF method for minimising a nonconvex objective function, where function evaluations are expensive and subject to noise. The level of noise is assumed to be controlled by available error bounds satisfying condition (5.5). To account for the presence of noise, we extend Gutmann's well-established RBF method for deterministic functions by using a regularised least-squares criterion for constructing suitable radial basis function approximants. Uniqueness of the approximants can therefore be guaranteed under the same weak conditions as in the case of interpolation, while the additional regularisation parameter is adjusted such that the approximants satisfy the assumed error bounds. Moreover, by employing the regularised least-squares criterion in similar fashion as the semi-norm within Jones's general response surface technique, see Jones [1996], we are able to define new evaluation points by means of a target value, cf. also Žilinskas [2010]. On the analogy of Gutmann's method, the corresponding subproblem for finding new points can then be reformulated, which facilitates to show convergence of the method for surface spline type radial basis functions under particular (simplified) assumptions on the level of noise and the choice of target values. Eventually, by applying the method to a number of modified test problems and to the calibration of the Hull-White interest rate model under the VSAA strategy, it can be shown that it yields promising results compared to its deterministic analogue.

In order to introduce the RBF method for noisy objective functions, we follow Banholzer et al. [2017b] and proceed as in the earlier presentation of the deterministic counterpart in that we first give a general description of the method. In Subsection 5.3.2, we then establish the convergence of the method and state several convergence results. In Subsection 5.3.3, we elaborate on practical issues concerning the implementation of the method. Finally, Subsection 5.3.4 demonstrates the applicability of the method to solve relevant test problems and to calibrate the Hull-White model in the presence of noise.

5.3.1 Description of Method

To minimise a general noisy objective function $\hat{f} : \mathcal{X} \to \mathbb{R}$ on a compact set $\mathcal{X}$, let ϕ be a chosen conditionally positive definite radial basis function of order m and $\mathcal{P}_m^d$ be the space of polynomials of degree at most $m-1$ with basis $\{p_j\}_{j=1}^{\widetilde{m}}$. Further, assume that points $x_1, \dots, x_n$ have been sampled in earlier iterations of the method, forming a $\mathcal{P}_m^d$-unisolvent set, along with their noisy function values $\hat{f}(x_1), \dots, \hat{f}(x_n)$. Eventually, assume that error bounds $\epsilon_i > 0$ are available for the observations $\hat{f}(x_i)$ according to condition (5.5), and that there are some positive weights $w_1, \dots, w_n$ for taking care of potential heteroscedasticity in the data. A general iteration of the method for noisy objective function values can then be described as follows.

5.3.1.1 Construction of Response Surface

Given the $\mathcal{P}_m^d$-unisolvency of $\{x_1, \dots, x_n\}$, corresponding function values $\hat{f}(x_1), \dots, \hat{f}(x_n)$ and weights $w_1, \dots, w_n$, Theorem 5.1 states that for any positive regularisation parameter there exists a unique approximant from $\mathcal{A}_\phi(\mathcal{X})$ minimising the regularised least-squares criterion in problem (5.12). Let this approximant be denoted by

$$s_n^{\gamma_n}(x) = \sum_{i=1}^{n} \lambda_i \phi(\|x - x_i\|_2) + p(x), \quad x \in \mathbb{R}^d, \tag{5.18}$$

where we use the superscript 'γ_n' to emphasise the dependence on the regularisation parameter, and thus to distinguish it from its interpolating counterpart.

Theorem 5.1 thus provides a general construction of the regularised least-squares approximant (5.18) for any $\gamma_n > 0$. To adequately specify the latter in the current iteration of the method, we make use of the fact that we are predominantly interested in finding a rather smooth approximant that deviates at most by the error bounds from the noisy function values to recover the underlying function f, see assumption (5.5). Accordingly, we first observe that the smoothness of the approximant $s_n^{\gamma_n}$ can alternatively be characterised in terms of the parameter γ_n, which seems intuitively clear from the formulation of problem (5.12). More formally, it can be justified by the following proposition, for which we define by $\mathcal{R}(P)$ the range of the polynomial basis matrix $P \in \mathbb{R}^{n\times\widetilde{m}}$ and recall that $\hat{f}_X = (\hat{f}(x_1), \ldots, \hat{f}(x_n))^\top$.

Proposition 5.2. *Let ϕ be a conditionally positive definite radial basis function of order m, and let $\{x_1, \ldots, x_n\} \subset \mathcal{X}$ be a $\mathcal{P}_m^d$-unisolvent set with corresponding function values $\hat{f}(x_1), \ldots, \hat{f}(x_n)$. For $\gamma > 0$, let $s_n^\gamma \in \mathcal{A}_\phi(\mathcal{X})$ denote the unique solution to the regularised least-squares problem* (5.12)*. Then, it holds:*

(a) s_n^γ depends continuously on γ.

(b) For any $\hat{f}_X \in \mathbb{R}^n$, $\|s_n^\gamma\|_\phi$ is monotonically decreasing and $1/n \sum_{i=1}^n w_i(s_n^\gamma(x_i) - \hat{f}(x_i))^2$ is monotonically increasing in γ. If $\hat{f}_X \notin \mathcal{R}(P)$, then the functions are strictly monotonically decreasing and increasing in γ, respectively.

Proof. Since the associated matrix on the left-hand side of the linear system (5.13) is nonsingular and depends continuously on γ, so does its inverse, which trivially establishes (a).

To show (b), let $0 < \gamma_1 < \gamma_2$ be fix. By the optimality of the corresponding minimisers $s_n^{\gamma_1}, s_n^{\gamma_2} \in \mathcal{A}_\phi(\mathcal{X})$, we then have

$$\gamma_1\|s_n^{\gamma_1}\|_\phi^2 + \frac{1}{n}\sum_{i=1}^n w_i\big(s_n^{\gamma_1}(x_i) - \hat{f}(x_i)\big)^2 \le \gamma_1\|s_n^{\gamma_2}\|_\phi^2 + \frac{1}{n}\sum_{i=1}^n w_i\big(s_n^{\gamma_2}(x_i) - \hat{f}(x_i)\big)^2, \tag{5.19}$$

and

$$\gamma_2\|s_n^{\gamma_2}\|_\phi^2 + \frac{1}{n}\sum_{i=1}^n w_i\big(s_n^{\gamma_2}(x_i) - \hat{f}(x_i)\big)^2 \le \gamma_2\|s_n^{\gamma_1}\|_\phi^2 + \frac{1}{n}\sum_{i=1}^n w_i\big(s_n^{\gamma_1}(x_i) - \hat{f}(x_i)\big)^2.$$

Adding both inequalities and cancelling equal terms yields

$$\gamma_1\|s_n^{\gamma_1}\|_\phi^2 + \gamma_2\|s_n^{\gamma_2}\|_\phi^2 \le \gamma_1\|s_n^{\gamma_2}\|_\phi^2 + \gamma_2\|s_n^{\gamma_1}\|_\phi^2,$$

which provides

$$(\gamma_2 - \gamma_1)\|s_n^{\gamma_2}\|_\phi^2 \le (\gamma_2 - \gamma_1)\|s_n^{\gamma_1}\|_\phi^2.$$

Hence, $\|s_n^\gamma\|_\phi^2$ is monotonically decreasing in γ, and by inequality (5.19) it further follows that

$$0 \le \gamma_1\Big(\|s_n^{\gamma_1}\|_\phi^2 - \|s_n^{\gamma_2}\|_\phi^2\Big) \le \frac{1}{n}\sum_{i=1}^n w_i\big(s_n^{\gamma_2}(x_i) - \hat{f}(x_i)\big)^2 - \frac{1}{n}\sum_{i=1}^n w_i\big(s_n^{\gamma_1}(x_i) - \hat{f}(x_i)\big)^2, \tag{5.20}$$

which thus shows that $1/n \sum_{i=1}^{n} w_i(s_n^\gamma(x_i) - \hat{f}(x_i))^2$ is monotonically increasing in γ.

To establish the strict monotonicity of both functions in case $\hat{f}_X \notin \mathcal{R}(P)$, we further show that both minimisers $s_n^{\gamma_1}$ and $s_n^{\gamma_2}$ cannot be identical for $0 < \gamma_1 < \gamma_2$, which thus implies that $\|s_n^{\gamma_1}\|_\phi \neq \|s_n^{\gamma_2}\|_\phi$. To this end, assume $s_n^{\gamma_1} \equiv s_n^{\gamma_2}$ and observe that the linear system (5.13) provides

$$s_n^{\gamma_1}(x_i) - \hat{f}(x_i) = -n\gamma_1 w_i^{-1}\lambda_i^{\gamma_1} \qquad \text{and} \qquad s_n^{\gamma_1}(x_i) - \hat{f}(x_i) = -n\gamma_2 w_i^{-1}\lambda_i^{\gamma_1}, \tag{5.21}$$

$i = 1, \ldots, n$, where $\lambda_i^{\gamma_1}$ denotes the i-th coefficient of $s_n^{\gamma_1}$. The latter in turn yields

$$n(\gamma_2 - \gamma_1)w_i^{-1}\lambda_i^{\gamma_1} = 0,$$

and therefore $\lambda_i^{\gamma_1} = 0$ for $i = 1 \ldots, n$. This, however, implies that $s_n^{\gamma_1} \in \mathcal{P}_m^d$, such that the function values $\hat{f}(x_1), \ldots, \hat{f}(x_n)$ in (5.21) are interpolated by a polynomial from the linear space $\mathcal{P}_m^d$, which contradicts the assumption $\hat{f}_X \notin \mathcal{R}(P)$. Hence, $\|s_n^\gamma\|_\phi$ is strictly monotonically decreasing in γ, from which the strict monotonicity of $1/n \sum_{i=1}^{n} w_i(s_n^\gamma(x_i) - \hat{f}(x_i))^2$ follows by (5.20). □

Accordingly, Proposition 5.2 allows to control the smoothness of the approximant $s_n^{\gamma_n}$ by the parameter γ_n, whose relation can be identified even uniquely under the weak assumption that $\hat{f}_X \notin \mathcal{R}(P)$. Hence, finding the smoothest approximant $s_n^{\gamma_n}$ such that it deviates at the considered points x_i from the noisy function values $\hat{f}(x_i)$ by at most ϵ_i can be stated as the *auxiliary problem*

$$\begin{aligned} &\max_{\gamma > 0} \quad \gamma \\ &\text{s.t.} \quad \left|s_n^\gamma(x_i) - \hat{f}(x_i)\right| \leq \epsilon_i, \quad i = 1, \ldots, n. \end{aligned} \tag{5.22}$$

Problem (5.22) consists of a linear objective function in one dimension which is subject to n nonlinear inequality constraints. Since $s_n^{\gamma_n}$ converges to the interpolant of $\hat{f}(x_1), \ldots, \hat{f}(x_n)$ for $\gamma_n \to 0$, as can be read off from the regularised system (5.13), a feasible solution to problem (5.22) exists. However, unlike the sum-of-squares function, the individual constraints are non-monotonic in γ_n, and each evaluation of the constraints requires to solve the linear system (5.13). This renders the problem difficult to solve and unnecessarily time-consuming if a solution is sought that is as exact as possible. Nonetheless, as γ_n is readjusted in each iteration upon the addition of a new point, searching for an approximate solution is sufficient. Latter can be obtained by an efficient backtracking strategy, which starts with a large enough γ_n and successively decreases this value by a suitable discretisation, until all constraints are met for the first time. Ideally, this algorithm then terminates with an approximant as smooth as possible and satisfying the admissible error bounds. For a more detailed description of the algorithm, see the practical Section 5.3.3.

5.3.1.2 Determination of Next Evaluation Point

To determine the next point of evaluation x_{n+1}, we continue in a way similar to Jones's technique and assume that a target value f_n^* with the usual functionality has been chosen.

Latter value is supposed to be completely noise-free as it will be set in relation to the surface $s_n^{\gamma_n}$, meant to capture the underlying deterministic function f. Keeping $\gamma_n > 0$ fixed at the value determined in problem (5.22), we then let x_{n+1} be the point $y \in \mathcal{X} \backslash \{x_1, \dots, x_n\}$ such that the augmented approximant $s_y^{\gamma_n} \in \mathcal{A}_\phi(\mathcal{X})$ minimises the regularised least-squares criterion to previous sample points and interpolates f_n^* at the new y. In formal terms, we thus require that $s_y^{\gamma_n}$ solves

$$\begin{aligned} \min_{s \in \mathcal{A}_\phi(\mathcal{X})} \quad & \gamma_n \|s\|_\phi^2 + \frac{1}{n} \sum_{i=1}^{n} w_i \big(s(x_i) - \hat{f}(x_i)\big)^2 \\ \text{s.t.} \quad & s(y) = f_n^*, \end{aligned} \tag{5.23}$$

which is a strictly convex optimisation problem on $\mathcal{A}_\phi(\mathcal{X})$ and thus admits a unique solution, cf. Theorem 5.1.

To simplify problem (5.23) in terms of the sought new point $y \in \mathcal{X} \backslash \{x_1, \dots, x_n\}$, we first rewrite the augmented approximant $s_y^{\gamma_n}$ according to

$$s_y^{\gamma_n}(x) = s_n^{\gamma_n}(x) + \left[f_n^* - s_n^{\gamma_n}(y)\right] l_n^{\gamma_n}(y, x), \quad x \in \mathbb{R}^d, \tag{5.24}$$

where $l_n^{\gamma_n}(y, \cdot\,) \in \mathcal{A}_\phi(\mathcal{X})$ is the radial basis function approximant that solves the constrained regularised least-squares problem

$$\begin{aligned} \min_{l(y, \cdot\,) \in \mathcal{A}_\phi(\mathcal{X})} \quad & \gamma_n \|l(y, \cdot\,)\|_\phi^2 + \frac{1}{n} \sum_{i=1}^{n} w_i \big(l(y, x_i)\big)^2 \\ \text{s.t.} \quad & l(y, y) = 1. \end{aligned} \tag{5.25}$$

Representation (5.24) is valid since for any $y \in \mathcal{X} \backslash \{x_1, \dots, x_n\}$ both $s_n^{\gamma_n}$ and $l_n^{\gamma_n}(y, \cdot\,)$ are uniquely defined as solutions to the problems (5.12) and (5.25), respectively. Hence, the right-hand side of (5.24) is a unique well-defined element of $\mathcal{A}_\phi(\mathcal{X})$, and also satisfies the interpolation constraint in problem (5.23). Moreover, similar to Theorem 5.1, it can be shown that the approximating function $l_n^{\gamma_n}(y, \cdot\,)$ has the form

$$l_n^{\gamma_n}(y, x) = \sum_{i=1}^{n} \alpha_i(y) \phi(\|x - x_i\|_2) + \beta(y) \phi(\|x - y\|_2) + \sum_{j=1}^{\widetilde{m}} b_j(y) p_j(x), \quad x \in \mathbb{R}^d,$$

where the coefficients $\alpha(y) = (\alpha_1(y), \dots, \alpha_n(y))^\top \in \mathbb{R}^n$, $\beta(y) \in \mathbb{R}$ and $b(y) = (b_1(y), \dots, b_{\widetilde{m}}(y))^\top \in \mathbb{R}^{\widetilde{m}}$ are defined by the linear system

$$\begin{pmatrix} \Phi + n\gamma_n W^{-1} & m_n(y) & P \\ m_n(y)^\top & \phi(0) & \pi(y)^\top \\ P^\top & \pi(y) & 0 \end{pmatrix} \begin{pmatrix} \alpha(y) \\ \beta(y) \\ b(y) \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \tag{5.26}$$

for the familiar matrices $\Phi \in \mathbb{R}^{n \times n}$, $P \in \mathbb{R}^{n \times \widetilde{m}}$ and $W \in \mathbb{R}^{n \times n}$, and the corresponding vectors $m_n(y) = (\phi(\|x_1 - y\|_2), \dots, \phi(\|x_n - y\|_2))^\top \in \mathbb{R}^n$ and $\pi(y) = (p_1(y), \dots, p_{\widetilde{m}}(y))^\top \in \mathbb{R}^{\widetilde{m}}$.

By inserting representation (5.24) into the objective function of (5.23), the latter can then be reformulated as

$$\begin{aligned}
&\gamma_n \|s_y^{\gamma_n}\|_\phi^2 + \frac{1}{n}\sum_{i=1}^n w_i \big(s_y^{\gamma_n}(x_i) - \hat{f}(x_i)\big)^2 \\
&\quad = \gamma_n \Big(\|s_n^{\gamma_n}\|_\phi^2 + 2\big[f_n^* - s_n^{\gamma_n}(y)\big]\langle s_n^{\gamma_n}, l_n^{\gamma_n}(y,\,\cdot\,)\rangle_\phi + \big[f_n^* - s_n^{\gamma_n}(y)\big]^2 \|l_n^{\gamma_n}(y,\,\cdot\,)\|_\phi^2 \Big) \\
&\qquad + \frac{1}{n}\sum_{i=1}^n w_i \Big(\big(s_n^{\gamma_n}(x_i) - \hat{f}(x_i)\big)^2 + \big[f_n^* - s_n^{\gamma_n}(y)\big]^2 \big(l_n^{\gamma_n}(y, x_i)\big)^2 \\
&\qquad + 2\big(s_n^{\gamma_n}(x_i) - \hat{f}(x_i)\big)\big[f_n^* - s_n^{\gamma_n}(y)\big] l_n^{\gamma_n}(y, x_i) \Big) \\
&\quad = \gamma_n \|s_n^{\gamma_n}\|_\phi^2 + \frac{1}{n}\sum_{i=1}^n w_i \big(s_n^{\gamma_n}(x_i) - \hat{f}(x_i)\big)^2 \\
&\qquad + \big[f_n^* - s_n^{\gamma_n}(y)\big]^2 \Big(\gamma_n \| \, l_n^{\gamma_n}(y,\,\cdot\,)\|_\phi^2 + \frac{1}{n}\sum_{i=1}^n w_i \big(l_n^{\gamma_n}(y, x_i)\big)^2 \Big),
\end{aligned} \tag{5.27}$$

where the last equation holds by definition of the semi-inner product (3.17) and the relation $s_n^{\gamma_n}(x_i) - \hat{f}(x_i) = -n\gamma_n w_i^{-1}\lambda_i$, $i = 1, \dots, n$, due to (5.13), which both together yield

$$\begin{aligned}
\langle s_n^{\gamma_n}, l_n^{\gamma_n}(y,\,\cdot\,)\rangle_\phi &= \sum_{i=1}^n \lambda_i l_n^{\gamma_n}(y, x_i) \\
&= -\frac{1}{n\gamma_n}\sum_{i=1}^n w_i \big(s_n^{\gamma_n}(x_i) - \hat{f}(x_i)\big) l_n^{\gamma_n}(y, x_i).
\end{aligned}$$

Now, the first two terms on the right-hand side of equation (5.27) are independent of y and correspond to the objective function for constructing the approximant $s_n^{\gamma_n}$, cf. problem (5.12). It thus suffices to consider the last term in (5.27) to find the new point y. However, by the semi-inner product (3.17) and the linear system (5.26), implying $l_n^{\gamma_n}(y, x_i) = -n\gamma_n w_i^{-1}\alpha_i(y)$ for $i = 1, \dots, n$, it holds

$$\begin{aligned}
&\|l_n^{\gamma_n}(y,\,\cdot\,)\|_\phi^2 + \frac{1}{n\gamma_n}\sum_{i=1}^n w_i \big(l_n^{\gamma_n}(y, x_i)\big)^2 \\
&\quad = \sum_{i=1}^n \alpha_i(y) l_n^{\gamma_n}(y, x_i) + \beta(y) l_n^{\gamma_n}(y, y) + \frac{1}{n\gamma_n}\sum_{i=1}^n w_i \big(l_n^{\gamma_n}(y, x_i)\big)^2 \\
&\quad = \beta(y).
\end{aligned}$$

Therefore, we can conclude that solving the required problem (5.23) is equivalent to minimising the nonnegative utility function

$$g_n^{\gamma_n}(y) := \mu_n^{\gamma_n}(y)\big[f_n^* - s_n^{\gamma_n}(y)\big]^2, \qquad y \in \mathcal{X}\backslash\{x_1, \dots, x_n\}, \tag{5.28}$$

where the function $\mu_n^{\gamma_n} : \mathcal{X}\backslash\{x_1, \dots, x_n\} \to \mathbb{R}$ is defined for $\gamma_n > 0$ by

$$\mu_n^{\gamma_n}(y) := \|l_n^{\gamma_n}(y,\,\cdot\,)\|_\phi^2 + \frac{1}{n\gamma_n}\sum_{i=1}^n w_i \big(l_n^{\gamma_n}(y, x_i)\big)^2 = \beta(y). \tag{5.29}$$

Note the resemblance of the functions $g_n^{\gamma_n}$ and $\mu_n^{\gamma_n}$ to their deterministic counterparts (3.28) and (3.29), respectively. In particular, since $l_n^{\gamma_n}(y,\cdot)$ is well-defined for $y \in \mathcal{X}\backslash\{x_1,\dots,x_n\}$, so are both functions $g_n^{\gamma_n}$ and $\mu_n^{\gamma_n}$.

Remark 5.3. *By the same argument as given in Remark 3.12, definition* (5.29) *implies that the function $\mu_n^{\gamma_n}$ is positive on $\mathcal{X}\backslash\{x_1,\dots,x_n\}$. Thus, assuming the existence of a point $y_0 \in \mathcal{X}\backslash\{x_1,\dots,x_n\}$ with $\mu_n^{\gamma_n}(y_0)=0$, definition* (5.29) *and the $\mathcal{P}_m^d$-unisolvency of $\{x_1,\dots,x_n\}$ yield $l_n^{\gamma_n}(y_0,y_0)\equiv 0$, in contradiction to the interpolation constraint $l_n^{\gamma_n}(y_0,y_0)=1$.*

Moreover, using (5.29) *and Cramer's rule to solve the linear system* (5.26)*, we have*

$$\mu_n^{\gamma_n}(y) = \frac{\det A_n^{\gamma_n}}{\det A_n^{\gamma_n}(y)}, \qquad y \in \mathcal{X}\backslash\{x_1,\dots,x_n\},$$

where $A_n^{\gamma_n}$ and $A_n^{\gamma_n}(y)$ denote the nonsingular matrices on the left-hand sides of the linear systems (5.13) *and* (5.26)*, respectively, that is*

$$A_n^{\gamma_n} := \begin{pmatrix} \Phi + n\gamma_n W^{-1} & P \\ P^\top & 0 \end{pmatrix} \quad \text{and} \quad A_n^{\gamma_n}(y) := \begin{pmatrix} \Phi + n\gamma_n W^{-1} & m_n(y) & P \\ m_n(y)^\top & \phi(0) & \pi(y)^\top \\ P^\top & \pi(y) & 0 \end{pmatrix}. \tag{5.30}$$

However, since $\det A_n^{\gamma_n}$ is a nonzero constant and $\lim_{y\to x_i}\det A_n^{\gamma_n}(y) \neq 0$ for any $i \in \{1,\dots,n\}$, it holds

$$\lim_{y\to x_i} \mu_n^{\gamma_n}(y) < \infty, \qquad i = 1,\dots,n,$$

as opposed to the divergence of $\mu_n(y)$ as y tends to any x_i, cf. property (3.31)*. Hence, even though $\mu_n^{\gamma_n}$ is not defined at the sample points $x_1,\dots,x_n$, it can be continuously extended at these points by the positive and finite values*

$$\mu_n^{\gamma_n}(x_i) = \frac{\det A_n^{\gamma_n}}{\det A_n^{\gamma_n}(x_i)}, \qquad i = 1,\dots,n, \tag{5.31}$$

due to the continuity of the determinant.

5.3.1.3 Choice of Target Value

Since the target value f_n^* has the same functionality as in the case of an interpolation, see Section 3.3.1.3, its choice similarly determines the location of the next evaluation point x_{n+1}, minimising $g_n^{\gamma_n}$ on $\mathcal{X}\backslash\{x_1,\dots,x_n\}$ for fixed $\gamma_n > 0$. Accordingly, if $f_n^* \in [\min_{y\in\mathcal{X}} s_n^{\gamma_n}(y), \max_{y\in\mathcal{X}} s_n^{\gamma_n}(y)]$, then any point $y \neq \{x_1,\dots,x_n\}$ with $s_n^{\gamma_n}(y) = f_n^*$ also yields $g_n^{\gamma_n}(y)=0$, such that the existence of a global minimiser of $g_n^{\gamma_n}$ on $\mathcal{X}\backslash\{x_1,\dots,x_n\}$ is not necessarily guaranteed. Hence, to ensure an existing x_{n+1} on $\mathcal{X}\backslash\{x_1,\dots,x_n\}$, we must at least require that

$$f_n^* \in \big[-\infty, \min_{y\in\mathcal{X}} s_n^{\gamma_n}(y)\big], \tag{5.32}$$

where the choice $f_n^* = \min_{y\in\mathcal{X}} s_n^{\gamma_n}(y)$ is only admissible if none of the x_i is a global minimiser of $s_n^{\gamma_n}$, i.e. if $f_n^* < s_n^{\gamma_n}(x_i)$, $i=1,\dots,n$. Unfortunately, however, it remains unclear at this point whether a choice $f_n^* < \min_{y\in\mathcal{X}} s_n^{\gamma_n}(y)$ is also sufficient to guarantee that a global minimiser of $g_n^{\gamma_n}$ on $\mathcal{X}\backslash\{x_1,\ldots,x_n\}$ exists, i.e. that a global minimiser of $g_n^{\gamma_n}$ over $\mathcal{X}$ does not coincide with any of the sample points $x_1,\ldots,x_n$, as in the case of interpolation, or whether a further condition needs to be imposed. The main issue here is due to the fact that $g_n^{\gamma_n}$ is continuously extendable at the points x_i, $i=1,\ldots,n$, by a finite value, cf. equation (5.31), which then implies that $g_n^{\gamma_n}(y)$ does not tend to infinity anymore as y approaches any x_i.

In any case, though, it is to be noted that for an admissible choice $f_n^* < \min_{y\in\mathcal{X}} s_n^{\gamma_n}(y)$, Remark 3.13 also allows to draw the same conclusions for $g_n^{\gamma_n}$ as for g_n, i.e. that for $f_n^* = -\infty$, the minimisation of $g_n^{\gamma_n}$ on $\mathcal{X}\backslash\{x_1,\ldots,x_n\}$ reduces to minimising $\mu_n^{\gamma_n}$ on the same feasible set. Again, however, it is unclear whether a global minimiser of $\mu_n^{\gamma_n}$ over $\mathcal{X}$ is distinct from $x_1,\ldots,x_n$.

As we have been unable to resolve above issue, yet our practical experience indicates that a global minimiser of $g_n^{\gamma_n}$ over $\mathcal{X}$ hardly coincides with any $x_1,\ldots,x_n$, we will implicitly assume throughout the remainder of this chapter that an admissible choice of f_n^* according to (5.32) leads to a well-defined x_{n+1} away from any x_i, $i=1,\ldots,n$, as in the exact setup.

5.3.1.4 Algorithm

All in all, the RBF method for minimising a noisy objective function $\hat{f}:\mathcal{X}\to\mathbb{R}$ on a compact set $\mathcal{X}$ can be formulated by the following consolidated algorithm, provided that positive error bounds ϵ_i satisfying condition (5.5) and positive weights w_i are available along with the i-th evaluation of $\hat{f}$.

Algorithm 5.4. *(RBF Method for Noisy Objective Functions).*

0. ***Initial step:***

- *Choose a conditionally positive definite radial basis function ϕ of order m.*
- *Generate a $\mathcal{P}_m^d$-unisolvent set of points $\{x_1,\ldots,x_{n_0}\}\subset\mathcal{X}$.*
- *Evaluate $\hat{f}$ at the points $x_1,\ldots,x_{n_0}$, and set $n=n_0$.*

1. ***Iteration step:***

while $n \le n^{\max}$ ***do***

- *Construct the approximant $s_n^{\gamma_n}\in\mathcal{A}_\phi(\mathcal{X})$ solving*

$$\min_{s\in\mathcal{A}_\phi(\mathcal{X})} \gamma_n\|s\|_\phi^2 + \frac{1}{n}\sum_{i=1}^n w_i\big(s(x_i)-\hat{f}(x_i)\big)^2,$$

where γ_n is determined according to

$$\begin{aligned} &\max_{\gamma>0} \quad \gamma \\ &s.t. \quad \big|s_n^\gamma(x_i)-\hat{f}(x_i)\big| \le \epsilon_i, \quad i=1,\ldots,n. \end{aligned}$$

```
        - Choose an admissible target value $f_n^* \in \big[-\infty, \min_{y\in\mathcal{X}} s_n^{\gamma_n}(y)\big]$.
        - Determine $x_{n+1}$, which is the value of $y$ solving
              $\min_{y\in\mathcal{X}\setminus\{x_1,\dots,x_n\}} \mu_n^{\gamma_n}(y)\big[f_n^* - s_n^{\gamma_n}(y)\big]^2.$
        - Evaluate $\hat{f}$ at $x_{n+1}$, and set $n = n+1$.
    end while
```

5.3.2 Convergence of Method

As Gutmann's RBF method, the RBF method for noisy objective functions is a purely deterministic sequential sampling algorithm. For a given set of function values, the construction of an approximant and the subsequent selection of a new evaluation point is carried out independent of any source of randomness. This means that starting off with the same initialisation and using the same scheme for evaluating $\hat{f}$ (i.e., by fixing the seed if $\hat{f}$ arises from a stochastic simulation), the method will always generate the same sequence of sample points. To show convergence of the method to the global minimum of any general continuous function, our main task is thus to establish the density of the sequence of generated iterates $\{x_n\}$ in $\mathcal{X}$. Based on Törn and Žilinskas [1989], Theorem 1.3, and in line with Theorem 3.16 for the original RBF method, we may therefore state the following theorem for the convergence of Algorithm 5.4.

Theorem 5.5. *Algorithm 5.4 converges for every continuous function f if and only if it generates a sequence of points $\{x_n\}$ that is dense in $\mathcal{X}$.*

Algorithm 5.4 and thus Theorem 5.5 have been stated in terms of a general noisy objective function $\hat{f}$. If we consider the special case of the VSAA approach where $\hat{f}$ is formed by a sequence of objective functions $\{\hat{f}_{N_k}\}$ that arises from a Monte Carlo simulation, Theorem 5.5 has to be reformulated to account for the underlying probabilistic setup. Specifically, for a particular realisation $\omega_\text{V} \in \Omega_\text{V}$ on the probability space $(\Omega_\text{V}, \mathcal{F}_\text{V}, \mathbb{Q}_\text{V})$, each surface $s_n^{\gamma_n}$ is then built by use of the function values $\hat{f}_{N_k}(x_k) = \hat{f}_{N_k}(x_k(\omega_\text{V}), \omega_\text{V})$, $k = 1, \dots, n$, where the corresponding uniformly error bounds[30] ϵ_{N_k} may be adopted under the assumptions of Proposition 2.34 in form of the right-hand side of inequality (2.38). In particular, since each point $x_n(\omega_\text{V})$ selected by Algorithm 5.4 depends on the sample path ω_V, so does the sequence of generated iterates $\{x_n(\omega_\text{V})\}$. Consequently, Theorem 5.5 must read as follows, and requires all assertions that are made in the following for a general noisy function $\hat{f}$ (e.g., the assertion that $n\gamma_n \to 0$ as $n \to \infty$) to be reformulated as '$\mathbb{Q}_\text{V}$-almost surely'.

Theorem 5.6. *Algorithm 5.4 converges $\mathbb{Q}_\text{V}$-almost surely for every continuous function f if and only if it generates a sequence of points $\{x_n\}$ that is dense in $\mathcal{X}$, $\mathbb{Q}_\text{V}$-almost surely.*

[30] Note that the bounds only apply from a finite number $k^*(\omega_\text{V}) \in \mathbb{N}$ on. This, however, does not affect the main convergence analysis.

In order to show the convergence of the method for a noisy function $\hat{f}$ in the sequel, we first address the role of the error bounds to this effect. Since this, however, is still an unresolved issue, we then present the main convergence results followed by a proof of convergence under rather general assumptions.

5.3.2.1 Assumption on Error Bounds

Given the functionality of the noisy RBF method, one approach to establish the required density of the iterates is to resort to the available convergence results of Gutmann's deterministic method and show that these pertain if the respective exact function values $f(x_i)$ are replaced by the noisy observations $\hat{f}(x_i)$ for $i = 1, 2, \ldots$. An indispensable assumption is thus that the involved level of noise decreases to zero over the course of the optimisation procedure, i.e. that $\epsilon_n \to 0$, as $n \to \infty$. Unfortunately, however, as opposed to our intuition, we have not been able to conclude from such an assumption via the auxiliary problem (5.22) that the sequence $\{n\gamma_n\}$ also goes to zero for $n \to \infty$. As one may already conjecture from the construction of regularised least-squares approximants through (5.13), this will be required to adopt Gutmann's proof of convergence for noisy function values. Nevertheless, under the simplified assumption that the bounds were adjusted in each iteration n according to $\epsilon_i^{(n)}$, $i = 1, \ldots, n$, and we required $\max_{1 \le i \le n} \epsilon_i^{(n)} \to 0$ as $n \to \infty$, it can be shown that the sequence $\{n\gamma_n\}$ converges to zero for $n \to \infty$, as shown in the following theorem.

Theorem 5.7. *Let ϕ be a conditionally positive definite radial basis function spline of order m, and let $\{x_1, \ldots, x_n\} \subset \mathcal{X}$ be a $\mathcal{P}_m^d$-unisolvent set with corresponding function values $\hat{f}(x_1), \ldots, \hat{f}(x_n)$ such that $\hat{f}_X = (\hat{f}(x_1), \ldots, \hat{f}(x_n))^\top \in \mathcal{R}(P)$ for the related polynomial basis matrix P. Let $s_n^{\gamma_n} \in \mathcal{A}_\phi(\mathcal{X})$ denote the unique optimal solution of the regularised least-squares problem* (5.12)*, where the regularisation parameter $\gamma_n > 0$ solves*

$$\begin{aligned} &\max_{\gamma > 0} \quad \gamma \\ &\text{s.t.} \quad \left| s_n^\gamma(x_i) - \hat{f}(x_i) \right| \le \epsilon_i^{(n)}, \quad i = 1, \ldots, n, \end{aligned}$$

for some positive error bounds $\epsilon_i^{(n)}$. Further, assume that $\max_{1 \le i \le n} \epsilon_i^{(n)} \to 0$ as $n \to \infty$. Then, the sequence $\{n\gamma_n\}$ converges to zero as $n \to \infty$.

Proof. For fixed $n \in \mathbb{N}$, note that problem (5.12) with $\gamma > 0$ constitutes a scalarisation of the bi-objective optimisation problem

$$\min_{s \in \mathcal{A}_\phi(\mathcal{X})} \Big(\|s\|_\phi^2, \sum_{i=1}^n w_i \big(s(x_i) - \hat{f}(x_i)\big)^2 \Big), \tag{5.33}$$

with corresponding weight vector $(\gamma, 1/n)^\top$. Consequently, since $\gamma > 0$, any optimal solution s^γ of the scalarised problem (5.12) is Pareto optimal for problem (5.33), and since $\mathcal{A}_\phi(\mathcal{X})$ is a convex set and both objective functions in (5.33) are convex in s^γ, there is some positive weight vector for any Pareto optimal point of (5.33) such that it is an optimal solution of the scalarisation (5.12), see, e.g., Ehrgott [2005], Theorem 4.1.

Since for any $\gamma_n > 0$, the optimal solution $s_n^{\gamma_n}$ of (5.12) is unique, the mapping

$$G_n : \gamma_n \mapsto \Big(\|s_n^{\gamma_n}\|_\phi^2 , \sum_{i=1}^n w_i \big(s_n^{\gamma_n}(x_i) - \hat{f}(x_i) \big)^2 \Big),$$

assigning γ_n to each point of the set of Pareto optimal solutions, is well-defined. Due to Proposition 5.2, the functions $\|s_n^{\gamma_n}\|_\phi^2$ and $\sum_{i=1}^n w_i(s_n^{\gamma_n}(x_i) - \hat{f}(x_i))^2$ are also continuous and strictly monotonic in γ_n, such that each function value of G_n is attained uniquely by moving along γ_n. In particular, by increasing γ_n, we strictly decrease $\|s_n^{\gamma_n}\|_\phi^2$ and strictly increase $\sum_{i=1}^n w_i(s_n^{\gamma_n}(x_i) - \hat{f}(x_i))^2$, and vice versa, where the extreme points of G_n are given for $\gamma_n \to 0$ by the interpolant to $\hat{f}(x_1), \dots, \hat{f}(x_n)$, and for $\gamma_n \to \infty$ by the Pareto optimal solution $s_n^{\gamma_n}$ with $\|s_n^{\gamma_n}\|_\phi^2 = 0$.

The graphs of G_n can further be identified with a function $\varphi : (0, \infty) \to \mathbb{R}_{\geq 0}$, where for a given $\xi_n = \|s_n^{\gamma_n}\|_\phi^2$ with unique $\gamma_n = \gamma_n(\xi_n)$, we set $\varphi(\xi_n) = \sum_{i=1}^n w_i(s_n^{\gamma_n}(x_i) - \hat{f}(x_i))^2$. Similarly, for given γ_n, we let $\xi_n(\gamma_n)$ be the unique ξ_n with $\gamma_n = \gamma_n(\xi_n(\gamma_n))$. By definition, both functions $\varphi(\xi_n)$ and $\xi_n(\gamma_n)$ are thus continuous, strictly convex and monotonically decreasing.

Now, for each $\epsilon^{(n)}$, there exists a unique point on the image of φ with $\varphi(\xi_n) = \zeta_n$, say. Since, by assumption, $\max_{1 \leq i \leq n} \epsilon_i^{(n)} \to 0$ as $n \to \infty$, it follows $\zeta_n \to 0$, and consequently $\gamma_n \to 0$ as $n \to \infty$. In turn, this yields that $\xi_n \to \lim_{n \to \infty} \|s_n\|_\phi^2$ for $n \to \infty$. Moreover, a subderivative of φ at ξ_n is given by $-n\gamma_n$. Due to the uniqueness of the weight vector $(\gamma_n, 1/n)$, the subderivative is even unique such that φ is continuously differentiable with $\varphi'(\xi_n) = -n\gamma_n$. However, by the continuity of φ', it also holds $\varphi'(\lim_{n \to \infty} \|s_n\|_\phi^2) = 0$, which thus implies that $\{n\gamma_n\}$ must converge to zero as $n \to \infty$. □

5.3.2.2 Convergence Results

Besides the assumption that $n\gamma_n \to 0$ as $n \to \infty$, we further require the target values f_n^* to be set sufficiently low compared to the approximating surfaces $s_n^{\gamma_n}$ in order to achieve convergence of the RBF method for noisy objective functions, cf. condition (3.33) for exact function values. Due the presence of noise, however, the critical thresholds for f_n^* need to be adjusted marginally to eventually guarantee convergence of the method in a similar fashion as Gutmann. To this end, we let, for infinitely many $n \in \mathbb{N}$, the target values f_n^* satisfy

$$f_n^* < \min_{y \in \mathcal{X}} \Big[s_n^{\gamma_n}(y) - \tau \|s_n^{\gamma_n}\|_\infty \big[\Delta_n(y) + \widetilde{w}_n^{-1/2}(y) \big]^{\rho/2} \Big], \tag{5.34}$$

where, as in the noise-free counterpart, $\tau > 0$ and $\rho \geq 0$ are constants with $\rho < 1$, for $\phi(r) = r$, and $\rho < 2$, otherwise, and Δ_n denotes the minimum distance function (3.34). For given $y \in \mathcal{X}$, the function $\widetilde{w}_n(y)$ gives the weight w_i of the sample point x_i that is closest to y, i.e. for $i(y) = \operatorname{argmin}_{1 \leq i \leq n} \|y - x_i\|_2$, we have

$$\widetilde{w}_n(y) := w_{i(y)}, \tag{5.35}$$

with the convention that the largest $i(y)$ is selected among the minimising indices if argmin is not unique. Gutmann's main convergence result stating that the generated sequence is dense in $\mathcal{X}$, cf. Theorem 3.17, can then be formulated in the noisy setup for spline type radial basis functions as follows.

Theorem 5.8. *Let ϕ be a conditionally positive definite surface spline of order m from Table 3.1, and let $\{x_n\}$ be the sequence of iterates generated by Algorithm 5.4. Further, let $s_n^{\gamma_n}$ with $\gamma_n > 0$ be the optimal regularised least-squares approximant from $\mathcal{A}_\phi(\mathcal{X})$ to the data $(x_i, \hat{f}(x_i))$, $i = 1, \dots, n$, with corresponding weights w_i bounded away from zero. Assume that, for infinitely many $n \in \mathbb{N}$, the choice of f_n^* satisfies* (5.34)*, where τ, Δ_n, ρ and $\widetilde{w}_n$ are given as above, and that $n\gamma_n \to 0$ as $n \to \infty$. Then, the sequence $\{x_n\}$ is dense in $\mathcal{X}$.*

In light of Corollary 3.18 for Gutmann's RBF method, we can conclude the following particular convergence result from Theorems 5.5 and 5.8, due to the finiteness of the right-hand side in assumption (5.34) for any $n \in \mathbb{N}$.

Corollary 5.9. *Let ϕ and m be as in Theorem 5.8. Further, let f be continuous, and assume that, for infinitely many $n \in \mathbb{N}$, it holds $f_n^* = -\infty$, and that $n\gamma_n \to 0$ as $n \to \infty$. Then, Algorithm 5.4 converges.*

To derive a further convergence result that applies to noisy functions in the native space, i.e. to functions $\hat{f}$ for which the noise behaves sufficiently well such that $\hat{f} \in \mathcal{N}_\phi(\mathcal{X})$, we first show that for sufficiently large n the maximum norm of the approximating surface can be bounded uniformly, cf. Lemma 3.19 for the equivalent case of interpolation.

Lemma 5.10. *Let $\{x_n\}$ be a sequence in $\mathcal{X}$ with pairwise different points such that $\{x_1, \dots, x_{n_0}\}$ is $\mathcal{P}_m^d$-unisolvent. For $n \geq n_0$, let $s_n^{\gamma_n}$ with $\gamma_n > 0$ denote the optimal regularised least-squares approximant to $\hat{f}$ at $x_1, \dots, x_n$, where the respective weights $w_1, \dots, w_n$ are bounded away from zero. Further, let $\hat{f} \in \mathcal{N}_\phi(\mathcal{X})$, and assume that $n\gamma_n \leq n_0\gamma_{n_0}$ for sufficiently large n. Then, for n large enough, $\|s_n^{\gamma_n}\|_\infty$ is bounded above by a number that depends only on $x_1, \dots, x_{n_0}$, on γ_{n_0} and on $\hat{f}$.*

Proof. Fix $n \in \mathbb{N}$, and let y be any point in $\mathcal{X} \backslash \{x_1, \dots, x_n\}$. For $\gamma_n > 0$, let $\tilde{s}_n^{\gamma_n}$ be the optimal regularised least-squares approximant from $\mathcal{A}_\phi(\mathcal{X})$ to $(x_i, \hat{f}(x_i))$, $i = 1, \dots, n$, and $(y, \hat{f}(y))$, which solves

$$\min_{s \in \mathcal{A}_\phi(\mathcal{X})} \gamma_n \|s\|_\phi^2 + \frac{1}{n} \sum_{i=1}^{n} w_i \big(s(x_i) - \hat{f}(x_i)\big)^2 + w_y \big(s(y) - \hat{f}(y)\big)^2, \tag{5.36}$$

for weights w_i and w_y, respectively, bounded away from zero. By analogy with Subsection 5.3.1, the approximant can thus be rewritten as

$$\tilde{s}_n^{\gamma_n}(x) = s_n^{\gamma_n}(x) + \big[\hat{f}(y) - s_n^{\gamma_n}(y)\big] l_n^{\gamma_n}(y, x), \qquad x \in \mathbb{R}^d, \tag{5.37}$$

where $l_n^{\gamma_n}(y,\cdot)$ is the optimal regularised least-squares approximant to $(x_i,0)$ and $(y,1)$, with respective weights w_i and w_y. By substituting (5.37) into (5.36) and simplifying the resulting expression similar to Subsection 5.3.1, it then follows that

$$\begin{aligned}&\gamma_n\|\tilde{s}_n^{\gamma_n}\|_\phi^2+\frac{1}{n}\sum_{i=1}^n w_i\big(\tilde{s}_n^{\gamma_n}(x_i)-\hat{f}(x_i)\big)^2+w_y\big(\tilde{s}_n^{\gamma_n}(y)-\hat{f}(y)\big)^2\\&\qquad=\gamma_n\|s_n^{\gamma_n}\|_\phi^2+\frac{1}{n}\sum_{i=1}^n w_i\big(s_n^{\gamma_n}(x_i)-\hat{f}(x_i)\big)^2+\big[\hat{f}(y)-s_n^{\gamma_n}(y)\big]^2\gamma_n\mu_n^{\gamma_n}(y),\end{aligned}\tag{5.38}$$

where the positive function $\mu_n^{\gamma_n}$ of the approximant $l_n^{\gamma_n}(y,\cdot)$ is given by

$$\mu_n^{\gamma_n}(y)=\|l_n^{\gamma_n}(y,\cdot)\|_\phi^2+\frac{1}{n\gamma_n}\sum_{i=1}^n w_i\big(l_n^{\gamma_n}(y,x_i)\big)^2+\frac{1}{\gamma_n}w_y\big(l_n^{\gamma_n}(y,y)-1\big)^2.$$

Equality (5.38) thus yields

$$\big[\hat{f}(y)-s_n^{\gamma_n}(y)\big]^2\le\frac{\gamma_n\|\tilde{s}_n^{\gamma_n}\|_\phi^2+\frac{1}{n}\sum_{i=1}^n w_i\big(\tilde{s}_n^{\gamma_n}(x_i)-\hat{f}(x_i)\big)^2+w_y\big(\tilde{s}_n^{\gamma_n}(y)-\hat{f}(y)\big)^2}{\gamma_n\mu_n^{\gamma_n}(y)}.\tag{5.39}$$

The right-hand side of inequality (5.39) can further be bounded. On the one hand, the optimality of the approximant $\tilde{s}_n^{\gamma_n}$ provides

$$\begin{aligned}&\gamma_n\|\tilde{s}_n^{\gamma_n}\|_\phi^2+\frac{1}{n}\sum_{i=1}^n w_i\big(\tilde{s}_n^{\gamma_n}(x_i)-\hat{f}(x_i)\big)^2+w_y\big(\tilde{s}_n^{\gamma_n}(y)-\hat{f}(y)\big)^2\\&\qquad\le\gamma_n\|\tilde{s}_n\|_\phi^2+\frac{1}{n}\sum_{i=1}^n w_i\big(\tilde{s}_n(x_i)-\hat{f}(x_i)\big)^2+w_y\big(\tilde{s}_n(y)-\hat{f}(y)\big)^2\\&\qquad\le\gamma_n\big\|\hat{f}\big\|_{\mathcal{N}_\phi}^2,\end{aligned}\tag{5.40}$$

where $\tilde{s}_n$ is the optimal interpolant to the data $(x_i,\hat{f}(x_i))$, $i=1,\dots,n$, and $(y,\hat{f}(y))$, whose semi-norm is bounded by $\|\hat{f}\|_{\mathcal{N}_\phi}$ as $\hat{f}\in\mathcal{N}_\phi(\mathcal{X})$, see Definition 3.10. On the other hand, we have for sufficiently large $n\ge n_0$ with $n\gamma_n\le n_0\gamma_0$ that

$$\begin{aligned}\mu_n^{\gamma_n}(y)&\ge\|l_n^{\gamma_n}(y,\cdot)\|_\phi^2+\frac{1}{n_0\gamma_{n_0}}\sum_{i=1}^{n_0} w_i\big(l_n^{\gamma_n}(y,x_i)\big)^2+\frac{1}{\gamma_{n_0}}w_y\big(l_n^{\gamma_n}(y,y)-1\big)^2\\&\ge\|l_{n_0}^{\gamma_{n_0}}(y,\cdot)\|_\phi^2+\frac{1}{n_0\gamma_{n_0}}\sum_{i=1}^{n_0} w_i\big(l_{n_0}^{\gamma_{n_0}}(y,x_i)\big)^2+\frac{1}{\gamma_{n_0}}w_y\big(l_{n_0}^{\gamma_{n_0}}(y,y)-1\big)^2\\&=\mu_{n_0}^{\gamma_{n_0}}(y),\end{aligned}$$

where $l_{n_0}^{\gamma_{n_0}}(y,\cdot)$ is the optimal approximant to $(x_1,0),\dots,(x_{n_0},0)$ and $(y,1)$, with respective weights $w_1,\dots,w_{n_0}$ and w_y, and regularisation parameter $\gamma_{n_0}>0$. By Cramer's rule, the positive function $\mu_{n_0}^{\gamma_{n_0}}$ can then be computed as $\mu_{n_0}^{\gamma_{n_0}}(y)=\det A_{n_0}^{\gamma_{n_0}}/\det A_{n_0}^{\gamma_{n_0}}(y)$, where the nonsingular matrices $A_{n_0}^{\gamma_{n_0}}$ and $A_{n_0}^{\gamma_{n_0}}(y)$ are given in (5.30) for $n = n_0$, respectively, and the $(n_0 + 1)$-th diagonal entry of the latter matrix becomes $\phi(0) + \gamma_{n_0} w_y^{-1}$. Now, $\det A_{n_0}^{\gamma_{n_0}}$ is a nonzero constant and $\det A_{n_0}^{\gamma_{n_0}}(y)$ is bounded on $\mathcal{X}$, as a continuous function. It thus follows that $\mu_{n_0}^{\gamma_{n_0}}(y)$ is bounded away from zero. Hence, there exists a constant $\alpha_2 > 0$, depending on $x_1, \ldots, x_{n_0}$ and on γ_{n_0}, such that

$$\mu_n^{\gamma_n}(y) \geq \alpha_2, \qquad \forall\, y \in \mathcal{X} \backslash \{x_1, \ldots, x_n\},\ n \geq n_0. \tag{5.41}$$

Consequently, by (5.40) and (5.41), we get that inequality (5.39) reduces to

$$\big|\hat{f}(y) - s_n^{\gamma_n}(y)\big| \leq \frac{1}{\sqrt{\alpha_2}} \big\|\hat{f}\big\|_{\mathcal{N}_\phi}, \qquad y \in \mathcal{X} \backslash \{x_1, \ldots, x_n\},$$

which, as $\hat{f}$ is bounded on $\mathcal{X}$, results in

$$\big|s_n^{\gamma_n}(y)\big| \leq \frac{1}{\sqrt{\alpha_2}} \big\|\hat{f}\big\|_{\mathcal{N}_\phi} + \big\|\hat{f}\big\|_\infty, \qquad y \in \mathcal{X} \backslash \{x_1, \ldots, x_n\}.$$

If $y \in \{x_1, \ldots, x_n\}$, then the optimality of the approximant $s_n^{\gamma_n}$ implies

$$\frac{1}{n} w_y \big(s_n^{\gamma_n}(y) - \hat{f}(y)\big)^2 \leq \gamma_n \|s_n^{\gamma_n}\|_\phi^2 + \frac{1}{n} \sum_{i=1}^n w_i \big(s_n^{\gamma_n}(x_i) - \hat{f}(x_i)\big)^2 \leq \gamma_n \|s_n\|_\phi^2,$$

where s_n is the interpolant to the data $(x_i, \hat{f}(x_i))$, $i = 1, \ldots, n$. Hence, for $\hat{f} \in \mathcal{N}_\phi(\mathcal{X})$, it follows

$$\big|s_n^{\gamma_n}(y) - \hat{f}(y)\big| \leq \sqrt{n \gamma_n w_y^{-1}}\, \big\|\hat{f}\big\|_{\mathcal{N}_\phi},$$

which, for sufficiently large n, yields

$$\big|s_n^{\gamma_n}(y)\big| \leq \frac{1}{\sqrt{\alpha_3}} \big\|\hat{f}\big\|_{\mathcal{N}_\phi} + \big\|\hat{f}\big\|_\infty,$$

where $\alpha_3 > 0$ is a constant that depends on n_0 and on γ_{n_0}, as the weights are bounded away from zero. □

By combining Lemma 5.10 with Theorems 5.5 and 5.8, the following convergence result for noisy objective functions in the native space of surface spline type radial basis functions can be established. Note that we directly assume $\hat{f} \in \mathcal{N}_\phi(\mathcal{X})$ since there is no known criterion for a noisy function to be in the associated native space, in contrast to the deterministic case, cf. Theorem 3.11 and Corollary 3.20.

Corollary 5.11. *Let ϕ and m be as in Theorem 5.8. Further, let f be continuous, and let $\hat{f} \in \mathcal{N}_\phi(\mathcal{X})$. Assume that, for infinitely many $n \in \mathbb{N}$, it holds*

$$f_n^* < \min_{y \in \mathcal{X}} \Big[s_n^{\gamma_n}(y) - \tau \big[\Delta_n(y) + \widetilde{w}_n^{-1/2}(y)\big]^{\rho/2}\Big],$$

where τ, Δ_n, ρ and $\widetilde{w}_n$ are given as above, and that $n\gamma_n \to 0$ as $n \to \infty$. Then, Algorithm 5.4 converges.

5.3.2.3 Proof of Convergence

To prove the main theorem establishing the density of the generated sequence, we require some lemmas on the behaviour of the functions $\mu_n^{\gamma_n}$, $n \in \mathbb{N}$. They essentially generalise Lemmas 3.21 - 3.23, as used by Gutmann to show convergence of the original RBF method, to account for the presence of noise. Correspondingly, the first two lemmas are concerned with the limit of the sequence $\{\mu_n^{\gamma_n}(x_n)\}$.

Lemma 5.12. *Let ϕ be a conditionally positive definite radial basis function of order m from Table 3.1, and let $\{z_1, \dots, z_k\}$ be a $\mathcal{P}_m^d$-unisolvent set in a compact set $\mathcal{X} \subset \mathbb{R}^d$. Let $\{x_n\}$ and $\{y_n\}$ be convergent sequences in $\mathcal{X}$ that have the same limit $x^* \notin \{z_1, \dots, z_k\}$ and satisfy $x_n \neq y_n$, $n \in \mathbb{N}$. Further, let $\tilde{l}_n^{\gamma_n}(x_n, \cdot)$ with $\gamma_n > 0$ be the optimal regularised least-squares approximant to the data $(z_1, 0), \dots, (z_k, 0), (y_n, 0)$ and subject to $\tilde{l}_n^{\gamma_n}(x_n, x_n) = 1$, where the corresponding weights $w_1, \dots, w_k, w_n$ are bounded away from zero. If $n\gamma_n \to 0$ as $n \to \infty$, then*

$$\lim_{n\to\infty} \left[\|y_n - x_n\|_2 + w_n^{-1/2}\right]^\rho \tilde{\mu}_n^{\gamma_n}(x_n) = \infty, \tag{5.42}$$

where $\tilde{\mu}_n^{\gamma_n}$ is the function defined by (5.29) for the approximant $\tilde{l}_n^{\gamma_n}(x_n, \cdot)$, and where $0 \leq \rho < 1$, for $\phi(r) = r$, and $0 \leq \rho < 2$, otherwise.

Proof. For $\gamma_n > 0$, consider the optimal approximant $\tilde{l}_n^{\gamma_n}(x_n, \cdot)$ to $(z_1, 0), \dots, (z_k, 0), (y_n, 0)$, with corresponding weights $w_1, \dots, w_k, w_n$, and interpolating $(x_n, 1)$. For sufficiently large n, neither x_n nor y_n is in the set $\{z_1, \dots, z_k\}$, so that Cramer's rule may be applied to compute the function $\tilde{\mu}_n^{\gamma_n}$ associated to $\tilde{l}_n^{\gamma_n}(x_n, \cdot)$ by

$$\tilde{\mu}_n^{\gamma_n}(x_n) = \frac{\det A_n^{\gamma_n}}{\det A_n^{\gamma_n}(x_n)},$$

where the nonsingular matrices $A_n^{\gamma_n}$ and $A_n^{\gamma_n}(x_n)$ are of the form (5.30) for the points $z_1, \dots, z_k, y_k$ and $z_1, \dots, z_k, y_k, x_n$, respectively. In particular, the latter matrix is thus written as

$$A_n^{\gamma_n}(x_n) = \begin{pmatrix} \Phi + n\gamma_n W^{-1} & m_k(y_n) & m_k(x_n) & P \\ m_k(y_n)^\top & \phi(0) + n\gamma_n w_n^{-1} & \phi(\|y_n - x_n\|_2) & \pi(y_n)^\top \\ m_k(x_n)^\top & \phi(\|y_n - x_n\|_2) & \phi(0) & \pi(x_n)^\top \\ P^\top & \pi(y_n) & \pi(x_n) & 0 \end{pmatrix},$$

where $\Phi \in \mathbb{R}^{k\times k}$ and $P \in \mathbb{R}^{k\times \widetilde{m}}$ correspond to the interpolation and polynomial basis matrix of $\{z_1, \dots, z_k\}$, respectively, $W = \operatorname{diag}(w_1, \dots, w_k)$, and $m_k(y) = (\phi(\|z_1 - y\|_2), \dots, \phi(\|z_k - y\|_2))^\top$ and $\pi(y) = (p_1(y), \dots, p_{\widetilde{m}}(y))^\top$ for any $y \in \mathcal{X}$.

By the continuity of the determinant and the assumption $n\gamma_n \to 0$ as $n \to \infty$ with weights bounded away from zero, it follows that $\lim_{n\to\infty} \det A_n^{\gamma_n} = \det A^* \neq 0$, where A^* denotes the nonsingular interpolation matrix given in form of the left-hand side of (3.11) for the points $z_1, \dots, z_k, x^*$. In order to show assertion (5.42), it therefore remains to consider expression

$$\left[\|y_n - x_n\|_2 + w_n^{-1/2}\right]^{-\rho} \det A_n^{\gamma_n}(x_n), \tag{5.43}$$

for which we show in the following that it converges to zero as $n \to \infty$. First note that the $(k+1)$-th and $(k+2)$-th rows of the matrix $A_n^{\gamma_n}(x_n)$, given by

$$\begin{array}{llll} \big(m_k(y_n)^\top & \phi(0) + n\gamma_n w_n^{-1} & \phi(\|y_n - x_n\|_2) & \pi(y_n)^\top\big), \quad \text{and} \\ \big(m_k(x_n)^\top & \phi(\|y_n - x_n\|_2) & \phi(0) & \pi(x_n)^\top\big), \end{array}$$

have the same limit for $n \to \infty$, as the weights are bounded away from zero and $n\gamma_n \to 0$ for $n \to \infty$. Consequently, $\det A_n^{\gamma_n}(x_n) \to 0$ as $n \to \infty$, and hence, for $\rho = 0$, assertion (5.42) follows immediately.

For $\rho > 0$, note that the determinant of $A_n^{\gamma_n}(x_n)$ does not change if the $(k+1)$-th row of the matrix $A_n^{\gamma_n}(x_n)$ is replaced by the difference between the $(k+1)$-th and the $(k+2)$-th row, and, subsequently, the $(k+1)$-th column is replaced by the difference between the $(k+1)$-th and the $(k+2)$-th column. Therefore, $\det A_n^{\gamma_n}(x_n)$ can equally be computed as

$$\begin{vmatrix} \Phi + n\gamma_n W^{-1} & m_k(y_n) - m_k(x_n) & m_k(x_n) & P \\ m_k(y_n)^\top - m_k(x_n)^\top & 2[\phi(0) - \phi(\|y_n - x_n\|_2)] + n\gamma_n w_n^{-1} & \phi(\|y_n - x_n\|_2) - \phi(0) & \pi(y_n)^\top - \pi(x_n)^\top \\ m_k(x_n)^\top & \phi(\|y_n - x_n\|_2) - \phi(0) & \phi(0) & \pi(x_n)^\top \\ P^\top & \pi(y_n) - \pi(x_n) & \pi(x_n) & 0 \end{vmatrix}.$$

To deduce the convergence of expression (5.43) to zero, we then divide the $(k+1)$-th row and the $(k+1)$-th column of the latter determinant by $[\|y_n - x_n\|_2 + w_n^{-1/2}]^{\rho/2}$, and make the following remarks on the newly formed $(k+1)$-th column.

For all choices of ϕ, the functions $\phi(\|z_i - \cdot\|_2)$, $i = 1, \ldots, k$, are Lipschitz continuous on $\mathcal{X}$. This implies for $\rho < 2$ that

$$\lim_{n\to\infty} \frac{\phi(\|z_i - y_n\|_2) - \phi(\|z_i - x_n\|_2)}{\big[\|y_n - x_n\|_2 + w_n^{-1/2}\big]^{\rho/2}} = 0, \qquad i = 1, \ldots, k,$$

such that $m_k(y_n) - m_k(x_n) \to 0$ as $n \to \infty$. Similarly, for the same choice of ρ, the Lipschitz continuity of the polynomials yields

$$\lim_{n\to\infty} \frac{p_j(y_n) - p_j(x_n)}{\big[\|y_n - x_n\|_2 + w_n^{-1/2}\big]^{\rho/2}} = 0, \qquad j = 1, \ldots, \widetilde{m},$$

resulting in $\pi(y_n) - \pi(x_n) \to 0$ as $n \to \infty$. Further, we have

$$\lim_{n\to\infty} \frac{\phi(\|y_n - x_n\|_2) - \phi(0)}{\big[\|y_n - x_n\|_2 + w_n^{-1/2}\big]^{\rho}} = 0,$$

for $\rho < \nu$ in the case of surface splines and for $\rho < 2$ in the other cases. This follows directly in the case of surface splines, due to their form, and by the second order Taylor expansion in the other cases, as $\phi'(0) = 0$ and $\phi''(r)$ is bounded for small r.

Eventually, by assuming that $n\gamma_n \to 0$ as $n \to \infty$ and since w_n is bounded away from zero, we observe for $\rho < 2$,

$$\lim_{n\to\infty} \frac{n\gamma_n w_n^{-1}}{\big[\|y_n - x_n\|_2 + w_n^{-1/2}\big]^{\rho}} = 0.$$

Altogether, we therefore have that expression (5.43) converges for the given choices of ρ to zero as $n \to \infty$, proving that assertion (5.42) also holds in case $\rho > 0$. $\square$

Lemma 5.13. *Let ϕ and m be chosen as in Lemma 5.12, where ρ takes a value as indicated. Let $\{x_n\}$ be a sequence in $\mathcal{X}$ with pairwise different points such that $\{x_1, \dots, x_{n_0}\}$ is $\mathcal{P}_m^d$-unisolvent. For any $y \in \mathcal{X} \backslash \{x_1, \dots, x_n\}$, let $l_n^{\gamma_n}(y, \cdot)$ with $\gamma_n > 0$ be the optimal regularised least-squares approximant to the data $(x_1, 0), \dots, (x_n, 0)$ and subject to $l_n^{\gamma_n}(y, y) = 1$, where the corresponding weights $w_1, \dots, w_n$ are bounded away from zero. If $n\gamma_n \to 0$ as $n \to \infty$, then for every convergent subsequence $\{x_{n_k}\}_{k \in \mathbb{N}}$ of $\{x_n\}$ it holds*

$$\lim_{k \to \infty} \left[\Delta_{n_k - 1}(x_{n_k}) + \widetilde{w}_{n_k-1}^{-1/2}(x_{n_k})\right]^{\rho} \mu_{n_k-1}^{\gamma_{n_k-1}}(x_{n_k}) = \infty,$$

where Δ_{n_k-1}, $\mu_{n_k-1}^{\gamma_{n_k-1}}$ and $\widetilde{w}_{n_k-1}$ are the functions given by (3.34), (5.29) *and* (5.35), *respectively, for $n = n_k - 1$.*

Proof. For $n \geq 2$, let $i(x_n) = \operatorname{argmin}_{1 \leq i \leq n-1} \|x_n - x_i\|_2$, where we choose the largest $i(x_n)$ among the minimising indices if argmin is not unique, and let the sequence $\{y_n\}_{n \in \mathbb{N}}$ be defined as

$$y_n := \begin{cases} x_2, & n = 1, \\ x_{i(x_n)}, & n \geq 2. \end{cases}$$

Further, let $\{x_{n_k}\}$ be any subsequence of $\{x_n\}$ that converges to a point $x^* \in \mathcal{X}$. The choice of $\{y_n\}$ and convergence thus yield $\lim_{k\to\infty} \|x_{n_k} - y_{n_k}\|_2 = 0$. Also note that there always exists a $\mathcal{P}_m^d$-unisolvent set $\{\bar{x}_1, \dots, \bar{x}_l\}$, $l \in \mathbb{N}$, in the sequence $\{x_n\}$ that does not contain the limit point x^*. If $x^* = x_i$ for some $i \in \{1, \dots, n_0\}$, then we can pick x_{n_i} in a neighbourhood of x^* such that the initial set $\{x_1, \dots, x_{i-1}, x_{n_i}, x_{i+1}, \dots, x_{n_0}\}$ is $\mathcal{P}_m^d$-unisolvent.

For sufficiently large $k \in \mathbb{N}$ such that $y_{n_k} \notin \{\bar{x}_1, \dots, \bar{x}_l\}$ and for any $y \in \mathcal{X} \backslash \{x_1, \dots, x_{n_k-1}\}$, let $\bar{l}_k^{\gamma_k}(y, \cdot)$ with $\gamma_k > 0$ be the optimal regularised least-squares approximant to the data $(\bar{x}_1, 0), \dots, (\bar{x}_l, 0)$, $(y_{n_k}, 0)$, with corresponding weights $\bar{w}_1, \dots, \bar{w}_l$, w_{n_k} bounded away from zero, and subject to $\bar{l}_k^{\gamma_k}(y, y) = 1$. Likewise, let $l_{n_k-1}^{\gamma_{n_k-1}}(y, \cdot)$ with $\gamma_{n_k-1} > 0$ be the optimal regularised least-squares approximant to $(x_1, 0), \dots, (x_{n_k-1}, 0)$, with corresponding weights $w_1, \dots, w_{n_k-1}$ bounded away from zero, and subject to $l_{n_k-1}^{\gamma_{n_k-1}}(y, y) = 1$. Observe that, for k large enough, $l_{n_k-1}^{\gamma_{n_k-1}}(y, \cdot)$ approximates $(\bar{x}_i, 0)$, $i = 1, \dots, l$, and $(y_{n_k}, 0)$, along with their given weights and subject to the same interpolation condition. Hence, for sufficiently large k, the functions $\bar{\mu}_k^{\gamma_k}$ and $\mu_{n_k-1}^{\gamma_{n_k-1}}$ associated to $\bar{l}_k^{\gamma_k}(y, \cdot)$ and $l_{n_k-1}^{\gamma_{n_k-1}}(y, \cdot)$ via (5.29), respectively, and the optimality of $\bar{l}_k^{\gamma_k}(y, \cdot)$ imply

$$\begin{aligned} \bar{\mu}_k^{\gamma_k}(y) &= \left\|\bar{l}_k^{\gamma_k}(y, \cdot)\right\|_\phi^2 + \frac{1}{(l+1)\gamma_k} \left[\sum_{i=1}^{l} \bar{w}_i \big(\bar{l}_k^{\gamma_k}(y, \bar{x}_i)\big)^2 + w_{n_k} \big(\bar{l}_k^{\gamma_k}(y, y_{n_k})\big)^2\right] \\ &\leq \left\|l_{n_k-1}^{\gamma_{n_k-1}}(y, \cdot)\right\|_\phi^2 + \frac{1}{(l+1)\gamma_k} \sum_{i=1}^{n_k-1} w_i \big(l_{n_k-1}^{\gamma_{n_k-1}}(y, x_i)\big)^2 \\ &\leq \mu_{n_k-1}^{\gamma_{n_k-1}}(y), \end{aligned} \tag{5.44}$$

where the last inequality follows from the assumption that $n\gamma_n \to 0$ as $n \to \infty$.

Eventually, by definition of the sequence $\{y_n\}$ and applying Lemma 5.12 with the set of points $\{z_1, \dots, z_k\}$ being $\{\bar{x}_1, \dots, \bar{x}_l\}$, the weights $w_1 \dots, w_k$ being replaced by $\overline{w}_1, \dots, \overline{w}_l$, and setting $n = n_k$, we obtain

$$\lim_{k\to\infty} \left[\Delta_{n_k-1}(x_{n_k}) + \widetilde{w}_{n_k-1}^{-1/2}(x_{n_k})\right]^\rho \bar{\mu}_k^{\gamma_k}(x_{n_k}) = \lim_{k\to\infty} \left[\|x_{n_k} - y_{n_k}\|_2 + w_{n_k}^{-1/2}\right]^\rho \bar{\mu}_k^{\gamma_k}(x_{n_k}) = \infty,$$

for the given choice of ρ. Consequently, by setting $y = x_{n_k}$ in (5.44), it follows that $[\Delta_{n_k-1}(x_{n_k}) + \widetilde{w}_{n_k-1}^{-1/2}(x_{n_k})]^\rho \mu_{n_k-1}^{\gamma_{n_k-1}}(x_{n_k})$ tends to infinity for $k \to \infty$, as claimed. □

Akin to Lemma 3.23, the following lemma states that $\{\mu_n^{\gamma_n}(y)\}$ is uniformly bounded if y is bounded away from the points in the sequence $\{x_n\}$. Note that this result only holds in the surface spline case, as Theorem 3.11 is required, and no result is known for other types of radial basis functions.

Lemma 5.14. *Let ϕ be a conditionally positive definite surface spline of order m from Table 3.1, and let $\{x_n\}$ be a sequence in $\mathbb{R}^d$ with pairwise different points such that $\{x_1, \dots, x_{n_0}\}$ is $\mathcal{P}_m^d$-unisolvent. Further, let $y_0 \in \mathbb{R}^d$ satisfy $\|y_0 - x_n\|_2 \geq \delta$, $n \in \mathbb{N}$, for some $\delta > 0$. Then, there exists $\widetilde{C} > 0$, depending only on y_0 and δ, such that*

$$\mu_n^{\gamma_n}(y_0) \leq \widetilde{C}, \qquad \forall\, n \geq n_0,$$

where $\mu_n^{\gamma_n}$ with $\gamma_n > 0$ is the function given by (5.29).

Proof. Let $B_\delta(y_0) = \{x \in \mathbb{R}^d : \|x - y_0\|_2 < \delta\}$. There exists a compactly supported function $\varphi \in C^\infty(\mathbb{R}^d)$ that takes the value 1 at y_0 and 0 on $\mathbb{R}^d \backslash B_\delta(y_0)$. It follows from Theorem 3.11 that $\varphi \in \mathcal{N}_\phi(\mathbb{R}^d)$.

For any $n \geq n_0$, let $l_n(y_0, \cdot\,)$ be the optimal interpolant to the data $(x_1, 0), \dots, (x_n, 0)$ and $(y_0, 1)$, such that $l_n(y_0, x_i) = \varphi(x_i) = 0$, $i = 1, \dots, n$, and $l_n(y_0, y_0) = \varphi(y_0) = 1$. Similarly, for any $n \geq n_0$, let $l_n^{\gamma_n}(y_0, \cdot\,)$ with $\gamma_n > 0$ denote the optimal regularised least-squares approximant to $(x_1, 0), \dots, (x_n, 0)$, with corresponding weights $w_1, \dots, w_n$, and subject to $l_n^{\gamma_n}(y_0, y_0) = 1$. By definition of $\mu_n^{\gamma_n}$ and the optimality of $l_n^{\gamma_n}(y_0, \cdot\,)$, we then have

$$\begin{aligned}
\mu_n^{\gamma_n}(y_0) &= \|l_n^{\gamma_n}(y_0, \cdot\,)\|_\phi^2 + \frac{1}{n\gamma_n} \sum_{i=1}^n w_i \big(l_n^{\gamma_n}(y_0, x_i)\big)^2 \\
&\leq \|l_n(y_0, \cdot\,)\|_\phi^2 + \frac{1}{n\gamma_n} \sum_{i=1}^n w_i \big(l_n(y_0, x_i)\big)^2 \\
&= \|l_n(y_0, \cdot\,)\|_\phi^2,
\end{aligned}$$

which is bounded by $\widetilde{C} := \|\varphi\|_{\mathcal{N}_\phi}^2$, see Definition 3.10. □

On the basis of the previous lemmas, we can now give the main proof of Theorem 5.8, stating that the sequence generated by Algorithm 5.4 is dense in $\mathcal{X}$. Because of the established similarity of the algorithm to Gutmann's RBF method, it can be carried out along the lines of the proof of Theorem 3.17.

Proof of Theorem 5.8. Assume that there is $y_0 \in \mathcal{X}$ and $\delta > 0$, such that $B_\delta(y_0) = \{x \in \mathbb{R}^d : \|x - y_0\|_2 < \delta\}$ does not contain any x_n, $n \in \mathbb{N}$. According to the iteration step of Algorithm 5.4, it then holds

$$g_n^{\gamma_n}(x_{n+1}) \leq g_n^{\gamma_n}(y_0), \qquad n \geq n_0,$$

where $\gamma_n > 0$. Moreover, since f_n^* is assumed to satisfy condition (5.34) for infinitely many $n \in \mathbb{N}$, there exists a subsequence $\{x_{n_k}\}_{k\in\mathbb{N}}$ such that

$$s_{n_k-1}^{\gamma_{n_k-1}}(x_{n_k}) - f_{n_k-1}^* > \tau \big\| s_{n_k-1}^{\gamma_{n_k-1}} \big\|_\infty \big[\Delta_{n_k-1}(x_{n_k}) + \widetilde{w}_{n_k-1}^{-1/2}(x_{n_k})\big]^{\rho/2},$$

for the specified quantities τ and ρ, and the functions Δ_{n_k-1} and $\widetilde{w}_{n_k-1}$ as given by (3.34) and (5.35), respectively, for $n = n_k - 1$. Now, the sequence $\{x_{n_k}\}$ has a convergent subsequence which, without loss of generality, shall be denoted again by $\{x_{n_k}\}$. Since each x_{n_k}, $k \in \mathbb{N}$, minimises g_{n_k-1} on $\mathcal{X}\backslash\{x_1, \ldots, x_{n_k-1}\}$, the same reasoning as in the proof of Theorem 3.17, see inequalities (3.37) - (3.39), then leads to the inequality

$$\begin{aligned} &\mu_{n_k-1}^{\gamma_{n_k-1}}(x_{n_k})\big[\Delta_{n_k-1}(x_{n_k}) + \widetilde{w}_{n_k-1}^{-1/2}(x_{n_k})\big]^{\rho} \\ &\qquad \leq \mu_{n_k-1}^{\gamma_{n_k-1}}(y_0)\left[\big[\Delta_{n_k-1}(x_{n_k}) + \widetilde{w}_{n_k-1}^{-1/2}(x_{n_k})\big]^{\rho/2} + \frac{2}{\tau}\right]^2, \end{aligned} \tag{5.45}$$

which renders a contradiction by virtue of Lemmas 5.12 - 5.14. In particular, on the one hand, Lemma 5.13 reveals that the left-hand side of (5.45) converges to infinity for $k \to \infty$. On the other hand, Lemma 5.14 shows that $\mu_n^{\gamma_n}(y_0)$ is bounded above by some constant independent of n, which together with the uniform boundedness of $\Delta_{n_k-1}(x_{n_k})$ on $\mathcal{X}$ and the weights being bounded away from zero implies that the right-hand side of inequality (5.45) is bounded above by a constant independent of k. Hence, due to this contradiction, we can deduce that $B_\delta(y_0)$ must contain a point of the sequence $\{x_n\}$, so that, eventually, $\{x_n\}$ is dense in the compact set $\mathcal{X}$. □

5.3.3 Practical Issues

On the analogy of Section 3.3.3 for Gutmann's original RBF method, we now also present some important practical aspects on the implementation of the RBF method for noisy objective functions in Algorithm 5.4. Due to the way in which the method is constructed, some of these aspects may be handled in the same manner as in the original method, while others require considerably modifications as a result of the incorporation of noise.

5.3.3.1 Initialisation

As the initialisation of Algorithm 5.4 is unaffected by noise, the same issues as delineated for Gutmann's RBF method essentially apply. Most importantly, the user is thus confronted with the choice of a suitable radial basis function ϕ in accordance with the smoothness of the underlying objective function $\hat{f}$, where surface splines provide the most convenient option if no information is available. Also, the generation of a unisolvent set of initial points by some strategy is required, where a transformation of the parameter space with its sample points to the unit hypercube is highly favoured if box constraints are used.

5.3.3.2 Optimisation of Subproblems

As in Gutmann's RBF method, the determination of a new evaluation point x_{n+1} in an iteration of Algorithm 5.4 requires to solve several subproblems. Here, however, the construction of as suitable regularised least-squares approximant $s_n^{\gamma_n}$, which is as smooth as possible but within the designated distances to the noisy observations, needs to derived by a different technique than an interpolant, due to the additionally introduced auxiliary problem. Once constructed, the minimisation of the surface $s_n^{\gamma_n}$ as well as the resulting functions $\mu_n^{\gamma_n}$ and $g_n^{\gamma_n}$ on $\mathcal{X}$ may then be treated in a similar manner as in the method for deterministic objective functions, subject to minor adjustments.

Construction of Response Surface

To construct a regularised least-squares approximant $s_n^{\gamma_n}$, recall that we need to solve the auxiliary problem (5.22) whose constraints depend on the solution of the regularised linear system (5.13). In particular, the system entails that the residual vector $(s_n^{\gamma_n}(x_1) - \hat{f}(x_1), \dots, s_n^{\gamma_n}(x_n) - \hat{f}(x_n))^\top$ equals $-\gamma W^{-1}\lambda$, such that we may rewrite the auxiliary problem in vector notation as

$$\begin{aligned} \max_{\gamma>0} \quad & \gamma \\ \text{s.t.} \quad & \gamma W^{-1}\lambda \preceq \epsilon \\ & -\gamma W^{-1}\lambda \preceq \epsilon, \end{aligned} \tag{5.46}$$

where '$\preceq$' denotes the componentwise 'less than or equal to' sign between vectors, and $\epsilon = (\epsilon_1, \dots, \epsilon_n)^\top$.

Since problem (5.46) does not allow for a solution in closed form and an exact one is unnecessary and expensive to compute (each new evaluation of the constraints requires $\mathcal{O}(n^3)$ operations for a Cholesky factorisation of (5.13), followed by $\mathcal{O}(n^2)$ to obtain the coefficients), we intend to solve (5.46) approximately by a backtracking strategy, see Algorithm 5.15. Main idea is to start with a sufficiently large value of γ and successively decrease it by some discretisation, until all constraints are satisfied for the first time or a lower bound is reached. However, to improve its efficiency and derive a suitable discretisation, we first observe that by use of the QR decomposition of the polynomial basis matrix P, cf. (3.40), we have

$$-\gamma W^{-1}\lambda = -\gamma W^{-1} Q_2 \big(Q_2^\top (\Phi + \gamma W^{-1}) Q_2\big)^{-1} Q_2^\top \hat{f}_X, \tag{5.47}$$

where $Q_2 \in \mathbb{R}^{n \times (n-\widetilde{m})}$ and $\hat{f}_X = (\hat{f}(x_1), \dots, \hat{f}(x_n))^\top \in \mathbb{R}^n$. The right-hand side of (5.47) may then be further dismantled by the eigenvalue decompositions (both matrices on the left-hand sides are positive definite and symmetric)

$$Q_2^\top W^{-1} Q_2 = V D_{W^{-1}} V^\top, \qquad \text{and} \qquad \widetilde{Q}_2^\top \Phi \widetilde{Q}_2 = \widetilde{V} D_\Phi \widetilde{V}^\top, \tag{5.48}$$

where $D_{W^{-1}}$ and D_Φ are diagonal with the eigenvalues of $Q_2^\top W^{-1} Q_2$ and $\widetilde{Q}_2^\top \Phi \widetilde{Q}_2$, respectively, V and $\widetilde{V}$ are orthogonal with the corresponding eigenvectors, and $\widetilde{Q}_2 = Q_2 V D_{W^{-1}}^{-1/2}$. Specifically, by first decomposing $Q_2^\top W^{-1} Q_2$, factoring out the term $V D_{W^{-1}} V^\top$, and subsequently decomposing $\widetilde{Q}_2^\top \Phi \widetilde{Q}_2$, we get

$$\begin{aligned} -\gamma W^{-1}\lambda &= -\gamma W^{-1} \widetilde{Q}_2 \big(\widetilde{Q}_2^\top \Phi \widetilde{Q}_2 + \gamma I \big)^{-1} \widetilde{Q}_2^\top \hat{f}_X \\ &= -W^{-1} \widetilde{Q}_2 \widetilde{V} \operatorname{diag}\Big(\frac{\gamma}{\tilde{\lambda}_1 + \gamma}, \ldots, \frac{\gamma}{\tilde{\lambda}_{n-\tilde{m}} + \gamma} \Big) \widetilde{V}^\top \widetilde{Q}_2^\top \hat{f}_X, \end{aligned} \tag{5.49}$$

where $\tilde{\lambda}_1 \geq \ldots \geq \tilde{\lambda}_{n-\tilde{m}}$ denote the eigenvalues of the matrix $\widetilde{Q}_2^\top \Phi \widetilde{Q}_2$. In particular, computing (5.49) thus requires to perform the eigenvalue decompositions (5.48) in an iteration n only once, using $\mathcal{O}(n^3)$ operations, and each update of γ can then be carried out in $\mathcal{O}(n^2)$ operations. Moreover, it follows from representation (5.49) that the term $-\gamma W^{-1}\lambda$ is of order $\mathcal{O}(\gamma/(\tilde{\lambda}_{n-\tilde{m}} + \gamma))$, which may be used to advantage to reparameterise the discretisation of γ according to $\tilde{\gamma} = \gamma/(\tilde{\lambda}_{n-\tilde{m}} + \gamma)$. Hence, altogether, we have the following algorithm for approximately solving the auxiliary problem (5.46).

Algorithm 5.15. *(Auxiliary Problem).*

```
Choose $\tilde{\gamma}^{\min} > 0$, $\tilde{\gamma}^{\max} \in (\tilde{\gamma}^{\min}, 1)$ and a step size $\Delta\tilde{\gamma} \in (\tilde{\gamma}^{\min}, \tilde{\gamma}^{\max})$.

Set $\tilde{\gamma} = \tilde{\gamma}^{\max}$, and compute the eigenvalue decompositions (5.48).

while $\tilde{\gamma} \geq \tilde{\gamma}^{\min}$ do

    Set $\gamma = \tilde{\gamma}/(\tilde{\lambda}_{n-\tilde{m}} - \tilde{\gamma})$, and compute $\gamma W^{-1}\lambda$ using (5.49).
    if $-\epsilon \preceq -\gamma W^{-1}\lambda \preceq \epsilon$ or $\tilde{\gamma} = \tilde{\gamma}^{\min}$
        Set $\gamma^* = \gamma$ and stop.
    else
        Set $\tilde{\gamma} = \max\{\tilde{\gamma} - \Delta\tilde{\gamma}, \tilde{\gamma}^{\min}\}$.
    end if

end while
```

Note that upon terminating Algorithm 5.15, the coefficient c still needs to be calculated. Using the QR decomposition of P, this can be done by solving $Rc = Q_1^\top(\hat{f}_X - (\Phi + \gamma^* W^{-1})\lambda)$.

Determination of Next Evaluation Point

Unlike in the case of interpolation where the function μ_n is not defined at the sample points $x_1, \dots, x_n$ due to infinite discontinuities, the undefined points of $\mu_n^{\gamma_n}$ can be removed continuously by considering the finite values $\mu_n^{\gamma_n}(x_i)$, $i = 1, \dots, n$, according to (5.31). It is therefore not absolutely necessary to derive an equivalent form of the function $\mu_n^{\gamma_n}$ (and thus also of $g_n^{\gamma_n}$), which is then defined on the entire feasible domain $\mathcal{X}$. Nevertheless, in an analogous manner to Proposition 3.24, we can state the following identity, giving $\mu_n^{\gamma_n}$ (and $g_n^{\gamma_n}$) a more intuitive interpretation and allowing for a more efficient computation of the respective subproblem in the RBF method for noisy objective functions.

Proposition 5.16. *For $\gamma_n > 0$, the function $v_n^{\gamma_n}$ defined by*

$$v_n^{\gamma_n}(y) := \left[\phi(0) - \begin{pmatrix} m_n(y) \\ \pi(y) \end{pmatrix}^\top \begin{pmatrix} \Phi + n\gamma_n W^{-1} & P \\ P^\top & 0 \end{pmatrix}^{-1} \begin{pmatrix} m_n(y) \\ \pi(y) \end{pmatrix}\right], \qquad y \in \mathbb{R}^d,$$

is identical to $1/\mu_n^{\gamma_n}$ on $y \in \mathbb{R}^d \backslash \{x_1, \dots, x_n\}$. Moreover, $v_n^{\gamma_n}$ can be continuously extended at the sample points $x_1, \dots, x_n$ by the finite values $v_n^{\gamma_n}(x_i) = 1/\mu_n^{\gamma_n}(x_i)$, $i = 1, \dots, n$.

Proof. The proof is the same as the one for Proposition 3.24, upon replacing the interpolation system (3.26) by the regularised linear system (5.26), and using definition (5.29). The continuous extensions follow immediately from relation (5.31). □

Given $\gamma_n > 0$, $\mu_n^{\gamma_n}(y)$ is positive and finite for $y \in \mathcal{X}$, so that $v_n^{\gamma_n}(y) = 1/\mu_n^{\gamma_n}(y) > 0$. Thus, in the same way as v_n, the function $v_n^{\gamma_n}$ can be regarded as a measure of uncertainty in the approximating model $s_n^{\gamma_n}$. However, besides the distance to the sample points $x_1, \dots, x_n$, as quantified by the chosen radial basis function, the error $v_n^{\gamma_n}(y)$ at y is further influenced by the inherent noise resulting from inexact function values. In particular, this is reflected by the fact that it does not vanish at the sample points, i.e. by $v_n^{\gamma_n}(x_i) > 0$ for $i = 1, \dots, n$.

The minimisers of $\mu_n^{\gamma_n}$ on $\mathcal{X}$ correspond to the maximisers of $v_n^{\gamma_n}$ on $\mathcal{X}$. Hence, if $f_n^* = -\infty$, the subproblem of minimising $\mu_n^{\gamma_n}$ on $\mathcal{X}$ for finding the next evaluation point x_{n+1} can be replaced with maximising $v_n^{\gamma_n}$ on $\mathcal{X}$. If $-\infty < f_n^* < \min_{y \in \mathcal{X}} s_n^{\gamma_n}(y)$, then we can maximise the utility function

$$h_n^{\gamma_n}(y) := \frac{v_n^{\gamma_n}(y)}{\left[s_n^{\gamma_n}(y) - f_n^*\right]^2}, \qquad y \in \mathcal{X},$$

as it corresponds to minimising $g_n^{\gamma_n}$ on $\mathcal{X}$. Equivalently, the minimisation of $-\log h_n^{\gamma_n}$ on $\mathcal{X}$ may be employed to avoid numerical issues when $h_n^{\gamma_n}$ is very small.

Eventually, since the functions $s_n^{\gamma_n}$, $v_n^{\gamma_n}$ and $h_n^{\gamma_n}$ pose the same smoothness as $\phi(\|\cdot\|_2)$, we may also conclude that

$$s_n^{\gamma_n}, v_n^{\gamma_n}, h_n^{\gamma_n} \in \begin{cases} C^{\nu-1}(\mathbb{R}^d), & \text{for surface splines,} \\ C^\infty(\mathbb{R}^d), & \text{for (inverse) multiquadrics and Gaussians,} \end{cases}$$

which thus allows to use essentially the same algorithms for solving the subproblems as applied in the deterministic case.

5.3.3.3 Choice of Target Value

Conforming with the choice of target values for the RBF method in Subsection 3.3.3.3, f_n^* is most effectively set by repeatedly employing short cycles, beginning with a low value to enhance a global search and ending with $f_n^* = \min_{y \in \mathcal{X}} s_n^{\gamma_n}(y)$ to initialise a purely local one. In particular, disregarding any potential adjustment that might be necessary to guarantee a well-defined global minimiser x_{n+1}, we may adopt both suggested strategies as well for the RBF method with the noisy function $\hat{f}$. However, one has to bear in mind that for an admissible choice $f_n^* = \min_{y \in \mathcal{X}} s_n^{\gamma_n}(y)$, we now have to make sure that $f_n^* < \min_{1 \le i \le n} s_n^{\gamma_n}(x_i)$. This can be checked in exactly the same way as in the case of an interpolation, by replacing $f_n^{\min}$ with $\min_{1 \le i \le n} s_n^{\gamma_n}(x_i)$ and using $\max_{1 \le i \le n} s_n^{\gamma_n}(x_i)$ instead of $f_n^{\max}$.

5.3.4 Numerical Analysis

In this section, we demonstrate the practicability of the noisy RBF method on problems that arise from approximatively solving an underlying deterministic optimisation problem by means of the Monte Carlo VSAA strategy. To show the suitability of the method on rather well-behaved problems, we first assess its performance on several relevant test problems and for different sample size schedules $\{N_k\}$. We then also examine the more intricate calibration of the Hull-White model at a selected date, where we additionally compare the VSAA strategy under different sample size schedules with the simpler SAA strategy.

Note that, as in the noise-free setup, we perform all numerical computations in MATLAB, Version 8.2 (R2013b), on the IRIDIS Compute Cluster at the University of Southampton.

Implementation of Method

To all considered problems, we apply the noisy RBF method with the radial basis function $\phi(r) = r^2 \log r$ and initialise the method at the corner points and the midpoint of the box-constrained parameter spaces $\mathcal{X}$, if not further specified. For the main optimisation procedure, the latter are transformed to the unit hybercube $[0,1]^d$ and only referred to for the evaluation of the objective function.

In each iteration of the method, we first construct the response surface $s_n^{\gamma_n}$ by solving the auxiliary problem (5.22), based on the set of available function values $\hat{f}_{N_k}(x_k)$, the derived error bounds $\epsilon(x_k)$ and the associated weights w_k, $k = 1, \dots, n$. As the applied error bounds depend on the considered problem class at hand, their exact forms are described in the respective sections below, using our insights of Subsection 2.3.2. Note, however, that we opt to use pointwise error bounds in both problem classes as they are easier to derive and provide tighter bounds, while the weights w_k will always be set equal to the reciprocal of the (estimated) variances of the respective objective function values. The auxiliary problem is then eventually solved via Algorithm 5.15, where we set the involved parameters as $\tilde{\gamma}^{\min} = 0.0001$, $\tilde{\gamma}^{\max} = 0.99$, and $\Delta\tilde{\gamma} = 0.0001$ for $\tilde{\gamma} \le 0.01$ and $\Delta\tilde{\gamma} = 0.01$ for $\tilde{\gamma} > 0.01$.

Upon the construction, the response surface $s_n^{\gamma_n}$ as well as either $-\log v_n^{\gamma_n}$ or $-\log h_n^{\gamma_n}$ are minimised on $[0,1]^d$ by the DIRECT algorithm (the maximum number of function evaluations and iterations are 300 and 30, respectively), enhanced by a local search from the best point found, and a local search from the sample point with the lowest response surface value. To carry out both local searches, we use the solver 'fmincon' with its default settings from the optimisation toolbox in MATLAB. Finally, concerning the choice of suitable target values, we set f_n^* for all problem instances according to strategy II with $l^{\text{cy}} = 4$ and $\tilde{\Delta}_n^{(b)}$, cf. Subsection 3.3.3.3.

5.3.4.1 Test Problems

We first assess the performance of the noisy RBF method on a number of inexpensive test functions that are itself constructed from a set of deterministic test functions to account for the presence of noise. The underlying set of deterministic test functions comprises the widely used test set proposed by Dixon and Szegö [1978], which has already been employed by Gutmann [2001b] for evaluating the original RBF method, as well as the Beale, the Freudenstein and Roth and the Bard test functions from the collection of Moré et al. [1981], which constitute least-squares functions of different nature (the Beale and Bard functions are fairly smooth, whereas the Freudenstein and Roth function is highly nonlinear). All deterministic test functions[31], along with their dimension, the domain and the number of local and global minima, are given in Table 5.1.

Function	Dimension	Domain	No. of local minima	No. of global minima
Beale	2	$[-4.5, 4.5]^2$	2	1
Branin (Mod.)	2	$[-5, 10] \times [0, 15]$	3	1
Freudensteinroth	2	$[-10, 10]^2$	2	1
Goldstein-Price	2	$[-2, 2]^2$	4	1
Bard	3	$[-0.25, 0.25] \times [0.1, 2.5]^2$	2	1
Hartman 3	3	$[0, 1]^3$	4	1
Shekel 5	4	$[0, 10]^4$	5	1
Shekel 7	4	$[0, 10]^4$	7	1
Shekel 10	4	$[0, 10]^4$	10	1
Hartman 6	6	$[0, 1]^6$	4	1

Table 5.1: Test functions, their dimension, the domain, and the number of local and global minima.

To introduce noise caused by the VSAA strategy into each of the considered deterministic problem instances, we distinguish between functions from the Dixon-Szegö test set and

[31] Note that we have modified the original Branin test function to allow for a unique global minimiser, as suggested by Forrester et al. [2008], Section A.2.

functions given in least-squares form. For the first class of test functions, we approximate the objective function f at the k-th evaluation by $\hat{f}_{N_k}(x_k) = f(x_k) + \frac{1}{N_k}\sum_{i=1}^{N_k} Z_i^k$, $x_k \in \mathcal{X}$, where the Z_i^k are independent and standard normally distributed random variables, independent of previous random samples. This entails that the error bounds may be computed as $\epsilon(x_k) = \sqrt{2\operatorname{Log}(N_k)}/\sqrt{N_k}$ and that the weights may be set as $w_k = N_k$. For the second class, we let the Monte Carlo approximation take place on the residuals, such that the Z_i^k's are standard normally distributed random vectors and the sample average $\frac{1}{N_k}\sum_{i=1}^{N_k} Z_i^k$ is added to the vector of residuals $r(x_k)$, rather than to the resulting objective function value $f(x_k)$. The error bounds are then derived as $\epsilon(x_k) = L_g(x_k)\sqrt{2\operatorname{Log}(N_k)}/\sqrt{N_k}$, where $L_g(x_k)$ denotes the (estimated) local Lipschitz constant of the squared 2-norm transformation g at x_k, and the weights are set to $w_k = N_k/(4\hat{r}_{N_k}(x_k)^\top \hat{r}_{N_k}(x_k))$ (using the delta method), where $\hat{r}_{N_k}(x_k) = r(x_k) + \frac{1}{N_k}\sum_{i=1}^{N_k} Z_i^k$.

Finally, to avoid overly large differences in function values that may negatively affect the quality of approximation, we use for all test functions not belonging to the Shekel and Hartman families a plog-transformation. To this end, the (estimated) variances of the objective function values and the (estimated) Lipschitz constants in the corresponding error bounds have to be adjusted accordingly, using the delta method and the gradient, respectively.

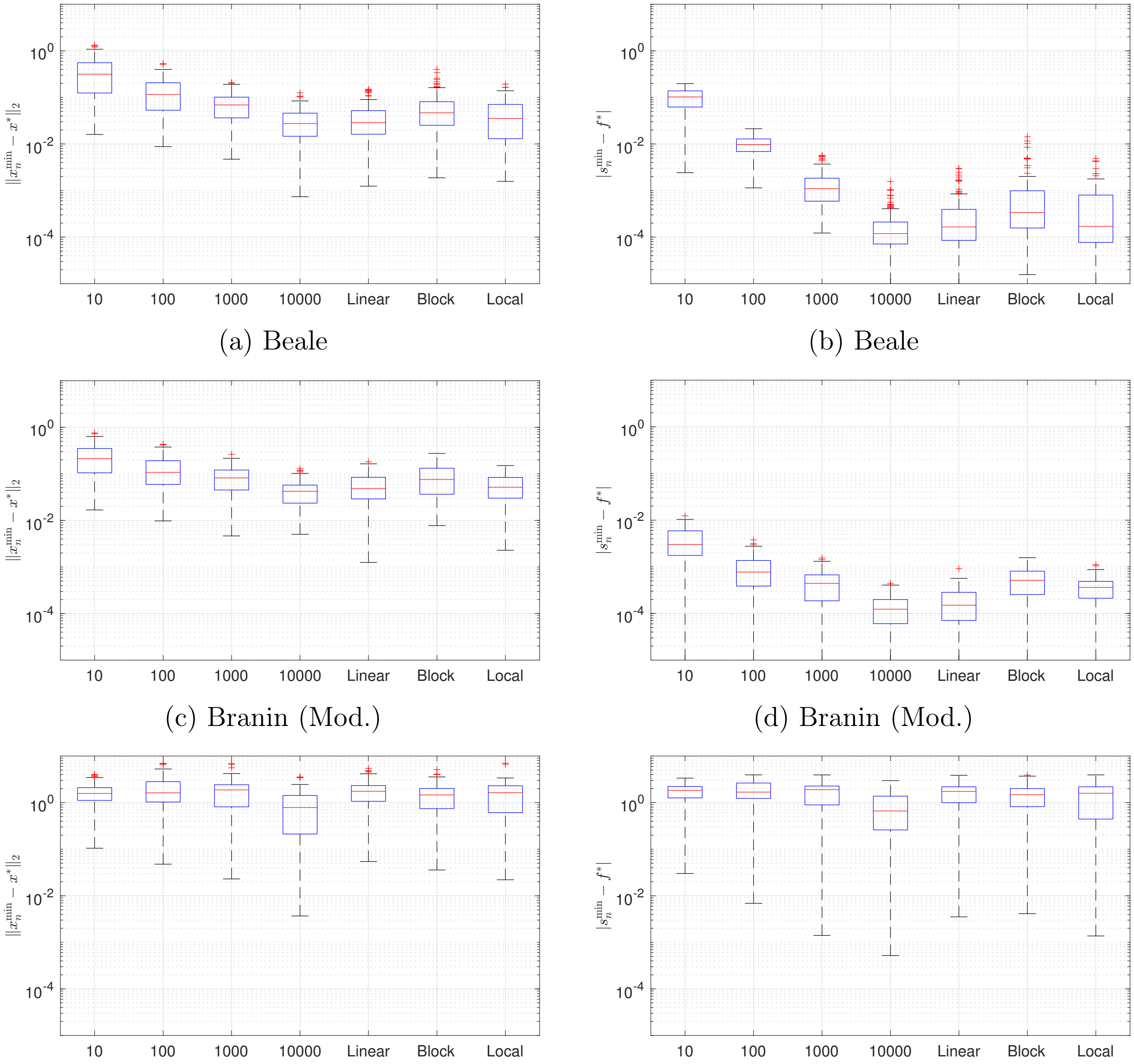
Numerical Results

On the modified test problems, we then apply the noisy RBF method under different sample size schedules $\{N_k\}$. The employed strategies are listed in Table 5.2 and essentially consist of four fixed strategies, where the schedule of sample sizes $\{N_k\}$ is fixed at various levels throughout the entire optimisation procedure (yet a new random sample is drawn at each function evaluation), and three variable strategies, where N_k is either increased linearly or blockwisely along the function evaluations or whenever the method is supposed to perform a local search step.

Strategy	Schedule
Fixed	Fix N_k at $\{10, 100, 1000, 10000\}$ each.
Linear	$N_k = N_{k-1} + \lfloor 10000/n^{\max} \rfloor$, with $N_0 := 0$.
Block	$N_k = \begin{cases} 10, & 1 \leq k \leq \lfloor \frac{1}{4} n^{\max} \rfloor, \\ 100, & \lfloor \frac{1}{4} n^{\max} \rfloor < k \leq \lfloor \frac{2}{4} n^{\max} \rfloor, \end{cases}$, and $N_k = \begin{cases} 1000, & \lfloor \frac{2}{4} n^{\max} \rfloor < k \leq \lfloor \frac{3}{4} n^{\max} \rfloor, \\ 10000, & \lfloor \frac{3}{4} n^{\max} \rfloor < k \leq n^{\max}. \end{cases}$
Local	$N_k = \begin{cases} 100, & (k - n_0) \bmod (l^{\text{cy}} + 1) \leq l^{\text{cy}} - 1, \\ 10000, & \text{otherwise.} \end{cases}$

Table 5.2: Sample size schedules for the VSAA strategy.

For each of the test problems and each strategy, we run the optimisation procedure 100 times, each time initialised with a different random seed. We stop a single optimisation after $n^{\max} = 200$ function evaluations and record the minimal response surface value (for $n = n^{\max}$) by $s_n^{\min} := \min_{1\leq k\leq n} s_n^{\gamma_n}(x_k)$ and the corresponding sample point $x_n^{\min} := \operatorname{argmin}_{1\leq k\leq n} s_n^{\gamma_n}(x_k)$, where the point x_k with the largest index k is chosen if argmin is not unique. Both quantities are then used to compute the distances $\|x_n^{\min} - x^*\|_2$ and $|s_n^{\min} - f^*|$ to the known global minimiser x^* and the minimum value f^* of the underlying deterministic test problems, respectively. Collected over all 100 runs, we obtain the results as shown in Figure 5.1 by means of box plots.



(a) Beale

(b) Beale

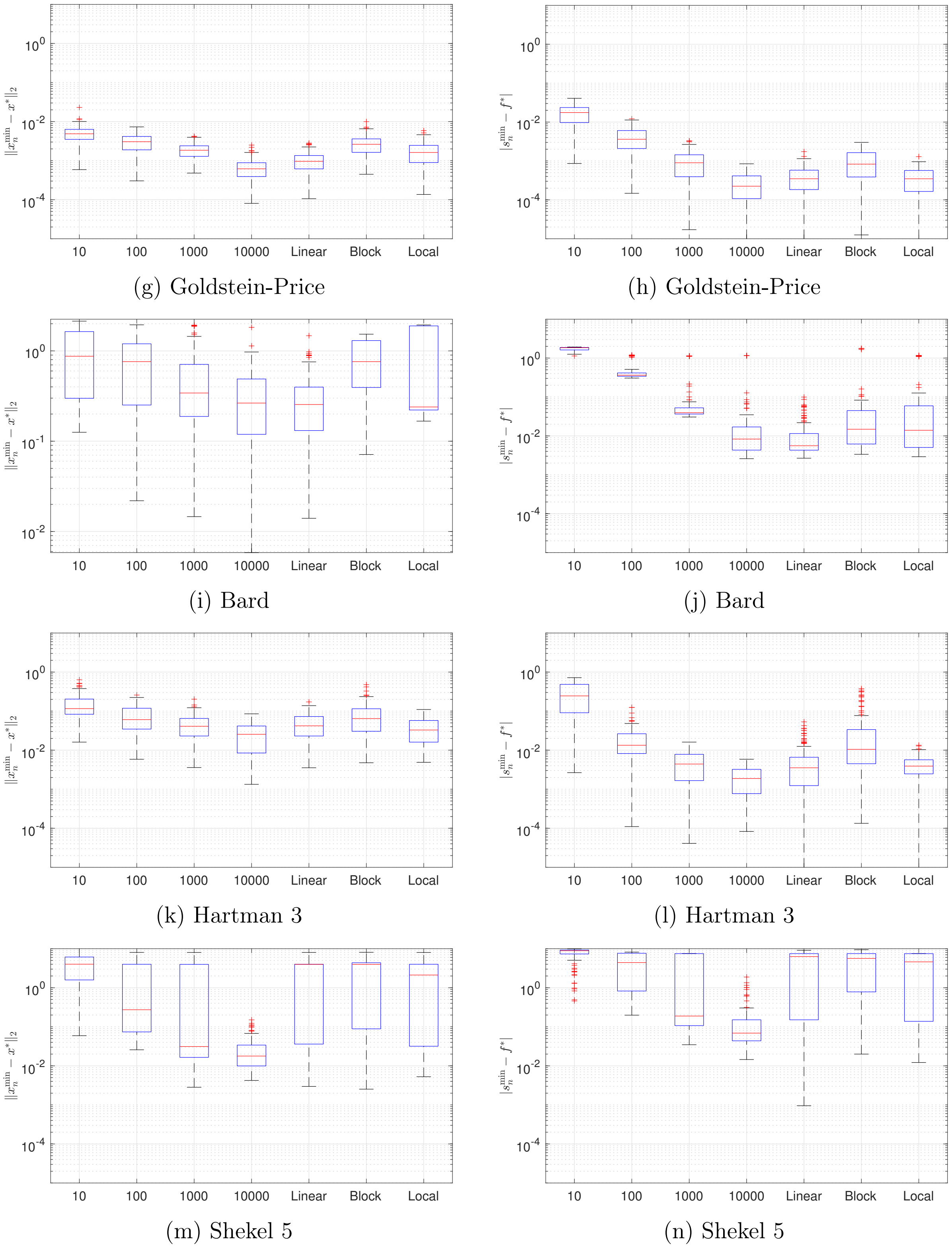
(c) Branin (Mod.)

(d) Branin (Mod.)

(e) Freudenstein and Roth

(f) Freudenstein and Roth

Figure 5.1: Box plots obtained by minimising the test functions with the noisy RBF method under the schedules given in Table 5.2. Each function is minimised with each schedule 100 times; a single optimisation procedure is stopped after 200 function evaluations.



(g) Goldstein-Price

(h) Goldstein-Price

(i) Bard

(j) Bard

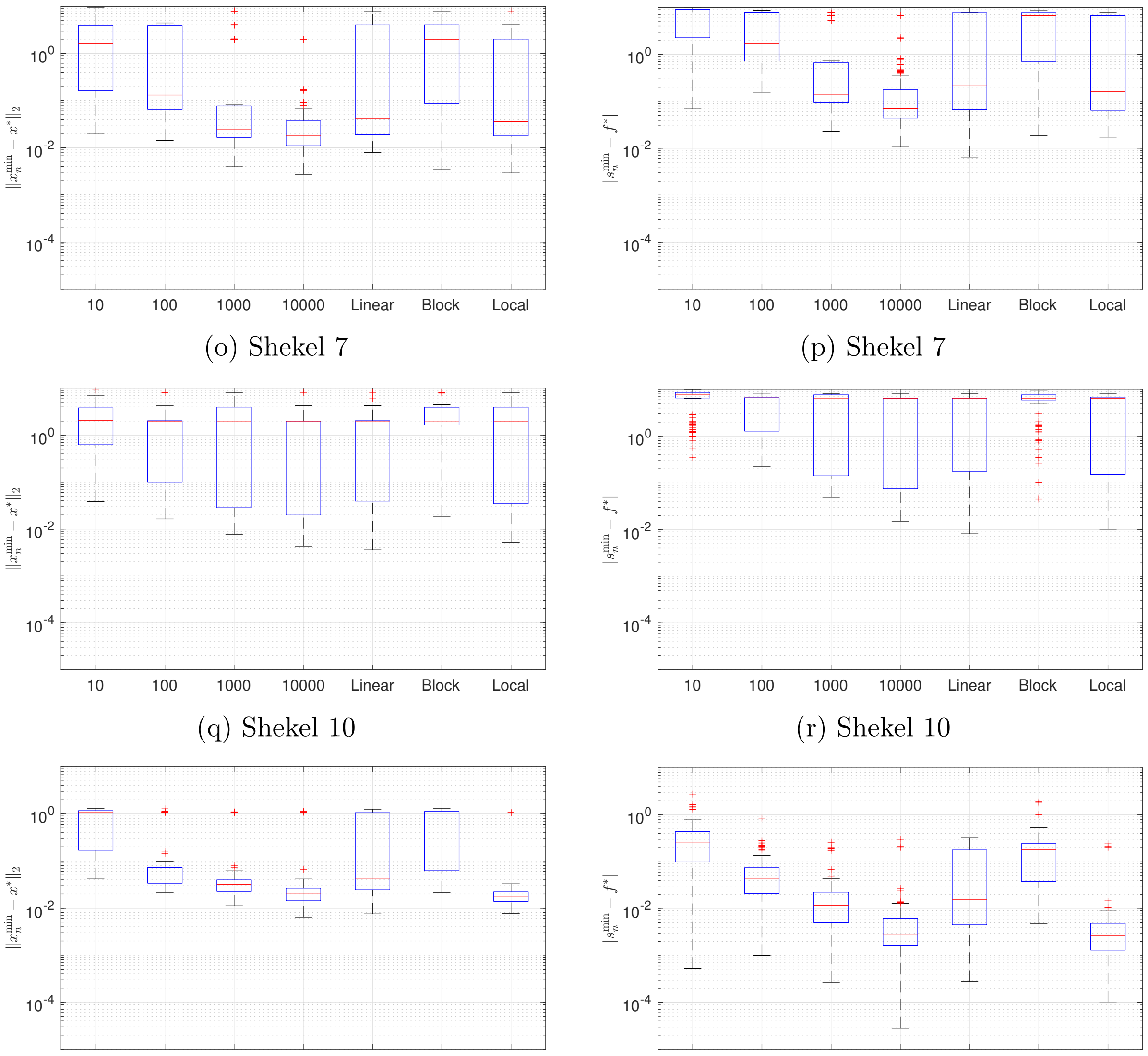
(k) Hartman 3

(l) Hartman 3

(m) Shekel 5

(n) Shekel 5

Figure 5.1: Box plots obtained by minimising the test functions with the noisy RBF method under the schedules given in Table 5.2. Each function is minimised with each schedule 100 times, where a single optimisation procedure is stopped after 200 function evaluations.



(o) Shekel 7

(p) Shekel 7

(q) Shekel 10

(r) Shekel 10

(s) Hartman 6

(t) Hartman 6

Figure 5.1: Box plots obtained by minimising the test functions with the noisy RBF method under the schedules given in Table 5.2. Each function is minimised with each schedule 100 times, where a single optimisation procedure is stopped after 200 function evaluations.

The obtained results clearly indicate that the noisy RBF method performs well on objective functions that are reasonably smooth, see Figures 5.1(a)-(d), (g)-(l) and (s)-(t). For these kind of problems, one may observe that the method is able to locate the neighbourhood of the global minimum already with a small fixed sample size, except for the Bard and the Hartman 6 test function. Moreover, with increasing sample size, the best sample point found by the method and its corresponding response surface value get closer to the global minimiser and the minimum function value, respectively, as it may be expected from a Monte Carlo sampling procedure. Considering the schedules with a variable sample size along a procedure, it is to be noted that by increasing the sample sizes linearly or whenever the method is supposed to perform a local search step, almost the same effect as through fixing the sample size at $N_k = 10000$ may be achieved (except for a linearly increasing schedule on the Hartman 6 test function and for a local schedule on the Bard function). Hence, if an objective function of similar type is indeed expensive[32] to evaluate such that drawing a large set of random vectors requires considerable more time than a small set, it may result worthwhile to employ either of these strategies.

However, similar to the original RBF method in the deterministic case (cf. Gutmann [2001b]), the noisy RBF method has difficulties in minimising test functions from the Shekel family, due to their particular nature with steep wells at the local minima but a largely flat behaviour of the objective functions otherwise. In particular, while the method may only reliably detect the vicinity of the global minimiser of the Shekel 5 and 7 test functions for a strategy with a fixed and sufficiently large sample size, it does not find the respective region of attraction on the Shekel 10 function for any of the applied strategies. Similar results may also be encountered for the Freudenstein and Roth test function and thus be expected for data-fitting problems exhibiting a more complex structure in general, as pointed out in Chapter 3 for deterministic objective function. This is further supported by internal test, which we have additionally carried out.

5.3.4.2 Calibration of Hull-White Model

Eventually, we use the noisy RBF method to calibrate the Hull-White model under the VSAA strategy, where we test the use of different sample size schedules $\{N_k\}$ and contrast the obtained results with those of the SAA strategy for various sizes of N. As in the calibration under the SAA strategy in Section 4.3.2, we choose to calibrate the Hull-White model via swaptions to the set of respective market prices at 28 May 2009 (see the corresponding section for a description of the data) and on the box constraints given in Table 4.3. However, due to the complex fitting structure and the above described difficulty of the noisy RBF method to solve these kind of problems, we opt to solve the resulting optimisation problems by additionally including the best point from the previous calibration into the initial construction of the response surface.

To deal with the noise arising from the application of the VSAA strategy, we we compute the error bounds at any approximate function evaluation $\hat{f}_{N_k}(x_k)$, $x_k \in \mathcal{X}$, $k \in \mathbb{N}$, by $\epsilon(x_k) = L_g(x_k)\sqrt{2\,\mathrm{Log}(N_k)}/\sqrt{N_k}\,\|\widehat{\Sigma}^{1/2}_{N_k}(x_k)\|_2$, where $L_g(x_k)$ is the (estimated) local Lipschitz constant of the squared 2-norm function g at x_k and $\|\widehat{\Sigma}^{1/2}_{N_k}(x_k)\|_2$ is the square root of the largest eigenvalue of the sample covariance matrix $\widehat{\Sigma}_{N_k}(x_k) = \frac{1}{N_k-1}\sum_{i=1}^{N_k}(h(x_k, Z_i^k) - \widehat{\Pi}_{N_k}(x_k))(h(x_k, Z_i^k) - \widehat{\Pi}_{N_k}(x_k))^\top$ at x_k. Further, since the difference $\hat{f}_{N_k}(x_k) - f(x_k)$ is approximatively normally distributed with mean zero and variance

[32] Note that by the way noise is generated in the test set, the time spent on each function evaluation is about the same for the considered sample sizes.

$\frac{1}{N_k}\nabla g(\Pi(x_k) - C^{\mathrm{mkt}})^\top \Sigma(x_k) \nabla g(\Pi(x_k) - C^{\mathrm{mkt}})$, we set the weights w_k used in the construction of approximants equal to the reciprocal of $\frac{4}{N_k}(\widehat{\Pi}_{N_k}(x_k) - C^{\mathrm{mkt}})^\top \widehat{\Sigma}_{N_k}(x_k)(\widehat{\Pi}_{N_k}(x_k) - C^{\mathrm{mkt}})$.

Numerical Results

We proceed in a similar way as for above test problems and apply the noisy RBF method with the VSAA sample size schedules listed in Table 5.2, with $n^{\max} = 150$. Specifically, for each schedule, we initialise an approximating optimisation procedure with underlying random sequence $\{Z_i^k\}$ at 100 different seeds and adopt the noisy RBF method to select new sample points. Each procedure is terminated after three minutes and the obtained minimum of the resulting response surface $x_n^{\min} = \operatorname{argmin}_{1 \le k \le n} s_n^{\gamma_n}(x_k)$ and its value $s_n^{\min} = \min_{1 \le k \le n} s_n^{\gamma_n}(x_k)$ are used to compute the differences $\|x_n^{\min} - x^*\|_2$ and $|\sqrt{\frac{1}{l} s_n^{\min}} - \sqrt{\frac{1}{l} f^*}|$, where x^* and f^* are the global minimiser and minimum function value of the original calibration problem, respectively, as found by the modified RBF method with extended local search, cf. Section 4.3.2. All results collected in this way are then shown in Figure 5.2 using box plots.

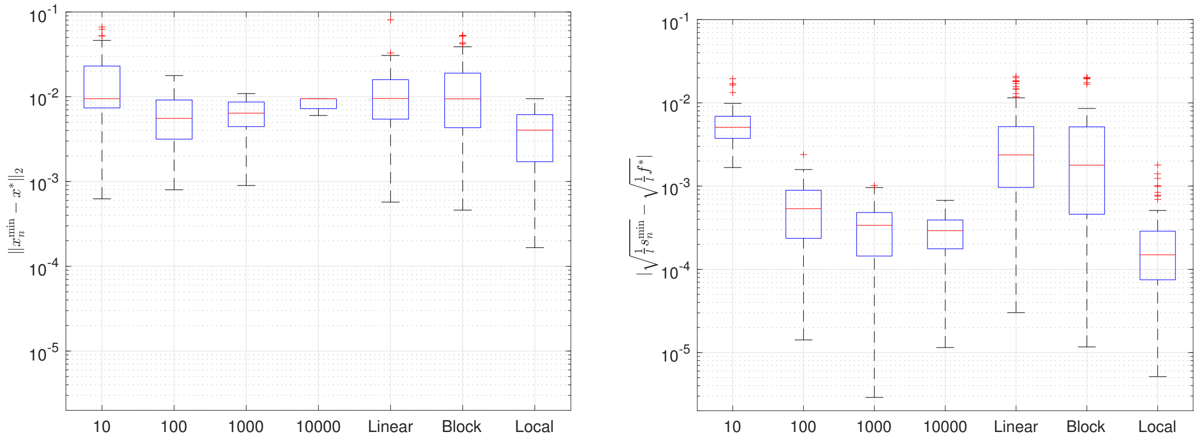


Figure 5.2: Box plots obtained by calibrating the Hull-White model under the VSAA strategy with different schedules $\{N_k\}$ by the noisy RBF method. For each schedule, the calibration problem is solved 100 times with a different seed, where a single optimisation procedure is stopped after a run time of three minutes.

Even if the best sample point of the previous calibration is used in the construction of the initial response surface, the box plots indicate that the noisy RBF method has difficulties to cope with the data-fitting structure of the Hull-White model calibration under the VSAA strategy in a satisfyingly manner. This becomes especially apparent in that the difference in optimal solutions only decreases very slowly to zero and and only if sufficiently many function evaluations can be made, whereas the difference in objective function values diminishes more noticeably, depending on the chosen schedule. In particular, the best approximation to the true global minimiser and the corresponding global minimum may be found by the method if a variable strategy with a local schedule is employed.

In addition to the results obtained under the VSAA strategy, Figure 5.3 shows the equivalent results if the Hull-White model is calibrated under the SAA strategy for various sample sizes N and the original RBF method is used to solve the deterministic optimisation problems. The obtained results are very similar to the ones under the VSAA strategy with a schedule of constant sample sizes $\{N_k\}$, both in terms of differences in parameters and in objective function values. As expected, however, neither the original RBF method under the SAA strategy nor the noisy RBF method under the VSAA strategy are able to compete with the calibration of the Hull-White model under the SAA strategy if the modified RBF method with extended local search is applied, cf. Figure 4.3.

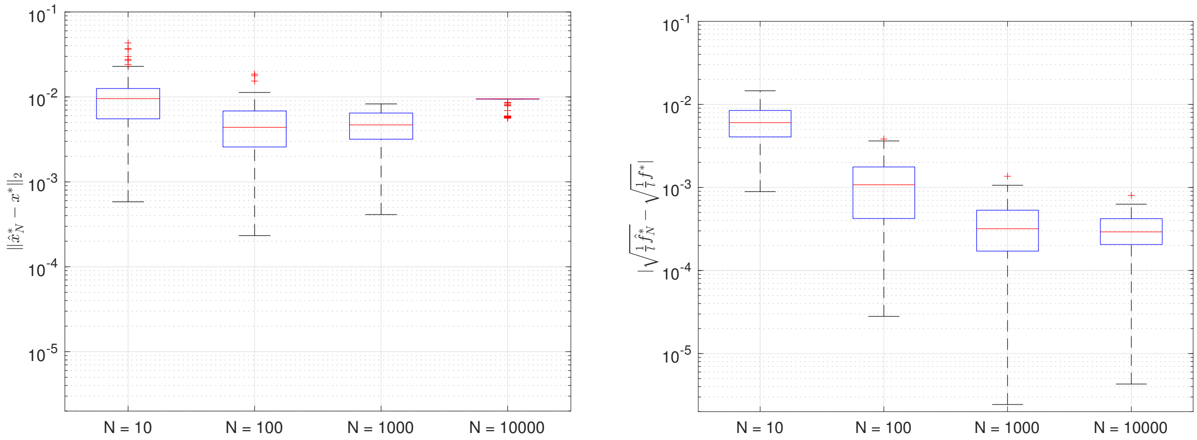


Figure 5.3: Box plots obtained by calibrating the Hull-White model under the SAA strategy with different sample sizes N by the RBF method. For each sample size, the calibration problem is solved 100 times with a different seed, where a single optimisation procedure is stopped after a run time of three minutes.

Chapter 6

Conclusion

Optimisation problems arise in a variety of different specifications from applications across many industries. In this thesis, we have considered the calibration of financial pricing models to market prices, where the corresponding model prices are approximated by standard Monte Carlo methods. By the complex nature of the pricing models, the resulting optimisation problems typically amount to a nonconvex data-fitting problem, thus demanding global optimisation techniques to be solved adequately. Moreover, depending on the use of Monte Carlo simulations, the objective function is expensive to evaluate and either deterministic or noisy, according to the chosen SAA or VSAA strategy for sampling along the optimisation procedure. Since response surface methods are specifically designed for the global optimisation of expensive objective functions, we have opted for this class of methods in the first place. In particular, we have focused on Gutmann's RBF method as it is theoretically well-established and is known to perform well in practice, compared to other global optimisation methods for expensive objective functions. However, the highly nonlinear data-fitting structure and the potentially noisy nature of the objective function require us to considerable extend the original method such that it can be applied suitably to the problems at hand.

To begin with, we have investigated in Chapter 2 the almost sure convergence properties of (optimal) estimators in the SAA and VSAA strategies for approximately solving the original optimisation problem. For the SAA strategy, we have derived almost sure rates of convergence in form of $\sqrt{\mathrm{LLog}(N)}/\sqrt{N}$ that complement the strong consistency of the estimators $\{\hat{f}_N^*\}$ and $\{\hat{x}_N^*\}$ to f^* and x^*, respectively, a matter which has not been investigated so far. To be able to infer these rates, we have applied a version of the LIL in Banach spaces, similar to the case of the functional CLT allowing to derive the asymptotic distributions and the involved rates of optimal estimators. Moreover, since the LIL also implies a certain rate of convergence in probability, this has been used to establish novel universal confidence sets under very mild conditions. By means of the established almost sure rates of convergence, we have then shown that the optimal estimators also converge under mild assumptions to their deterministic counterparts in L_1. However, while for optimal values it is possible to derive an asymptotic rate which essentially corresponds to the rate of the almost sure convergence, no rate can be given for optimal solutions in general. Eventually, from the convergence in mean, we have derived rates of convergence in probability for optimal values. These are rather weak but do not rely on the strong exponential moment conditions as in other approaches.

As for the VSAA strategy, we have shown by use of the SLLN for triangular arrays of Banach space valued random variables that the sequence of approximating objective functions $\{\hat{f}_{N_k}\}$ converges uniformly to f on $\mathcal{X}$, almost surely, as $k \to \infty$. Besides assuming a strictly monotonically increasing schedule $\{N_k\}$, we have been able to derive this result under very mild assumptions on the random function h, in contrast to existing results requiring exponential moments. Moreover, by applying the compact LIL for triangular arrays in Banach spaces, we have further derived, under the same schedule condition and a strong almost fourth centralised moment of h, uniform sample path bounds of the form $\sqrt{\mathrm{Log}(N_k)}/\sqrt{N_k}$ for the difference in objective functions. As far as we are aware, the fact that this result holds uniformly is new; and it may be used to show almost sure convergence of a sequential sampling method incorporating the VSAA strategy on an infinite parameter space $\mathcal{X}$. With respect to both results, it is to be noted that their derivation is based on triangular arrays, which are then transferred to general ones and thus require the schedules $\{N_k\}$ to be strictly monotonically increasing. However, while an improved result with a potentially weaker condition on $\{N_k\}$ can also be given in the case of the SLLN for general arrays, it does not seem to be known whether such a generalisation is also possible for the compact LIL. Eventually, as a by-product of both analysis, we have also been able to show the pointwise strong consistency of $\{\hat{f}_{N_k}\}$ to $f(x)$ on $\mathcal{X}$ and corresponding sample path bounds under assumptions that are comparable than those already given in the literature.

Beginning with Chapter 3, we have considered the global minimisation of expensive objective functions for which response surface methods are most suitable. We have outlined the common literature on this class of optimisation methods in case the objective function is deterministic and closely reviewed the interpolation of the latter by means of radial basis functions to construct adequate response surfaces. Eventually, we have described Gutmann's RBF method in sufficient detail for our later use, including convergence of the method and important practical aspects.

Since the RBF method performs well on reasonably smooth deterministic objective functions but exhibits very slow convergence for more complex ones because of a poorly functioning inherent local search in these cases, we have proposed in Chapter 4 a modified RBF method with extended local search that is more suited for data-fitting problems. Specifically, to exploit the particular data-fitting structure, the method proceeds by interpolating each residual function of the objective with an individual response surface, all of which are then combined to a universal surface by the transformation g. This setup further enables us to define new evaluation points x_{n+1} as Pareto optimal solutions under a particular parameterisation of a multi-objective optimisation problem with a single target value f_n^*, which is consistent with Gutmann's technique for selecting new points. Moreover, to additionally enforce its inherent local search mechanism, the modification is equipped with an external local search component, which is then initialised during optimisation if no sufficient progress towards a potential local minimum is made by the method itself. In a similar fashion as the RBF method, we have shown convergence of the suggested method and presented several convergence results. Eventually, we have conducted stringent computational tests on relevant test problems, the fitting of the Nelson-Siegel and Svensson models, as well as the calibration of the Hull-White model under the SAA strategy, to show the effectiveness of the proposed method.

In Chapter 5, we have addressed the global optimisation of an expensive and noisy objective function, to which end we have first complemented the initially given overview of response surface methods on deterministic functions by their noisy counterpart. Assuming that error bounds on the observed function values are available, we have further discussed several options to approximate a noisy function by radial basis functions for integration into a sequential sampling method. Based on the most suitable choice of regularised (weighted) least-squares approximants, we have then presented the noisy RBF method, where response surfaces are constructed in the smoothest possible way but such that they stay within the error bounds and new evaluation points x_{n+1} are determined by minimising a (weighted) least-squares criterion in terms of a target value f_n^*. Unfortunately, we have not been able to find a condition on f_n^* that guarantees new evaluation points to be well-defined and distinct from any of the previously evaluated points, such as in the case of interpolation. Practical experience, however, indicates that a critical situation rarely occurs, due to the use of error bounds controlling the involved noise. Further on, we have established convergence of the noisy RBF method and provided several convergence results. This, though, has only been possible under some simplified assumption on the error bounds, i.e. by assuming that the error bounds are iteratively readjusted according to $\epsilon_i^{(n)}$ and satisfy $\max_{1\leq i\leq n} \epsilon_i^{(n)} \to 0$ as $n \to \infty$, which then entails $n\gamma_n \to 0$ and thus facilitates to establish the density of the generated sequence $\{x_n\}$ in $\mathcal{X}$. Nevertheless, our intuition and numerical tests tell us that convergence of the method should also follow under the natural assumption that $\epsilon_n \to 0$ as $n \to \infty$. Finally, we have assessed the practicalness of the proposed method by application to relevant test problems and calibrating the Hull-White model under the VSAA strategy. In line with the original RBF method in the deterministic case, our obtained numerical results suggest that the method performs well on reasonably smooth objective functions but may converge only slowly on more intricated problems.

Besides some issues mentioned above, there are further questions related to our investigation in this thesis that still remain open, both from a theoretical and practical perspective.

Considering the optimisation with Gutmann's RBF method and our developments, for instance, it has to be pointed out that their proofs of convergence have only been established for spline type radial basis function and do not extend to (inverse) multiquadrics and Gaussians, as remarked by Gutmann [2001b]. It is still an unresolved issue whether convergence of the methods can be shown by a different technique allowing to also include the latter common basis functions. Moreover, on the practical side, the optimisation of the subproblems in each iteration of the RBF method and our suggested methods is still a matter which could be further improved to enhance their performance. Since all subproblems for determining new evaluation points, i.e. the minimisation of s_n, $-v_n$ and $-h_n$ on $\mathcal{X}$, are cheap to evaluate and possess a particular structure, they could be efficiently exploited by some specifically designed global optimisation method, even if they may have many local minima. Despite this potential for improvement, however, not much effort has been invested since the initial suggestions by Gutmann [2001b] and Björkman and Holmström [2000], see Section 3.3.3, which up to now seem to be the most employed approaches. In fact, to the best of our knowledge, this issue has only been further addressed by Edman [2016] who investigates the calculation of sophisticated lower bounds for a branch-and-bound routine, albeit with less promising numerical results.

Moreover, concerning the practical performance of the noisy RBF method in Chapter 5, our numerical results have shown that there is plenty of room for improvement when it is applied to data-fitting problems or other more involved problems. In particular, similar to the suggested modifications on the original RBF method for deterministic objectives in Chapter 4, it would be very beneficial to exploit the particular data-fitting structure of the objective function and/or to further enhance the local search mechanism of the method by an external technique. While the latter modification could be realised by employing a sophisticated local search algorithm for noisy function values, e.g., implicit filtering (see Kelley [2001]) with a switch akin to Algorithm 4.2, the former requires to replace the interpolation of residuals by a suitable approximation which can then be carried over to the resulting objective function. A further, more elegant, approach (which could also be applied in the deterministic case) would be to find a specific type of basis function that is able to adequately reproduce the data-fitting structure if linearly combined but still preserves the computational efficiency of radial basis functions.

Finally, it is to be mentioned that even though we have pointed out in the introduction of this thesis that an important aspect in the calibration of financial models is the stability of calibrated parameters, we have not investigated this issue any further, due to lack of time. However, one possible approach to tackle this problem could be by using the historical time series of calibrated parameters to describe the stochastic evolution of these parameters over time, which might then be incorporated into a calibration algorithm. In particular, this would result in a multi-objective optimisation problem, which takes into account the robustness requirements of users and where knowledge about the cost of computing objective and gradient values could additionally be built into.

Appendix A

Background on Financial Modelling

In this chapter, we review the calibration of the Hull-White one-factor model and the enhanced fitting of the Nelson-Siegel and Svensson models (as proposed by Banholzer et al. [2017c]), which are used to additionally assess the applicability of our developed methods in Chapters 4 and 5. Since the models apply within an interest rate market, we begin in Section A.1 by briefly recalling the underlying market and essential quantities for further use. In Section A.2, we then describe the Hull-White one-factor model for pricing interest rate derivatives and outline the derivation of (semi-)analytical and Monte Carlo pricing formulas for caps/floors and swaptions, as these are the main derivative instruments through which the model is calibrated. We state the related calibration problems for both pricing methods using the least-squares criterion. Eventually, in Section A.3, we introduce the Nelson-Siegel and Svensson models for modelling the term structure of interest rates, review their traditional fitting procedure and carry out a thorough analysis of the inherent instability of both models, leading to the employed enhanced fitting procedure.

To facilitate the readability of this chapter, we describe the interest rate market and the calibration/fitting problems of the respective models by adopting the standard notations as found in most textbooks, even if similar terminology has already been used earlier in a different context. Nevertheless, as this chapter is self-contained, this shall incur no ambiguity with the notation used previously.

A.1 Interest Rate Market

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space endowed with a filtration $\{\mathcal{F}_t\}_{t\in[0,T']}$ for a fixed horizon date $T' > 0$. Following Musiela and Rutkowski [2005], Chapter 9, we then consider an interest rate market in which an exogenously given risk-free *money-market account* exists. It is denoted by the adapted stochastic process $\{B_t\}_{t\in[0,T']}$, determined by

$$B_t = \exp\big(\int_0^t r_u \,\mathrm{d}u\big), \tag{A.1}$$

where $\{r_t\}_{t\in[0,T']}$ is an adapted process satisfying $\int_0^{T'}|r_u|\,\mathrm{d}u < \infty$, $\mathbb{P}$-almost surely. The rate r_t at which the money-market account accrues interest at time t is referred to as the *instantaneous interest rate*, or briefly as the *short rate*. It describes the short-time rate for risk-free borrowing or lending at time t over the infinitesimal time interval $[t, t+\mathrm{d}t]$. A further fundamental quantity assumed to exist in the market are zero-coupon bonds. By definition, a *zero-coupon bond* of maturity $T \leq T'$ is a financial security paying its holder one unit of cash at a prespecified date T in the future, without intermediate payments. The price at time $t \leq T$ of a zero-coupon maturing T is denoted by $P(t,T)$ and, assuming that it is default-free, we thus have $P(T,T) = 1$.

To guarantee a sufficiently rich and regular interest rate market in the sequel, let us assume that the money-market account and zero-coupon bonds are traded for all maturities $T \in [0,T']$, and that the mapping $T \mapsto P(t,T)$ is differentiable for any fixed $t \leq T$.

Term Structure of Interest Rates

For $0 \leq t \leq T \leq T'$, the continuous rate of return $Y(t,T)$ from holding a zero-coupon bond over the time period $[t,T]$ is known as the *continuously-compounded spot rate* (also termed *zero rate* or *yield*) and defined by

$$Y(t,T) := -\frac{\log P(t,T)}{T-t}.$$

Alternatively to continuous compounding, the rate of return of the zero-coupon bond can be assumed to be simple, resulting in the *simply-compounded spot rate* $L(t,T)$ at time t for the maturity T, and given by

$$L(t,T) := \frac{1}{T-t}\Big(\frac{1}{P(t,T)} - 1\Big). \tag{A.2}$$

Whereas continuously-compounded spot rates are mainly theoretical quantities, simply-compounded ones are directly observable in the market. In particular, the LIBOR and EURIBOR are simply-compounded market rates, describing the rates at which banks in the Eurocurrency markets are prepared to lend money to each other.

At any given date t, the *term structure of interest rates* (also referred to as *zero rate curve* or *yield curve*) is defined as the function $T \mapsto Y(t,T)$, relating the yield $Y(t,T)$ to the maturity T. As it is not directly observable in the market, it needs to be constructed from market data, which is typically done by cash rates, such as LIBOR or EURIBOR, and the prices of liquidly traded interest rate instruments, such as forward rate agreements, futures and swaps, see, e.g., Hirsa [2012], Section 7.7.

Forward Interest Rates

Letting $0 \leq t \leq T \leq S \leq T'$, the *continuously-compounded forward rate* $Y(t,T,S)$ describes the constant rate of interest, agreed on at time t, that accrues continuously during holding a zero-coupon bond over the future interval $[T, S]$, and is thus given by

$$Y(t,T,S) := -\frac{\log P(t,S) - \log P(t,T)}{S-T}.$$

Analogously, the *simply-compounded forward rate* prevailing at time t for the time period $[T, S]$ is defined by

$$L(t,T,S) := \frac{1}{S-T}\Big(\frac{P(t,T)}{P(t,S)} - 1\Big). \tag{A.3}$$

Considering the limit of the continuously-compounded forward rate $Y(t, T, S)$ for $S \to T^+$ leads to the *instantaneous forward rate* $F(t, T)$ prevailing at time t for the maturity T, hence given by

$$F(t,T) := \lim_{S\to T^+} Y(t,T,S) = -\frac{\partial \log P(t,T)}{\partial T}. \tag{A.4}$$

The rate $F(t, T)$ can be interpreted as the interest rate for riskless borrowing or lending over the infinitesimal time interval $[T, T + \mathrm{d}T]$ in the future, as seen from time t. In particular, by setting $T = t$, it must hold that $r_t = F(t, t)$.

No-arbitrage Pricing

Given the existence of an equivalent martingale measure $\mathbb{Q}$ in the market, discounted price processes are required to be martingales under $\mathbb{Q}$, such that the prices of interest rate derivatives are expectations of their discounted payoffs. In particular, denoting by π_t the no-arbitrage price at time t of an attainable contingent claim H settling at time T, i.e. an $\mathcal{F}_T$-measurable random variable, we have by the standard risk-neutral valuation

$$\pi_t(H) = B_t\, \mathbb{E}^{\mathbb{Q}}\big[B_T^{-1} H \,|\, \mathcal{F}_t\big], \qquad t \in [0,T], \tag{A.5}$$

see, e.g., Musiela and Rutkowski [2005], equation (9.26).

Formula (A.5), however, may not be particularly suited for pricing in a stochastic interest rate framework, due to the presence of the factor B_T^{-1} in the expectation and the involved correlation between B_T^{-1} and H. In fact, it is more convenient to apply a change of measure to the T-forward measure $\mathbb{Q}^T$, defined by $\mathrm{d}\mathbb{Q}^T/\mathrm{d}\mathbb{Q} = 1/(P(0, T)B_T)$, which then allows to price the claim equivalently by the formula

$$\pi_t(H) = P(t,T)\mathbb{E}^{\mathbb{Q}^T}\big[H \,|\, \mathcal{F}_t\big], \qquad t \in [0,T], \tag{A.6}$$

cf. Musiela and Rutkowski [2005], Proposition 9.6.2.

In the latter context, the price of a contingent claim settling at time $T \neq S$ may also be expressed under the S-forward measure, which is of particular advantage in pricing deferred and multiple payoffs or in simulating the underlying assets under a unique measure by means of Monte Carlo methods, see, e.g., Musiela and Rutkowski [2005], Corollary 9.6.1. Specifically, if $T \leq S \leq T'$, then the price at time $t \leq T$ equals

$$\pi_t(H) = P(t,S)\mathbb{E}^{\mathbb{Q}^S}\big[P^{-1}(T,S) H \,|\, \mathcal{F}_t\big], \tag{A.7}$$

and, if $S \leq T$ and H is $\mathcal{F}_S$-measurable, then we have for any $t \leq S$,

$$\pi_t(H) = P(t,S)\mathbb{E}^{\mathbb{Q}^S}\big[P(S,T)H \,|\, \mathcal{F}_t\big]. \tag{A.8}$$

A.2 The Hull-White Model

The Hull-White model (Hull and White [1990]) belongs to the most traditional stochastic interest rate models that are based on exogenously specifying the short-rate process $\{r_t\}$ in (A.1). It essentially extends the Vasicek model [1977] by introducing a time-varying parameter that allows for an exact fit to the currently observed term structure of interest rates, thus giving the model a desirable property. In particular, if the process $\{r_t\}$ is specified under the measure $\mathbb{Q}$, the valuation formula (A.5) provides an arbitrage-free family of zero-coupon bonds in only two parameters, from which the prices of the main calibration instruments may be obtained by no-arbitrage techniques, to eventually set up the related calibration.

Short-rate Dynamics

In the Hull-White model, the short-rate process $\{r_t\}$ is assumed to follow the stochastic differential equation (SDE)

$$\mathrm{d}r_t = \big(\vartheta(t) - \kappa r_t\big)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t^{\mathbb{Q}}, \tag{A.9}$$

where κ and σ are positive constants, ϑ is a deterministic function of time and $\{W_t^{\mathbb{Q}}\}$ is a standard Brownian motion under the risk-neutral measure $\mathbb{Q}$.

The specification in (A.9) allows for the following interpretation of the parameters. As a time-varying parameter, ϑ is typically chosen in order to fit the initial term structure of interest rates observed in the market. Moreover, by factoring out the parameter κ in the drift term, $\{r_t\}$ assumes the typical form of a mean-reverting process: it evolves around the time-dependent mean-reversion level $\vartheta(t)/\kappa$ and any significant deviation from it is pushed back to the mean at a rate of κ. Therefore, κ can be interpreted as the mean-reverting speed of $\{r_t\}$, while σ describes the volatility of the process and as such has an opposite effect to κ on the evolution of the rates.

It is well-known that equation (A.9) admits an explicit solution, as the following lemma shows, see Musiela and Rutkowski [2005], Lemma 10.1.2 and p. 395.

Lemma A.1. *The unique solution to the SDE* (A.9) *is given by the formula*

$$r_t = r_s \mathrm{e}^{-\kappa(t-s)} + \int_s^t \mathrm{e}^{-\kappa(t-u)}\vartheta(u)\,\mathrm{d}u + \sigma \int_s^t \mathrm{e}^{-\kappa(t-u)}\,\mathrm{d}W_u^{\mathbb{Q}}. \tag{A.10}$$

Hence, for any $s < t$, the conditional distribution of r_t under $\mathbb{Q}$ with respect to the σ-field $\mathcal{F}_s$ is Gaussian, with conditional expected value

$$\mathbb{E}^{\mathbb{Q}}[r_t \,|\, \mathcal{F}_s] = r_s \mathrm{e}^{-\kappa(t-s)} + \int_s^t \mathrm{e}^{-\kappa(t-u)}\vartheta(u)\,\mathrm{d}u,$$

and conditional variance

$$\mathbb{V}\mathrm{ar}^{\mathbb{Q}}[r_t \,|\, \mathcal{F}_s] = \frac{\sigma^2}{2\kappa}\big(1 - \mathrm{e}^{-2\kappa(t-s)}\big).$$

A.2.1 Zero-coupon Bonds

By means of above short-rate specification, we next establish a general closed-form expression for zero-coupon bond prices in the Hull-White model. By fitting the model to the initial term structure of interest rates, this can then be further reduced to a formula in the parameters κ and σ only. Given the involved dynamics of the bond prices, we eventually apply a change of numeraire to derive the short-rate dynamics under the T-forward measure, which will be useful for the pricing of interest rate derivatives and the simulation of the short rate.

General Zero-coupon Bond Prices

According to formula (A.5), the price at time t of a zero-coupon bond with maturity T is given in the Hull-White model by

$$P(t,T) = \mathbb{E}^{\mathbb{Q}}\big[\mathrm{e}^{-\int_t^T r_s \,\mathrm{d}s} \,|\, \mathcal{F}_t\big], \tag{A.11}$$

where the short rate r_s is given by (A.10). Integrating the exponent in (A.11), applying Fubini's theorems, and using the conditional distribution of the short rate as in Lemma A.1, it thus follows that

$$P(t,T) = A(t,T)\mathrm{e}^{-B(t,T)r_t}, \tag{A.12}$$

where

$$A(t,T) = \exp\Big(\frac{1}{2}\int_t^T \sigma^2 B^2(u,T)\,\mathrm{d}u - \int_t^T \vartheta(u)B(u,T)\,\mathrm{d}u\Big), \tag{A.13}$$
$$B(t,T) = \frac{1}{\kappa}\big(1 - \mathrm{e}^{-\kappa(T-t)}\big),$$

and r_t describes the short rate at time t. In particular, thus note that price of a zero-coupon bond is a monotonically decreasing function of r_t. Moreover, by applying Itô's formula (e.g., Musiela and Rutkowski [2005], Proposition A.9.1) to (A.12), the dynamics of the zero-coupon bond price under $\mathbb{Q}$ can be shown to be specified by

$$\mathrm{d}P(t,T) = P(t,T)\big(r_t\,\mathrm{d}t - \sigma B(t,T)\,\mathrm{d}W_t^{\mathbb{Q}}\big). \tag{A.14}$$

Fitting the Initial Term Structure of Interest Rates

To ensure that the Hull-White model fits the initial term structure of interest rates observed in the market, the unknown function ϑ in formula (A.12) needs to be specified. This is most conveniently done by equating the instantaneous forward rates, as implied by the model via definition (A.4), with their corresponding market rates $F^{\mathrm{M}}(0,T)$ for all maturities $T > 0$, and resulting in

$$\vartheta(T) = \frac{\partial F^{\mathrm{M}}(0,T)}{\partial T} + \kappa F^{\mathrm{M}}(0,T) + \frac{\sigma^2}{2\kappa}(1 - \mathrm{e}^{-2\kappa T}), \tag{A.15}$$

see, e.g., Musiela and Rutkowski [2005], pp. 396. Upon substituting (A.15) into (A.13) and integrating, the following explicit solution for the price of a zero-coupon bond can then be obtained, depending solely on the parameters κ and σ as well as on market implied quantities.

Proposition A.2. *Under the assumption that the function ϑ satisfies* (A.15) *for any $T > 0$ with fixed κ and σ, the price at time t of a zero-coupon bond maturing at time T is given in the Hull-White model by*

$$P(t,T) = A(t,T)\mathrm{e}^{-B(t,T)r_t}, \tag{A.16}$$

where

$$\begin{aligned} A(t,T) &= \frac{P^M(0,T)}{P^M(0,t)} \exp\big(B(t,T)F^M(0,t) - \frac{\sigma^2}{4\kappa}(1 - \mathrm{e}^{-2\kappa t})B^2(t,T)\big), \\ B(t,T) &= \frac{1}{\kappa}\big(1 - \mathrm{e}^{-\kappa(T-t)}\big), \end{aligned}$$

and r_t describes the short rate at time t.

In particular, thus note that $P(0,T) = P^{\mathrm{M}}(0,T)$ for any $T \in [0,T']$, i.e. zero-coupon bond prices are directly observable in the market at time $t = 0$, independent of κ and σ.

Change of Numeraire

From the SDE for zero-coupon bond prices in (A.14), we are further able to derive the dynamics of the process $\{r_t\}$ under the T-forward measure. Specifically, by solving (A.14) via Itô's formula and applying Girsanov's theorem (e.g., Musiela and Rutkowski [2005], Theorem A.15.1), it can be shown that the process $\{W_t^{\mathbb{Q}^T}\}$, given by

$$\mathrm{d}W_t^{\mathbb{Q}^T} = \mathrm{d}W_t^{\mathbb{Q}} + \sigma B(t,T)\,\mathrm{d}t, \tag{A.17}$$

is an $\mathcal{F}_t$-standard Brownian motion under the T-forward measure $\mathbb{Q}^T$ defined by $\mathrm{d}\mathbb{Q}^T/\mathrm{d}\mathbb{Q} = 1/(P(0,T)B_T)$. Combining equation (A.17) and (A.9) thus yields that the dynamics of the short rate process $\{r_t\}$ under $\mathbb{Q}^T$ can be expressed as

$$\mathrm{d}r_t = \big(\vartheta(t) - \kappa r_t - \sigma^2 B(t,T)\big)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t^{\mathbb{Q}^T}. \tag{A.18}$$

In particular, equation (A.18) shows that $\{r_t\}$ is also mean-reverting under the T-forward measure, but with a reversion level corrected by the term $-\sigma^2 B(t,T)/\kappa$, whereas the volatility remains the same. Similar to Lemma A.1, we can then conclude the following result.

Lemma A.3. *The unique solution to the SDE* (A.18) *is given by the formula*

$$r_t = r_s \mathrm{e}^{-\kappa(t-s)} + \int_s^t \mathrm{e}^{-\kappa(t-u)} \big(\vartheta(u) - \sigma^2 B(u,T)\big)\,\mathrm{d}u + \sigma \int_s^t \mathrm{e}^{-\kappa(t-u)}\,\mathrm{d}W_u^{\mathbb{Q}^T},$$

which by fitting the model to the initial term structure of interest rates, i.e. setting ϑ as in (A.15)*, becomes*

$$r_t = r_s \mathrm{e}^{-\kappa(t-s)} + \theta(t) - \theta(s)\mathrm{e}^{-\kappa(t-s)} - M^T(s,t) + \sigma \int_s^t \mathrm{e}^{-\kappa(t-u)}\,\mathrm{d}W_u^{\mathbb{Q}^T},$$

where

$$\theta(t) = F^M(0,t) + \frac{\sigma^2}{2\kappa^2}\big(1 - \mathrm{e}^{-\kappa t}\big)^2, \tag{A.19}$$

$$M^T(s,t) = \frac{\sigma^2}{\kappa^2}\big(1 - \mathrm{e}^{-\kappa(t-s)}\big) - \frac{\sigma^2}{2\kappa^2}\big(\mathrm{e}^{-\kappa(T-t)} - \mathrm{e}^{-\kappa(T+t-2s)}\big). \tag{A.20}$$

Hence, for any $s < t$, the conditional distribution of r_t under $\mathbb{Q}^T$ with respect to the σ-field $\mathcal{F}_s$ is Gaussian, with conditional expected value

$$\mathbb{E}^{\mathbb{Q}^T}[r_t \,|\, \mathcal{F}_s] = r_s \mathrm{e}^{-\kappa(t-s)} + \theta(t) - \theta(s)\mathrm{e}^{-\kappa(t-s)} - M^T(s,t), \tag{A.21}$$

and conditional variance

$$\mathbb{V}\mathrm{ar}^{\mathbb{Q}^T}[r_t \,|\, \mathcal{F}_s] = \frac{\sigma^2}{2\kappa}\big(1 - \mathrm{e}^{-2\kappa(t-s)}\big). \tag{A.22}$$

A.2.2 Pricing of Interest Rate Derivatives

We now provide (semi-)analytical and Monte Carlo pricing formulas for the main calibration instruments caps/floors and swaptions in the Hull-White model. Since their (semi-)analytical valuation can essentially be reduced to the pricing of zero-coupon bonds options, we will first state the respective formula for these products and then give analytical pricing formulas for caps/floors and semi-analytical expressions for swaptions, which nevertheless shall be treated as proper closed-form solutions. Eventually, we establish the respective Monte Carlo prices under convenient forward measures.

A.2.2.1 (Semi-)Analytical Derivative Prices

Zero-coupon Bond Options

A *European call (put) option written on a zero-coupon bond* is a financial derivative that gives its holder the right to buy (sell) a zero-coupon bond at some given future date T for a prespecified strike price K. Accordingly, since the holder of a call option will naturally only exercise his right at time T if the value of the zero-coupon bond $P(T,S)$ with maturity $S \geq T$ is larger than K, the payoff at time T of the option then amounts to

$$H = \big(P(T,S) - K\big)^+.$$

This thus implies by virtue of the valuation formula (A.6) that the no-arbitrage price at time $t \leq T$ of the call option is given under the T-forward measure by

$$\mathrm{ZBC}(t,T,S,K) = P(t,T)\mathbb{E}^{\mathbb{Q}^T}\Big[\big(P(T,S)-K\big)^+ \,\big|\, \mathcal{F}_t\Big]. \tag{A.23}$$

Given the distribution of r_T under the T-forward measure according to Lemma A.3, the expectation in (A.23) can be explicitly computed in the Hull-White model. Specifically, since r_T is conditionally normally distributed, it follows that the random variable $P(T,S)$ is (conditional) lognormal under $\mathbb{Q}^T$, where the mean and variance of $\log P(T,S)$ are given by $\mu_P = \log A(T,S) - B(T,S)\mathbb{E}^{\mathbb{Q}^T}[r_T \,|\, \mathcal{F}_t]$ and $\sigma_P^2 = B^2(T,S)\,\mathbb{V}\mathrm{ar}^{\mathbb{Q}^T}[r_T \,|\, \mathcal{F}_t]$, respectively. Standard calculations involving the lognormal distribution (e.g., Brigo and Mercurio [2006], Appendix D) then show that the price of the call option is given by the following proposition, where $\mathcal{N}(\,\cdot\,)$ denotes the standard normal cumulative distribution function. Note that the price of a corresponding put option is obtained likewise, either by a modification of the payoff or directly by applying the put-call-parity for bond options (e.g., Musiela and Rutkowski [2005], p. 515).

Proposition A.4. *The no-arbitrage price at time t of a European call option with maturity $T \geq t$ and strike K, written on a zero-coupon bond maturing at time $S \geq T$ is given in the Hull-White model by*

$$ZBC(t,T,S,K) = P(t,S)\mathcal{N}(d_1) - KP(t,T)\mathcal{N}(d_2),$$

where

$$\begin{aligned} d_1 &= \frac{\log\big(P(t,S)/(P(t,T)K)\big) + \frac{1}{2}\sigma_P^2}{\sigma_P}, \\ d_2 &= d_1 - \sigma_P, \end{aligned}$$

and

$$\sigma_P = \sqrt{\frac{\sigma^2}{2\kappa}\big(1 - e^{-2\kappa(T-t)}\big)}B(T,S).$$

The price of the corresponding European put option written on a zero-coupon bond is given by

$$ZBP(t,T,S,K) = KP(t,T)\mathcal{N}(-d_2) - P(t,S)\mathcal{N}(-d_1).$$

Interest Rate Caps/Floors

An *interest rate cap (floor)* is a financial derivative in which the buyer of the contract receives interest payments if a particular floating interest rate, typically indexed to the LIBOR, exceeds (falls below) an agreed level K at some prespecified future dates T_i, $i = 1, \dots, n$, and in which no payments occur otherwise. Accordingly, assuming that a cap is settled in arrears, the payoffs[33] at times T_i are

$$H = \tau_i \big(L(T_{i-1}, T_i) - K\big)^+,$$

where $\tau_i = T_i - T_{i-1}$ and the LIBOR $L(T_{i-1}, T_i)$ as defined by (A.2) is determined at the reset date T_{i-1}. Hence, by the valuation formula (A.6), the no-arbitrage price at time $t \leq T_0$ of a cap with reset and payment dates $\mathcal{T} = \{T_0, \dots, T_n\}$ may be written under the T_i-forward measures as

$$\mathrm{Cap}(t, \mathcal{T}, K) = \sum_{i=1}^{n} P(t, T_i) \mathbb{E}^{\mathbb{Q}^{T_i}} \Big[\tau_i \big(L(T_{i-1}, T_i) - K\big)^+ \,\big|\, \mathcal{F}_t \Big].$$

Now, since the payoff of the i-th caplet (i.e. the i-the leg of the cap) at time T_i is an $\mathcal{F}_{T_{i-1}}$-measurable random variable, it can be expressed under the $\mathbb{Q}^{T_{i-1}}$-forward measure, cf. formula (A.8). The price of a single caplet can thus be simplified together with the definition of the rate $L(T_{i-1}, T_i)$ into

$$\mathrm{Caplet}(t, T_{i-1}, T_i, K) = P(t, T_{i-1}) \mathbb{E}^{\mathbb{Q}^{T_{i-1}}} \Big[\big(1 - (1 + K\tau_i) P(T_{i-1}, T_i)\big)^+ \,\big|\, \mathcal{F}_t \Big]. \tag{A.24}$$

Latter expression then reveals that the i-th caplet is equivalent to a European put option with maturity T_{i-1}, strike rate $1/(1+K\tau_i)$, and written on a zero-coupon bond with maturity T_i and notional principal $(1 + K\tau_i)$. Consequently,

$$\mathrm{Caplet}(t, T_{i-1}, T_i, K) = (1 + K\tau_i)\, \mathrm{ZBP}\big(t, T_{i-1}, T_i, 1/(1 + K\tau_i)\big),$$

and summing up the prices of all underlying caplets yields the corresponding cap price. Since the price of a floor can be derived analogously, we arrive at the following valuation formulas for caps and floors, based on the zero-coupon bond option formulas of Proposition A.4.

Proposition A.5. *The no-arbitrage price at time t of a cap with reset and payment dates $\mathcal{T} = \{T_0, \dots, T_n\}$ and cap rate K is given in the Hull-White model by*

$$\mathit{Cap}(t, \mathcal{T}, K) = \sum_{i=1}^{n} (1 + K\tau_i)\, \mathit{ZBP}\big(t, T_{i-1}, T_i, 1/(1 + K\tau_i)\big),$$

where $\tau_i = T_i - T_{i-1}$ for $i = 1, \dots, n$. The price of the corresponding floor is given by

$$\mathit{Floor}(t, \mathcal{T}, K) = \sum_{i=1}^{n} (1 + K\tau_i)\, \mathit{ZBC}\big(t, T_{i-1}, T_i, 1/(1 + K\tau_i)\big).$$

[33] For ease of exposition, we assume the notional to be 1 for caps/floors as well as for swaptions.

Swaptions

A *European payer (receiver) swaption* is an option that gives its holder the right to enter an interest rate swap at a given future date T, in which then a fixed interest rate K is paid (received) in exchange for a floating interest rate, typically indexed to the LIBOR, at some prespecified future dates T_i, $i = 1, \ldots, n$. Considering a payer swaption in its most basic form, the cash flows of the underlying swap occurring at each T_i thus amount to $\tau_i(L(T_{i-1}, T_i) - K)$, where $\tau_i = T_i - T_{i-1}$ with $T_0 = T$, such that the payoff at time T of the payer swaption can be shown to be given by

$$H = \big(1 - \sum_{i=1}^{n} c_i P(T, T_i)\big)^+.$$

cf. Musiela and Rutkowski [2005], Lemma 13.1.1 and p. 522. Applying valuation formula (A.6) to the payoff H, the no-arbitrage price at time t of a payer swaption with expiry T then equals

$$\mathrm{PS}(t, T, \mathcal{T}, K) = P(t, T)\mathbb{E}^{\mathbb{Q}^T}\Big[\big(1 - \sum_{i=1}^{n} c_i P(T, T_i)\big)^+ \,\big|\, \mathcal{F}_t\Big], \tag{A.25}$$

where $\mathcal{T} = \{T_1, \ldots, T_n\}$, and $c_i = \tau_i K$, $i = 1, \ldots, n-1$, and $c_n = 1 + \tau_n K$.

Due to the fact that the price of a zero-coupon bond in the Hull-White model is a monotonically decreasing function of the short rate, cf. formula (A.12), the expected value in (A.25) can be further broken down into more elementary derivative products using a decomposition by Jamshidian [1989]. Specifically, given the monotonicity of $P(T, T_i; r_T)$ in r_T, the payoff H will be exercised at time T if and only if $r_T > r^*$, where the critical short rate r^* is determined (numerically) as the unique solution to

$$\sum_{i=1}^{n} c_i P(T, T_i; r^*) = 1. \tag{A.26}$$

Hence, letting $K_i = P(T, T_i; r^*)$, the latter condition (A.26) then implies that the payoff H can be rewritten as

$$H = \sum_{i=1}^{n} c_i \big(K_i - P(T, T_i; r_T)\big)^+,$$

such that the valuation of a payer swaption has been reduced to the pricing of multiple European put options on zero-coupon bonds with individual strikes K_i. In an analogous way receiver swaptions are evaluated, and we can thus conclude the following pricing formulas for swaptions, along with Proposition A.4.

Proposition A.6. *The no-arbitrage price at time t of a payer swaption with option maturity T, payment dates $\mathcal{T} = \{T_1, \ldots, T_n\}$, and strike K is given in the Hull-White model by*

$$PS(t, T, \mathcal{T}, K) = \sum_{i=1}^{n} c_i \, ZBP(t, T, T_i, K_i),$$

where $c_i = K\tau_i$, $i = 1, \ldots, n-1$, $c_n = 1 + K\tau_n$, $\tau_i = T_i - T_{i-1}$ *with* $T_0 = T$, *and the strikes* K_i *are defined by* $K_i(r^*) = P(T, T_i; r^*)$ *for* r^* *satisfying* (A.26)*. The price of the corresponding receiver swaption is given by*

$$RS(t, T, \mathcal{T}, K) = \sum_{i=1}^{n} c_i \, ZBC(t, T, T_i, K_i).$$

A.2.2.2 Monte Carlo Derivative Prices

Since the payoffs of caps/floors and swaptions are defined as expected values of a function of the short rate at various time points, we can also establish Monte Carlo pricing formulas for these derivatives in the Hull-White model. Accordingly, conforming with (1.3), we can approximate their prices π_t at the current time $t = 0$ by the scalar-valued Monte Carlo estimator

$$\widehat{\pi}_{0,N} = \frac{1}{N} \sum_{i=1}^{N} h(r_{t_1}^{(i)}, \ldots, r_{t_{d_Z}}^{(i)}), \tag{A.27}$$

where the concrete form of the discounted payoff function h as well as the number and distribution of the time points $0 < t_1 < \ldots < t_{d_Z}$ at which the short rate needs to be simulated depends on the considered derivative and the employed pricing measure. Yet, common to both derivative payoffs is the fact that h depends on the value of the short rate at future time points via the price of zero-coupon bonds according to formula (A.16).

Simulation of Short-Rate Path

In the Hull-White model, the exact conditional distribution of the short rate is known and can thus be used to construct a sample path at the required time points $t_1, \ldots, t_{d_Z}$, without any discretisation method such as the Euler or the Milstein scheme. Specifically, by Lemma A.3, we have that for any $0 < s < t$ the value r_t is conditionally normally distributed on r_s under the T-forward measure with expected value (A.21) and variance (A.22). Hence, given an initial value of the short rate r_{t_0} at time $t_0 = 0$, a path of $\{r_t\}$ may be simulated at the time instants $0 < t_1 < \ldots < t_{d_Z} = T^*$ under the forward measure $\mathbb{Q}^{T^*}$ by setting

$$r_{t_i} = r_{t_{i-1}} e^{-\kappa(t_i - t_{i-1})} + \mu_r(t_{i-1}, t_i) + \sigma_r(t_{i-1}, t_i) Z^{(i)}, \tag{A.28}$$

for

$$\begin{aligned}
\mu_r(t_{i-1}, t_i) &= \theta(t_i) - \theta(t_{i-1}) \mathrm{e}^{-\kappa(t_i - t_{i-1})} - M^{T^*}(t_{i-1}, t_i), \\
\sigma_r(t_{i-1}, t_i) &= \frac{\sigma^2}{2\kappa}\big(1 - \mathrm{e}^{-2\kappa(t_i - t_{i-1})}\big),
\end{aligned}$$

where the functions $\theta(\cdot)$ and $M^{T^*}(\cdot, \cdot)$ are given by (A.19) and (A.20), respectively, and $Z^{(1)}, \ldots, Z^{(d_Z)}$ are independent, standard normally distributed random variables. This scheme thus yields a single simulated short-rate path $r_{t_0}, r_{t_1}, \ldots, r_{t_{d_Z}}$, where N repetitions are necessary to construct the Monte Carlo estimator in (A.27) for a single derivative price.

Interest Rate Caps/Floors

Since a cap is a sum of caplets each being valued using formula (A.24), its intermediate discounted payoffs occurring at times T_i, $i = 1, \dots, n$, are given under the respective $\mathbb{Q}^{T_{i-1}}$-forward measures for $t = 0$ by

$$h^{\text{Caplet}}(0, T_{i-1}, T_i, K; r_{T_{i-1}}) = P(0, T_{i-1})\big(1 - (1 + K\tau_i)P(T_{i-1}, T_i; r_{T_{i-1}})\big)^+. \tag{A.29}$$

By applying (A.7) with $S = T_{n-1}$, these payoffs can be collected and put as a single discounted payoff under the 'terminal' forward measure $\mathbb{Q}^{T_{n-1}}$. Hence, the discounted payoff of a cap under $\mathbb{Q}^{T_{n-1}}$ can be written as

$$h^{\text{Cap}}(0, \mathcal{T}, K; r_{T_0}, \dots, r_{T_{n-1}}) = P(0, T_{n-1}) \sum_{i=1}^{n} \frac{\big(1 - (1 + K\tau_i)P(T_{i-1}, T_i; r_{T_{i-1}})\big)^+}{P(T_{i-1}, T_{n-1}; r_{T_{i-1}})},$$

requiring the short rate to be simulated under the $\mathbb{Q}^{T_{n-1}}$-forward measure at the n reset dates $T_0, \dots, T_{n-1}$ of the underlying LIBOR $L(T_{i-1}, T_i)$ in order to compute the Monte Carlo cap price (i.e. in scheme (A.28), we have $t_1 = T_0, \dots, t_{d_Z} = T_{n-1}$). In the same way, a single payoff h^{Floor} for a corresponding floor can be derived.

Swaptions

Due to the valuation formula (A.25), the discounted payoff at the maturity date T of a payer swaption is given under the $\mathbb{Q}^T$-forward measure for $t = 0$ by

$$h^{\text{PS}}(0, T, \mathcal{T}, K; r_T) = P(0, T)\big(1 - \sum_{i=1}^{n} c_i P(T, T_i; r_T)\big)^+. \tag{A.30}$$

Hence, to compute the Monte Carlo price of a payer swaption, the short rate only needs to be simulated under the $\mathbb{Q}^T$-forward measure at the swaption maturity T (i.e. in scheme (A.28), we have $t_1 = T$). Clearly, this also applies to a receiver swaption with discounted payoff h^{RS}.

A.2.3 Model Calibration

With the pricing formulas for the main calibration instruments in place, we can now set up the related calibration problems. To this end, we will match each for a set of caps/floors and swaptions with different contractual features, the model prices to their respective market prices in the least-squares sense. Note that, as it is market practice to quote cap/floor and swaption prices in terms of their implied (Black) volatilities, the (Black) market prices are obtained by transforming the implied volatilities retrieved from the market through a version of *Black's formula*[34], after Black [1976]. In particular, the implied volatilities are quoted for

[34] If forward rates are negative, then either the Bachelier model or a displaced log-normal model is used, in which case the Bachelier/normal volatilities or the size of the displacement and the corresponding Black volatilities are quoted, respectively, see, e.g., Kienitz [2017].

several contractual features of the calibration instruments and different kinds of moneyness, to sufficiently capture the term structure of volatilities, see, e.g., Brigo and Mercurio [2006], Section 1.6, or Björk [2009], Chapter 27, for further information.

Interest Rate Caps/Floors

Let us first consider the case of calibrating the Hull-White model through a set of caps/floors with different sets of reset and payment dates and cap/floor rates, where we additionally assume that the implied volatilities are quoted as so-called flat volatilities, i.e. the same implied volatility is used in all caplets/floorlets that constitute a single cap/floor (see, e.g., Björk [2009], Section 27.1).

Letting $\bar{\sigma}_{0,n}$ denote an implied volatility parameter that is retrieved at $t = 0$ from market quotes for a cap with reset and payment dates $\mathcal{T} = \{T_0, \dots, T_n\}$, associated year fractions $\tau_i = T_i - T_{i-1}$, $i = 1, \dots, n$, and cap rate K, it may then be transformed to the market price of the cap by the Black formula

$$\mathrm{Cap}^{\mathrm{Bl}}(0, \mathcal{T}, K; \bar{\sigma}_{0,n}) = \sum_{i=1}^{n} \tau_i P(0, T_i)\Big(L(0, T_{i-1}, T_i)\mathcal{N}(d_1^i) - K\mathcal{N}(d_2^i)\Big),$$

where

$$\begin{aligned} d_1^i &= \frac{\log\big(L(0, T_{i-1}, T_i)/K\big) + \bar{\sigma}_{0,n}^2 T_i/2}{\bar{\sigma}_{0,n}\sqrt{T_i}}, \\ d_2^i &= d_1^i - \bar{\sigma}_{0,n}\sqrt{T_i}, \end{aligned}$$

and $L(0, T_{i-1}, T_i)$ is the simply-compounded forward rate at time $t = 0$ for the time period $[T_{i-1}, T_i]$ as defined by (A.3). Analogously, for a quote $\bar{\sigma}_{0,n}$ of a floor with the same contractual features, the market price may be obtained by

$$\mathrm{Floor}^{\mathrm{Bl}}(0, \mathcal{T}, K; \bar{\sigma}_{0,n}) = \sum_{i=1}^{n} \tau_i P(0, T_i)\Big(K\mathcal{N}(-d_2^i) - L(0, T_{i-1}, T_i)\mathcal{N}(-d_1^i)\Big).$$

Altogether, given a set of implied volatilities $\bar{\sigma}_{0,n}^{(j)}$, $j = 1, \dots, l$, for caps (or floors) with different sets of reset and payment dates $\mathcal{T}^{(j)}$ and cap rates $K^{(j)}$, we may thus formulate the calibration of the Hull-White model through caps as the least-squares problem

$$\min_{x \in \mathcal{X}} \Big\{ f(x) = \big\|\Pi(x) - C^{\mathrm{mkt}}\big\|_2^2 \Big\}, \tag{A.31}$$

where $x = (\kappa, \sigma)^\top$, $\mathcal{X} \subset \mathbb{R}^2_{>0}$ is a suitably chosen compact set, and the corresponding model and market prices are given by $\Pi(x) = (\mathrm{Cap}(0, \mathcal{T}^{(1)}, K^{(1)}; x), \dots, \mathrm{Cap}(0, \mathcal{T}^{(l)}, K^{(l)}; x))^\top$ (whose dependence on the set of parameters x is now explicitly stated) and $C^{\mathrm{mkt}} = (\mathrm{Cap}^{\mathrm{Bl}}(0, \mathcal{T}^{(1)}, K^{(1)}; \bar{\sigma}_{0,n}^{(1)}), \dots, \mathrm{Cap}^{\mathrm{Bl}}(0, \mathcal{T}^{(l)}, K^{(l)}; \bar{\sigma}_{0,n}^{(l)}))^\top$, respectively.

If an approximation to problem (A.31) is sought by either the SAA or the VSAA strategy, then the respective Monte Carlo estimators for $\Pi(x)$ are built according to

$$\widehat{\Pi}_N(x) = \frac{1}{N}\sum_{i=1}^{N} h(x, Z_i), \qquad \text{and} \qquad \widehat{\Pi}_{N_k}(x) = \frac{1}{N_k}\sum_{i=1}^{N_k} h(x, Z_i^k),$$

where $h(x,\cdot) = (h^{\mathrm{Cap}}(0,\mathcal{T}^{(1)},K^{(1)};x,\cdot),\dots,h^{\mathrm{Cap}}(0,\mathcal{T}^{(l)},K^{(l)};x,\cdot))^\top$ is the vector of discounted payoff functions, and the respective samples of i.i.d. random vectors $Z_1,\dots,Z_N$ and $Z_1^k,\dots,Z_{N_k}^k$ have the same distribution as a vector Z of independent and standard normally distributed random variables. In particular, the dimension of Z corresponds to the number of distinct reset dates of all included caps, and each random vector is used to simulate a single short-rate path at the required reset dates under their respective (distinct) $T_{n-1}^{(j)}$-forward measures by the scheme (A.28). However, using pricing formula (A.7) with $S = T^*$, where T^* denotes the longest reset date of all included caps, it is also possible to simulate the short rate path at the required reset dates via (A.28) under the terminal forward measure $\mathbb{Q}^{T^*}$ only. In this case, the j-th calibration instrument of h, which may be formulated as

$$h^{\mathrm{Cap}}(0,\mathcal{T}^{(j)},K^{(j)};x,\cdot) = \sum_{i=1}^{n^{(j)}} h^{\mathrm{Caplet}}(0,T_{i-1}^{(j)},T_i^{(j)},K^{(j)};x,\cdot),$$

each caplet h^{Caplet} given by expression (A.29), must then be adjusted to

$$P(0,T^*)\sum_{i=1}^{n^{(j)}} P^{-1}(T_{i-1}^{(j)},T^*)\,\frac{h^{\mathrm{Caplet}}(0,T_{i-1}^{(j)},T_i^{(j)},K^{(j)};x,\cdot)}{P(0,T_{i-1}^{(j)})}.$$

Swaptions

Similar to the case of caps/floors, the Hull-White model can also be calibrated through a set of swaptions with different option maturities, (underlying swap) tenors and strikes. In particular, given a single implied volatility parameter $\bar{\sigma}_{0,n}$ as quoted in the market at $t = 0$ for a payer swaption with option maturity T, payment dates $\mathcal{T} = \{T_1,\dots,T_n\}$, associated year fractions $\tau_i = T_i - T_{i-1}$ with $T_0 = T$, and strike K, it may be transformed to the Black market price of a payer swaption by

$$\mathrm{PS}^{\mathrm{Bl}}(0,T,\mathcal{T},K;\bar{\sigma}_{0,n}) = \sum_{i=1}^{n} \tau_i P(0,T_i)\Big(S_{0,n}(0)\mathcal{N}(d_1) - K\mathcal{N}(d_2)\Big),$$

where

$$\begin{aligned} d_1 &= \frac{\log\big(S_{0,n}(0)/K\big) + \bar{\sigma}_{0,n}^2 T_0/2}{\bar{\sigma}_{0,n}\sqrt{T_0}}, \\ d_2 &= d_1 - \bar{\sigma}_{0,n}\sqrt{T_0}, \end{aligned}$$

and the forward swap rate $S_{0,n}(0)$ is given with $t = 0$ by

$$S_{0,n}(t) := \frac{P(t,T_0) - P(t,T_n)}{\sum_{i=1}^n \tau_i P(t,T_i)}. \tag{A.32}$$

For a corresponding implied volatility of a receiver swaption, the market price is obtained by

$$\mathrm{RS}^{\mathrm{Bl}}(0,T,\mathcal{T},K;\bar{\sigma}_{0,n}) = \sum_{i=1}^{n} \tau_i P(0,T_i)\Big(K\mathcal{N}(-d_2) - S_{0,n}(0)\mathcal{N}(-d_1)\Big).$$

To sum up, retrieving a set of implied volatilities $\bar{\sigma}_{0,n}^{(j)}$, $j = 1,\ldots,l$, for payer (or receiver) swaptions with different option maturities $T^{(j)}$, payment dates $\mathcal{T}^{(j)}$ and strikes $K^{(j)}$, we can thus state the usual calibration of the Hull-White model through payer swaptions by

$$\min_{x\in\mathcal{X}} \Big\{ f(x) = \big\|\Pi(x) - C^{\mathrm{mkt}}\big\|_2^2 \Big\}, \tag{A.33}$$

where $x = (\kappa,\sigma)^\top$, $\mathcal{X} \subset \mathbb{R}^2_{>0}$ is a suitably chosen compact set, and the vectors of model and market prices are denoted by $\Pi(x) = (\mathrm{PS}(0,T^{(1)},\mathcal{T}^{(1)},K^{(1)};x),\ldots,\mathrm{PS}(0,T^{(l)},\mathcal{T}^{(l)},K^{(l)};x))^\top$ and $C^{\mathrm{mkt}} = (\mathrm{PS}^{\mathrm{Bl}}(0,T^{(1)},\mathcal{T}^{(1)},K^{(1)};\bar{\sigma}_{0,n}^{(1)}),\ldots,\mathrm{PS}^{\mathrm{Bl}}(0,T^{(l)},\mathcal{T}^{(l)},K^{(l)};\bar{\sigma}_{0,n}^{(l)}))^\top$, respectively.

Moreover, approximating (A.33) by the SAA and VSAA strategy amounts to using the Monte Carlo estimators

$$\widehat{\Pi}_N(x) = \frac{1}{N}\sum_{i=1}^{N} h(x,Z_i), \qquad \text{and} \qquad \widehat{\Pi}_{N_k}(x) = \frac{1}{N_k}\sum_{i=1}^{N_k} h(x,Z_i^k),$$

respectively, where $h(x,\cdot) = (h^{\mathrm{PS}}(0,T^{(1)},\mathcal{T}^{(1)},K^{(1)};x,\cdot),\ldots,h^{\mathrm{PS}}(0,T^{(l)},\mathcal{T}^{(l)},K^{(l)};x,\cdot))^\top$ and the respective i.i.d. random vectors are drawn from the same distribution as a vector Z of independent and standard normally distributed random variables. Here, the dimension of Z equals the number of distinct option maturities of all included swaptions, and each random vector is then used to simulate a single short-rate path at the required maturities under their respective (distinct) $T^{(j)}$-forward measures by use of the scheme (A.28). Nevertheless, using formula (A.7) with $S = T^*$, where T^* denotes the longest option maturity of all included swaptions, the short rate may also be simulated at the required maturities via (A.28) under the terminal forward measure $\mathbb{Q}^{T^*}$ only. The j-th calibration instrument of the discounted payoff vector h, given by $h^{\mathrm{PS}}(0,T^{(j)},\mathcal{T}^{(j)},K^{(j)};x,\cdot)$ through formula (A.30), then has to be rewritten as

$$P(0,T^*)P^{-1}(T^{(j)},T^*)\,\frac{h^{\mathrm{PS}}(0,T^{(j)},\mathcal{T}^{(j)},K^{(j)};x,\cdot)}{P(0,T^{(j)})}.$$

A.3 The Nelson-Siegel and Svensson Models

The Nelson-Siegel and Svensson models are concerned with modelling the term structure of interest rates by means of simple parametric functions that rely on few parameters only. Yet, despite their simplicity, fitting them to market data is numerically challenging and various difficulties have been reported. Out of this reason, we present in this section a novel analysis leading to new insights into the inherent instability of both models, to which end we then suggest a new fitting approach based on a penalising objective function.

Let us first briefly introduce the models, where we consider an arbitrary but fixed time instant $t \leq T'$ and define the time to maturity by $\tau = (T - t)$, for $T \in [t, T']$.

Nelson-Siegel Model

Nelson and Siegel [1987] propose to model the instantaneous forward rate curve by a parsimonious three-component exponential approximation that is able to reproduce most of the stylised facts observed in historical records of forward rate curves, such as a monotonic, humped or S-shaped behaviour. The parametric function, which effectively consists of a constant and a Laguerre function, i.e. a polynomial times an exponential decay term, takes on the form

$$F_{\lambda,\beta}(t,T) = \beta_1 + \beta_2 \mathrm{e}^{-\lambda_1 (T-t)} + \beta_3 \lambda_1 (T-t) \mathrm{e}^{-\lambda_1 (T-t)},$$

where $\beta_1, \beta_2, \beta_3 \in \mathbb{R}$ denote the linear coefficients and $\lambda_1 > 0$ the shape parameter. With (A.2) and (A.4), one then obtains for the corresponding zero rate curve, now explicitly written as a function of the maturity τ, the expression

$$y_{\lambda,\beta}(\tau) = \beta_1 + \beta_2 \left(\frac{1 - \mathrm{e}^{-\lambda_1 \tau}}{\lambda_1 \tau} \right) + \beta_3 \left(\frac{1 - \mathrm{e}^{-\lambda_1 \tau}}{\lambda_1 \tau} - \mathrm{e}^{-\lambda_1 \tau} \right). \tag{A.34}$$

Although Nelson and Siegel's model is quite simple, it can assume a variety of shapes depending on the four parameters which have a clear interpretation: β_1 describes the long-term rate of $y_{\lambda,\beta}$, the sum $\beta_1 + \beta_2$ accounts for its short-term rate, and β_3 and λ_1 determine the height and position of the hump of $y_{\lambda,\beta}$, respectively. For more details on the model and its parameters including the subsequent Svensson model, we refer to De Pooter [2007].

Svensson Model

To allow for an even greater flexibility in the curves, Svensson [1995] proposes to extend Nelson and Siegel's model by adding a further term, giving the model a better fit to long maturities, see, e.g., Diebold and Rudebusch [2013]. Accordingly, the instantaneous forward rate curve is specified as

$$F_{\lambda,\beta}(t,T) = \beta_1 + \beta_2 \mathrm{e}^{-\lambda_1 (T-t)} + \beta_3 \lambda_1 (T-t) \mathrm{e}^{-\lambda_1 (T-t)} + \beta_4 \lambda_2 (T-t) \mathrm{e}^{-\lambda_2 (T-t)},$$

where $\beta_4 \in \mathbb{R}$ and $\lambda_2 > 0$ denote the additionally introduced parameters, and which can be integrated to obtain the related zero rate curve as

$$y_{\lambda,\beta}(\tau) = \beta_1 + \beta_2 \left(\frac{1 - \mathrm{e}^{-\lambda_1 \tau}}{\lambda_1 \tau} \right) + \beta_3 \left(\frac{1 - \mathrm{e}^{-\lambda_1 \tau}}{\lambda_1 \tau} - \mathrm{e}^{-\lambda_1 \tau} \right) + \beta_4 \left(\frac{1 - \mathrm{e}^{-\lambda_2 \tau}}{\lambda_2 \tau} - \mathrm{e}^{-\lambda_2 \tau} \right). \tag{A.35}$$

Note that Nelson and Siegel and Svensson, as well as several other authors, use a different specification of the models in which the shape parameters are defined as the reciprocals of λ_1 and λ_2, i.e. as $1/\lambda_1$ and $1/\lambda_2$, respectively. From an optimisation point of view, however, we opt for the presented modification of the original parametrisation of the Nelson-Siegel and the Svensson models. Moreover, unlike other authors, we do not impose any restrictions on the linear parameters β_1, β_2, β_3 and β_4 at this point. This is justified by the fact that interest rates may well become negative, as recent developments in financial markets have shown, see, e.g., Atkins [2014].

A.3.1 Fitting of Zero Rate Curves

In what follows, we first outline the standard approach for fitting the Nelson-Siegel and Svensson models to available market zero rates and briefly point out the difficulties arising from this procedure. We then show how this approach can be improved by exploiting the inherent structure of both models.

A.3.1.1 Traditional Approach

Given the descriptions of the Nelson-Siegel and the Svensson models in (A.34) and (A.35), respectively, the zero rate curves can be generalised to

$$y_{\lambda,\beta}(\tau) = \sum_{j=1}^{d_\beta} \beta_j \psi_j(\lambda;\tau), \tag{A.36}$$

where the continuously differentiable basis functions ψ_j have the form

$$\psi_1(\lambda;\tau) = 1, \qquad \psi_2(\lambda;\tau) = \frac{1-\mathrm{e}^{-\lambda_1\tau}}{\lambda_1\tau},$$
$$\psi_3(\lambda;\tau) = \frac{1-\mathrm{e}^{-\lambda_1\tau}}{\lambda_1\tau} - \mathrm{e}^{-\lambda_1\tau}, \qquad \psi_4(\lambda;\tau) = \frac{1-\mathrm{e}^{-\lambda_2\tau}}{\lambda_2\tau} - \mathrm{e}^{-\lambda_2\tau}.$$

Using (A.36) and letting $\tau_1, \ldots, \tau_l \in [0, T']$ denote some set of predefined maturities at which the market zero rates $y_1^{\mathrm{mkt}}, \ldots, y_l^{\mathrm{mkt}} \in \mathbb{R}$ are available, the fitting of the Nelson-Siegel and the Svensson models to market data can then be set up in the least-squares sense as

$$\min_{\lambda\in\mathcal{X}_\lambda,\, \beta\in\mathbb{R}^{d_\beta}} \Big\{ f(\lambda,\beta) := \big\| \Psi(\lambda)\beta - y^{\mathrm{mkt}} \big\|_2^2 \Big\}, \tag{A.37}$$

where $\beta \in \mathbb{R}^{d_\beta}$ and $\lambda \in \mathcal{X}_\lambda$ are unknown parameters, $\mathcal{X}_\lambda \subset \mathbb{R}_{>0}^{d_\lambda}$ is without loss of generality a compact set (see Section A.3.1.2 for the relevant reasoning), $\Psi(\lambda) \in \mathbb{R}^{l\times d_\beta}$ denotes the matrix of basis functions with entries $\{\Psi(\lambda)\}_{i,j} = \psi_j(\lambda;\tau_i)$, $i = 1, \ldots, l$, $j = 1, \ldots, d_\beta$, and $y^{\mathrm{mkt}} = (y_1^{\mathrm{mkt}}, \ldots, y_l^{\mathrm{mkt}})^\top \in \mathbb{R}^l$ presents the vector of available market rates. We also assume that $l > d_\lambda + d_\beta$ always holds, ensuring that (A.37) defines an overdetermined problem.

Brief Review of Existing Approaches

Although low-dimensional, problem (A.37) is numerically challenging to solve for several reasons. Firstly, due to the form of the Nelson-Siegel and the Svensson models, it is a highly *nonlinear nonconvex least-squares problem* and as such intrinsically difficult. The objective function f is typically rather insensitive in regions of low function values while it reacts very strongly to parameter combinations in other regions. Multiple local minima may further exist, which essentially requires global optimisation techniques to ensure convergence to an optimal solution. Secondly, as a consequence of the exponential expressions in both models, they suffer from severe multicollinearity in certain regions of the parameter space $\mathcal{X}_\lambda$, such that the fitting of these models to market data leads to an ill-conditioned problem.

To facilitate fitting and avoid some of the numerical issues, a straightforward approach adopted by several authors to solve (A.37) is to fix the nonlinear parameter λ at prespecified values and to use ordinary linear least-squares methods to obtain the remaining optimal linear parameters. This involves techniques where λ is fixed at a grid of different values in a reasonable interval, as considered by Nelson and Siegel [1987] and Annaert et al. [2013], for instance, and techniques where the parameter is set according to its interpretation to some value derived from the considered data set, see, e.g., Fabozzi et al. [2005], Diebold and Li [2006], De Pooter [2007] and De Rezende [2011]. In either case, however, the approach considerably limits the models and reduces some of the flexibility in reproducing different types of zero rate curves.

Even though in theory global optimisation methods are required to solve the fitting problem appropriately, a further approach frequently applied is to resort to standard local optimisation techniques, see Cairns and Pritchard [2001], De Pooter [2007], Gauthier and Simonato [2012] or Virmani [2012], for example. To mitigate the danger of getting stuck in a local optimum, these search techniques may then be 'globalised' by starting from a number of different initial values, e.g. as selected by some interpretive strategy or randomly. However, no guarantee of convergence to a global minimum can be given. As opposed to traditional optimisation techniques that are based on gradients, Gilli et al. [2010] argue in favour of using a heuristic global method, differential evolution, to solve the fitting problem and reach to some extent satisfying results if the parameter space is restricted substantially.

Eventually, to prevent a potential multicollinearity in which the linear parameter estimate β becomes extremely sensitive to the choice of λ while still achieving a good fit, the parameter space is commonly restricted in an approximate manner, as analysed, e.g., by De Pooter [2007], Gilli et al. [2010] and Annaert et al. [2013]. In most cases, though, this results in a parameter space that is too restrictive and thus may also exclude regions containing a possible global minimum with only moderately correlated factor loadings.

A.3.1.2 New Approach

Despite the fact that the main difficulties in fitting both models to market data by means of zero rates have been recognised in various source, no fully satisfying analysis has been presented in the literature so far. To this end, we propose in the following a novel analysis supplementing the existing approaches. The approach is based on the observation that problem (A.37) can be reformulated as a separable nonlinear least-squares problem, which renders the global optimisation problem computationally tractable as its dimension is reduced significantly and at the same time allows to avoid collinearity issues substantially. Even though the special structure of the objective function was already recognised by Angelini and Herzel [2002] and Gauthier and Simonato [2012], no theoretical justification in the sense of Theorem A.7 below was provided, not to mention the subsequent implications on the treatment of the ill-conditioning of the inner problem by a stability analysis, see Subsection A.3.2.

Main Idea

Since the zero rate curves $y_{\lambda,\beta}$ in both the Nelson-Siegel and the Svensson model are expressed as a linear combination of nonlinear basis functions in which the parameters λ and β form two disjoint sets, cf. formula (A.36), the original minimisation problem (A.37) evidently presents a *separable nonlinear least-squares problem*, see, e.g., Björck [1996], Section 9.4. Hence, if we know the nonlinear parameter λ, the corresponding optimal linear parameter $\beta^* = \beta^*(\lambda)$ will always exist for any $\lambda \in \mathcal{X}_\lambda$ and can be obtained uniquely by solving the standard linear least-squares problem

$$\min_{\beta \in \mathbb{R}^{d_\beta}} f(\lambda, \beta), \tag{A.38}$$

for fixed $\lambda \in \mathcal{X}_\lambda$. Its solution is given by

$$\beta^*(\lambda) = \Psi(\lambda)^\dagger y^{\mathrm{mkt}},$$

where $\Psi(\lambda)^\dagger$ denotes the Moore-Penrose pseudoinverse of $\Psi(\lambda)$, generalising the notion of the inverse of a square and invertible matrix to rectangular matrices, see, e.g., Björck [1996], Sections 1.1.4 and 1.2.5. Accordingly, if the columns of $\Psi(\lambda)$ are linearly independent, i.e. $\operatorname{rank}(\Psi(\lambda)) = d_\beta$, the unique least squares solution satisfies the normal equations $\Psi(\lambda)^\top \Psi(\lambda) \beta(\lambda) = \Psi(\lambda)^\top y^{\mathrm{mkt}}$ and is thus given by $\beta^*(\lambda) = \big(\Psi(\lambda)^\top \Psi(\lambda)\big)^{-1} \Psi(\lambda)^\top y^{\mathrm{mkt}}$. If $\operatorname{rank}(\Psi(\lambda)) < d_\beta$, there are many least squares solutions $\beta^*(\lambda)$ that have the same residual $\Psi(\lambda)\beta^*(\lambda) - y^{\mathrm{mkt}}$. In this case, the Moore-Penrose pseudoinverse assigns the solution with minimum length $\|\beta^*(\lambda)\|_2$, which is uniquely defined.

On substituting the optimal solution into the objective function f, the original problem (A.37) can be decomposed into an outer and inner optimisation problem

$$\min_{\lambda \in \mathcal{X}_\lambda,\, \beta \in \mathbb{R}^{d_\beta}} f(\lambda, \beta) = \min_{\lambda \in \mathcal{X}_\lambda} \underbrace{\min_{\beta \in \mathbb{R}^{d_\beta}} f(\lambda, \beta)}_{=: f^{\mathrm{sep}}(\lambda)} = \min_{\lambda \in \mathcal{X}_\lambda} f^{\mathrm{sep}}(\lambda), \tag{A.39}$$

where the objective function f^{sep} takes the semi-analytical form

$$f^{\mathrm{sep}}(\lambda) = f\big(\lambda, \beta^*(\lambda)\big) = \big\| \Psi(\lambda)\Psi(\lambda)^\dagger y^{\mathrm{mkt}} - y^{\mathrm{mkt}} \big\|_2^2, \tag{A.40}$$

thus eliminating the linear parameter β. The outer problem (A.39) is a nonconvex optimisation problem in the nonlinear parameter $\lambda \in \mathcal{X}_\lambda$. For each function evaluation of the objective function f^{sep} in (A.40), the inner problem (A.38) needs to be solved, being an unconstrained linear least-squares problem in β.

Theoretical Justification

The rationale for employing the proposed technique is given by the following theorem, due to Golub and Pereyra [1973], Theorem 2.1. It shows the relationship between critical points of the original objective f and the new objective f^{sep}, as well as between their global minimisers.

Theorem A.7. *Assume that in the open set $\widetilde{\mathcal{X}}_\lambda \subset \mathbb{R}^{d_\lambda}_{>0}$, the matrix $\Psi(\lambda)$ has constant rank $0 < r_\Psi \leq d_\beta$.*

(a) If λ^* is a critical point, resp. a global minimiser, of $f^{sep}(\lambda)$ in $\widetilde{\mathcal{X}}_\lambda$ and $\beta^* = \Phi(\lambda^*)^\dagger y^{mkt}$, then (λ^*, β^*) is a critical point, resp. a global minimiser, of $f(\lambda, \beta)$ for $\lambda \in \widetilde{\mathcal{X}}_\lambda$ and $f(\lambda^*, \beta^*) = f^{sep}(\lambda^*)$.

(b) If (λ^*, β^*) is a global minimiser of $f(\lambda, \beta)$ for $\lambda \in \widetilde{\mathcal{X}}_\lambda$, then λ^* is a global minimiser of $f^{sep}(\lambda)$ in $\widetilde{\mathcal{X}}_\lambda$ and $f^{sep}(\lambda^*) = f(\lambda^*, \beta^*)$. Furthermore, if there is a unique β^* among the minimising pairs of $f(\lambda, \beta)$, then β^* must satisfy $\beta^* = \Phi(\lambda^*)^\dagger y^{mkt}$.

The equivalence between the global minimisers of both objective functions relies on the assumption that the rank of the matrix $\Psi(\lambda)$ is locally constant on an open set $\widetilde{\mathcal{X}}_\lambda$, that is for every $\lambda \in \widetilde{\mathcal{X}}_\lambda$, the rank of the matrix $\Phi(\lambda)$ is constant. This guarantees that the Moore-Penrose pseudoinverse $\Psi(\lambda)^\dagger$ is continuous and differentiable on $\widetilde{\mathcal{X}}_\lambda$, see the subsequent theorem, cf. Golub and Pereyra [1973], Theorem 4.3.

Theorem A.8. *Let $\widetilde{\mathcal{X}}_\lambda \subset \mathbb{R}^{d_\lambda}_{>0}$ be an open set. For $\lambda \in \widetilde{\mathcal{X}}_\lambda$, let $\Psi(\lambda)$ be Fréchet differentiable of local constant rank $0 < r_\Psi \leq d_\beta$ in $\widetilde{\mathcal{X}}_\lambda$. Then, for any $\lambda \in \widetilde{\mathcal{X}}_\lambda$, the Fréchet derivative of $\Psi(\lambda)$, denoted by $\mathrm{D}\Psi(\lambda)$, satisfies*

$$\begin{aligned}\mathrm{D}\Psi(\lambda)^\dagger = &- \Psi(\lambda)^\dagger\, \mathrm{D}\Psi(\lambda)\Psi(\lambda)^\dagger + \left(\Psi(\lambda)^\top\Psi(\lambda)\right)^\dagger \mathrm{D}\Psi(\lambda)^\top\left(I - \Psi(\lambda)\Psi(\lambda)^\dagger\right) \\ &+ \left(I - \Psi(\lambda)^\dagger\Psi(\lambda)\right)\mathrm{D}\Psi(\lambda)^\top\left(\Psi(\lambda)\Psi(\lambda)^\top\right)^\dagger.\end{aligned} \tag{A.41}$$

From the differentiability of the Moore-Penrose pseudoinverse on $\widetilde{\mathcal{X}}_\lambda$, it immediately follows with (A.40) that the objective function f^{sep} is also differentiable on $\widetilde{\mathcal{X}}_\lambda$, so that formulas for its gradient can be established.

Corollary A.9. *Let $\widetilde{\mathcal{X}}_\lambda \subset \mathbb{R}^{d_\lambda}_{>0}$ be an open set. For $\lambda \in \widetilde{\mathcal{X}}_\lambda$, let $\Psi(\lambda)$ be the Fréchet differentiable matrix of basis functions of local constant rank $0 < r_\Psi \leq d_\beta$ in $\widetilde{\mathcal{X}}_\lambda$. Then, for any $\lambda \in \widetilde{\mathcal{X}}_\lambda$,*

$$\nabla f^{sep}(\lambda) = -2(y^{mkt})^\top\left(I - \Psi(\lambda)\Psi(\lambda)^\dagger\right)\left[\beta_3^*(\lambda)(\tau \circ \mathrm{e}^{-\lambda_1\tau}),\, \beta_4^*(\lambda)(\tau \circ \mathrm{e}^{-\lambda_2\tau})\right],$$

where '$\circ$' denotes the Hadamard product of componentwise vector multiplication.

Proof. Due to Golub and Pereyra [1973], p. 419, the gradient of f^{sep} can be written as

$$\nabla f^{\text{sep}}(\lambda) = -2(y^{\text{mkt}})^{\top}\big(I - \Psi(\lambda)\Psi(\lambda)^{\dagger}\big)\mathrm{D}\Psi(\lambda)\Psi(\lambda)^{\dagger}y^{\text{mkt}}. \tag{A.42}$$

Since $\mathrm{D}\Psi(\lambda) \in \mathbb{R}^{d_\lambda \times (l \times d_\beta)}$ is a tensor, its first and second slab of partial derivatives with respect to λ_1 and λ_2 have the matrix forms

$$\big[\mathrm{D}\Psi(\lambda)\big]_1 = \big[\mathbf{0}, \psi_2'(\lambda_1;\tau), \psi_2'(\lambda_1;\tau) + \tau \circ \mathrm{e}^{-\lambda_1\tau}, \mathbf{0}\big],$$

and

$$\big[\mathrm{D}\Psi(\lambda)\big]_2 = \big[\mathbf{0}, \mathbf{0}, \mathbf{0}, \psi_2'(\lambda_2;\tau) + \tau \circ \mathrm{e}^{-\lambda_2\tau}\big],$$

respectively, where $\psi_2'(x;\tau) = \mathrm{e}^{-x\tau}/x - (\mathbf{1} - \mathrm{e}^{-x\tau})/(x^2\tau)$ denotes the first derivative of the second basis function $\psi_2(\lambda;\tau)$ with respect to λ_1. Using $\beta^*(\lambda) = \Psi(\lambda)^{\dagger}y^{\text{mkt}}$, it follows that we can rewrite (A.42) as

$$\begin{aligned}\nabla f^{\text{sep}}(\lambda) = &-2(y^{\text{mkt}})^{\top}\big(I - \Psi(\lambda)\Psi(\lambda)^{\dagger}\big)\\ &\times \big[\big(\beta_2^*(\lambda) + \beta_3^*(\lambda)\big)\psi_2'(\lambda_1) + \beta_3^*(\lambda)(\tau \circ e^{-\lambda_1\tau}),\ \beta_4^*(\lambda)\psi_2'(\lambda_2) + \beta_4^*(\lambda)(\tau \circ \mathrm{e}^{-\lambda_2\tau})\big].\end{aligned}$$

Now, $\psi_2'(\lambda_1;\tau) = -\psi_3(\lambda;\tau)/\lambda_1$ and $\psi_2'(\lambda_2;\tau) = -\psi_4(\lambda;\tau)/\lambda_2$, and from the normal equations

$$\Psi(\lambda)^{\top}\Psi(\lambda)\beta^*(\lambda) = \Psi(\lambda)^{\top}y^{\text{mkt}},$$

it follows that any column of $\Psi(\lambda)$ is orthogonal to $(y^{\text{mkt}})^{\top}(I - \Psi(\lambda)\Psi(\lambda)^{\dagger})$. Hence, the gradient of f^{sep} can be simplified to

$$\nabla f^{\text{sep}}(\lambda) = -2(y^{\text{mkt}})^{\top}\big(I - \Psi(\lambda)\Psi(\lambda)^{\dagger}\big)\big[\beta_3^*(\lambda)(\tau \circ \mathrm{e}^{-\lambda_1\tau}),\ \beta_4^*(\lambda)(\tau \circ \mathrm{e}^{-\lambda_2\tau})\big].$$

□

Since the matrix norms of $\Psi(\lambda)$ and $\Psi(\lambda)^{\dagger}$ and the Fréchet derivatives $\mathrm{D}\Psi(\lambda)$ and $\mathrm{D}\Psi(\lambda)^{\top}$ are bounded on bounded domains for locally constant rank, the Fréchet derivative in (A.41) is bounded on $\widetilde{\mathcal{X}}_\lambda$. Hence, the Moore-Penrose pseudoinverse is locally Lipschitz continuous on $\widetilde{\mathcal{X}}_\lambda$, as well as the objective function f^{sep} as a composition of locally Lipschitz continuous functions. Thus, f^{sep} is globally Lipschitz continuous on any compact subset $\mathcal{X}_\lambda \subset \widetilde{\mathcal{X}}_\lambda$.

When fitting the models to data, an open feasible set $\widetilde{\mathcal{X}}_\lambda \subset \mathbb{R}^{d_\lambda}_{>0}$ and a compact subset $\mathcal{X}_\lambda \subset \widetilde{\mathcal{X}}_\lambda$ can be chosen in such a way that $\Psi(\lambda)$ always has full rank, and thus the assumptions of the previous theorems are satisfied. This is due to the subsequent theorem showing that global solutions with rank deficient models can be ruled out.

Theorem A.10. *Let λ^* be the global minimiser of $f^{sep}(\lambda)$ on $\mathbb{R}^{d_\lambda}_{>0}$ and $f^{sep}(\lambda^*) > 0$. Then, the matrix $\Psi(\lambda^*)$ has full rank.*

Using the fact that the set of rank deficient points of the Svensson model in $\mathbb{R}^2_{>0}$ is

$$\mathbb{R}_{>0} \times \{0\} \cup \{0\} \times \mathbb{R}_{>0} \cup \{\lambda \in \mathbb{R}^2_{>0} \,|\, \lambda_1 = \lambda_2\},$$

we can choose $\widetilde{\mathcal{X}}_\lambda$ as large as possible, i.e.

$$\widetilde{\mathcal{X}}_\lambda := \{\lambda \in \mathbb{R}^2_{>0} \,|\, 0 < \lambda_1 < \lambda_2\} \cup \{\lambda \in \mathbb{R}^2_{>0} \,|\, \lambda_1 > \lambda_2 > 0\}$$

in the Svensson model, and analogously in the Nelson-Siegel model. This allows to split the problem into two independent global problems which are combined afterwards to yield the overall global solution.

Proof of Theorem A.10. Without loss of generality, we consider the Svensson model and show that for arbitrary $\lambda_2 > 0$, there exists a $\lambda_1 > 0$ with $f^{\text{sep}}(0, \lambda_2) > f^{\text{sep}}(\lambda_1, \lambda_2)$. All remaining cases are completely analogous; similar arguments also hold for the Nelson-Siegel model where the parameters λ_2 and β_4 are omitted.

We first show that for any fixed λ_2, there exists a $\lambda_1 > 0$, $\lambda_1 \neq \lambda_2$ yielding a function value of f^{sep} that is at least as small as for $\lambda_1 = 0$. By definition of f^{sep}, we have

$$f^{\text{sep}}(0, \lambda_2) = \min_{\beta_1, \beta_2, \beta_3, \beta_4 \in \mathbb{R}} f(0, \lambda_2, \beta_1, \beta_2, \beta_3, \beta_4).$$

Since $\psi_1(0, \lambda_2; \tau) = \psi_2(0, \lambda_2; \tau) = \mathbf{1}$ and $\psi_3(0, \lambda_2; \tau) = \mathbf{0}$, we can rewrite the latter as

$$\begin{aligned} \min_{\beta_1, \beta_4 \in \mathbb{R}} f(0, \lambda_2, \beta_1, 0, 0, \beta_4) &= \min_{\beta_1, \beta_4 \in \mathbb{R}} f(\lambda_1, \lambda_2, \beta_1, 0, 0, \beta_4) \\ &\geq \min_{\beta_1, \beta_2, \beta_3, \beta_4 \in \mathbb{R}} f(\lambda_1, \lambda_2, \beta_1, \beta_2, \beta_3, \beta_4) \qquad \text{(A.43)} \\ &= f^{\text{sep}}(\lambda_1, \lambda_2), \end{aligned}$$

where the inequality immediately follows from allowing the minimisation to be additionally carried out over the parameters β_2 and β_3.

To show that (A.43) holds even strictly for suitably chosen λ_1, let $\beta^* = (\beta_1^*, 0, 0, \beta_4^*)^\top$ denote the minimiser of $f^{\text{sep}}(0, \lambda_2)$. It then follows that

$$\begin{aligned} \min_{\beta_1, \beta_4 \in \mathbb{R}} f(\lambda_1, \lambda_2, \beta_1, 0, 0, \beta_4) &= f(\lambda_1, \lambda_2, \beta_1^*, 0, 0, \beta_4^*) \\ &\geq \min_{\beta_2, \beta_3 \in \mathbb{R}} f(\lambda_1, \lambda_2, \beta_1^*, \beta_2, \beta_3, \beta_4^*). \qquad \text{(A.44)} \end{aligned}$$

Moreover, the normal equations imply that β^* satisfies $\Psi(0, \lambda_2)^\top(\Psi(0, \lambda_2)\beta^* - y^{\text{mkt}}) = 0$, meaning that the first and fourth column of $\Psi(0, \lambda_2)$ are orthogonal to the fitted residuals of $f^{\text{sep}}(0, \lambda_2)$ and, equivalently, to $(\Psi(\lambda_1, \lambda_2)\beta^* - y^{\text{mkt}})$ as these columns do not depend on λ_1.

Now note that any solution of $\min_{\beta_2, \beta_3 \in \mathbb{R}} f(\lambda_1, \lambda_2, \beta_1^*, \beta_2, \beta_3, \beta_4^*)$ is unique by the assumption that $\lambda_1 \neq \lambda_2$ and $\lambda_1 \neq 0$. Thus, β^* cannot be a minimiser of $f(\lambda_1, \lambda_2, \beta_1^*, \beta_2, \beta_3, \beta_4^*)$ unless the columns $\psi_2(\lambda_1, \lambda_2; \tau)$ and $\psi_3(\lambda_1, \lambda_2; \tau)$ are orthogonal to the fitted residuals of $f^{\text{sep}}(0, \lambda_2)$, in which case the inequality in (A.44) does not hold strictly. However, in these cases, since $\psi_2(\lambda_1, \lambda_2; \tau)$ and $\psi_3(\lambda_1, \lambda_2; \tau)$ are independent of λ_2, $\lambda_1 \neq \lambda_2$ can always be chosen such that $\psi_2(\lambda_1, \lambda_2; \tau)$ and $\psi_3(\lambda_1, \lambda_2; \tau)$ are not orthogonal to the fitted residuals of $f^{\text{sep}}(0, \lambda_2)$.

In particular, it is possible to find l different λ_1's for which the vectors $\psi_2(\lambda_1, \lambda_2; \tau)$ and $\psi_3(\lambda_1, \lambda_2; \tau)$ form a basis of $\mathbb{R}^l$, respectively, so that the orthogonality condition becomes unsatisfiable for both functions entirely and nonzero residuals (as $f^{\text{sep}}(\lambda^*) > 0$). $\square$

Remark A.11. *For the yearly spaced maturity vector $\tau = (1, \ldots, 15)^\top$ as used in Section 4.3.3, it can be shown that the λ_1's*

$$\{0.01, 0.02, 0.06, 0.12, 0.20, 0.32, 0.47, 0.65, 0.89, 1.18, 1.56, 2.06, 2.77, 3.99, 7.62\},$$

and

$$\{0.01, 0.04, 0.09, 0.17, 0.28, 0.42, 0.58, 0.79, 1.05, 1.36, 1.77, 2.31, 3.10, 4.43, 8.47\},$$

lead to vector sets $\psi_2(\lambda_1, \lambda_2; \tau)$ and $\psi_3(\lambda_1, \lambda_2; \tau)$ that provide a basis of $\mathbb{R}^{15}$, respectively.

A.3.2 Stability Analysis

One of the main issues arising in the optimisation of the reduced problem $\min_{\lambda \in \mathcal{X}_\lambda} f^{\text{sep}}(\lambda)$ is the stability of optimal solutions. To assess the quality of optimal solutions, note that the evaluation of f^{sep} only depends on the solution of the inner problem $\beta^*(\lambda) = \Psi(\lambda)^\dagger y^{\text{mkt}}$. Hence, the stability of optimal solutions of the separable least-squares problem can be analysed by applying perturbation theory to linear least-squares problems, e.g., Björck [1996], Section 1.4. Accordingly, there are two different situations in which optimal solutions $\beta^*(\lambda)$ may become too sensitive with respect to perturbations of either the data vector y^{mkt} or the matrix $\Psi(\lambda)$. The first one concerns the projection of y^{mkt} onto the span of $\Psi(\lambda)$ and turns out to be of relvance if both components are nearly orthogonal to each other. In such case, the projected y^{mkt} is much smaller than y^{mkt}, so that minor changes in y^{mkt} may affect the linear solution $\beta^*(\lambda)$ greatly. However, since the Nelson-Siegel and Svensson models are able to fit a variety of different zero rate curves with high accuracy, this situation in fact never occurs and the sensitivity to perturbations in y^{mkt} may be neglected[35]. The second issue pertains to the conditioning of the matrix $\Psi(\lambda)$ and is thus influenced solely by the factor loading structure that is imposed by the models. In this case, optimal solutions of the linear least-squares system respond strongly to perturbations in $\Psi(\lambda)$ if the matrix is ill-conditioned, i.e. if some of the columns of $\Psi(\lambda)$ are almost linearly dependent. As this is a more subtle issue, we provide a thorough analysis of the potential ill-conditioning of $\Psi(\lambda)$ and how it can be dealt with in the remaining part of this section. In particular, we use the condition number of $\Psi(\lambda)$ to measure the sensitivity of an optimal solution $(\lambda^*, \beta^*(\lambda^*))^\top$, which also corresponds to the condition of the problem of evaluating $f^{\text{sep}}(\lambda)$. In this way, we are able to quantify – and manage – the ill-conditioning with our enhanced approach, in contrast to previous approaches.

[35] If models are used where perturbations to y^{mkt} turn out to be relevant, the following analysis can be extended by adjusting the condition number to include y^{mkt}, see, e.g., Björck [1996], Subsection 1.4.3.

A.3.2.1 The Inherent Ill-conditioning of $\Psi(\lambda)$

To be able to quantify the degree of (ill-)conditioning of the rectangular matrix $\Psi(\lambda) \in \mathbb{R}^{l \times d_\beta}$ for a specified maturity vector τ,[36] we consider the following definition, which is based on the singular value decomposition of $\Psi(\lambda)$ (e.g., Björck [1996], Theorem 1.2.1) and generalises the condition number of a square nonsingular matrix, cf. Björck [1996], Definition 1.4.2.

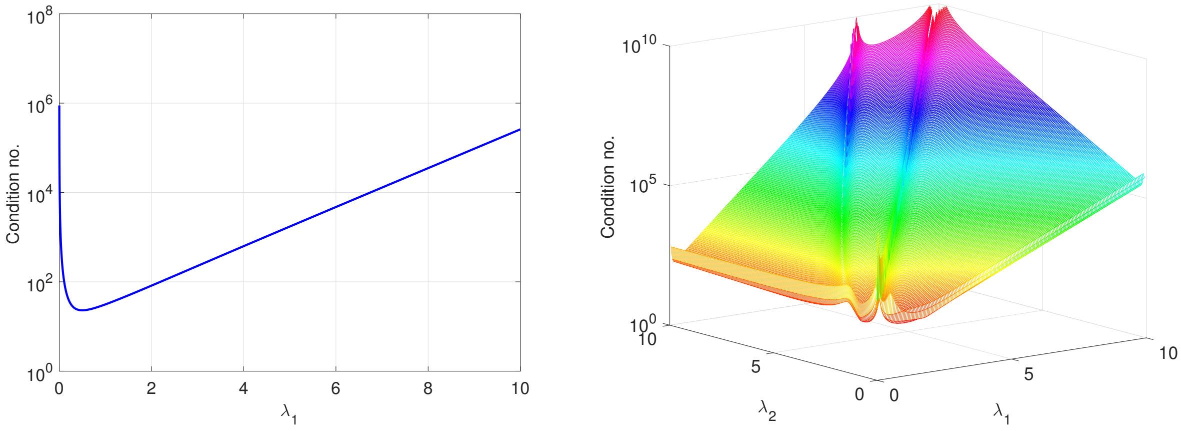
Definition A.12. *The condition number of* $\Psi(\lambda) \in \mathbb{R}^{l \times d_\beta}$ *is given by*

$$\kappa_\Psi(\lambda) := \|\Psi(\lambda)\|_2 \|\Psi(\lambda)^\dagger\|_2 = \frac{\sigma_1(\lambda)}{\sigma_{r_\Psi}(\lambda)},$$

where $0 < r_\Psi = \operatorname{rank}(\Psi(\lambda)) \leq d_\beta$*,* $\sigma_1(\lambda) \geq \ldots \geq \sigma_{r_\Psi}(\lambda) > 0$ *are the nonzero singular values of* $\Psi(\lambda)$*, and* $\|\cdot\|_2$ *denotes the matrix* 2*-norm.*

The condition number describes how solutions of linear least-squares problems are affected by small perturbations. If it is 'large', i.e. solutions are affected greatly, the problem is said to be ill-conditioned, see, e.g., Nocedal and Wright [2006], Chapter A.1. A more precise interpretation of ill-conditioning is subject to the problem at hand and depends on the application. For our setup, we will give a suitable idea of a large condition number in Subsection A.3.2.2.

The effect of having obtained an optimal nonlinear solution λ^* with ill-conditioned matrix $\Psi(\lambda^*)$ may become especially apparent in that some of the values of the corresponding linear parameter $\beta^*(\lambda^*)$ turn out to be very large (and offsetting), with values being proportional to the degree of ill-conditioning. This, though, is in contradiction to the intuitive economic interpretation that all model parameters have.



(a) Nelson-Siegel model

(b) Svensson model

Figure A.1: The condition number of $\Psi(\lambda)$ as a function of λ for the Nelson-Siegel and the Svensson models with maturity vector $\tau = (1, 2, \ldots, 15)^\top$.

[36] Note that for the data set used in the practical sections, we have $l = 15$ and $\tau = (1, 2, \ldots, 15)^\top$.

Figure A.1 depicts the condition number of the matrix $\Psi(\lambda)$ as a function of the parameter λ for the Nelson-Siegel model and Svensson's extension, respectively. From the subfigures, we can observe that the main difficulties in the fitting of both models arise when the shape parameter λ is either very small or becomes increasingly large, or, in the case of the Svensson model, when $\lambda_1 \approx \lambda_2$. These critical regions can be reconstructed intuitively by taking the limits of the factor loadings in each of the columns of the functional matrix

$$\Psi(\lambda) = \big[\psi_1(\lambda;\tau), \ldots, \psi_{d_\beta}(\lambda;\tau)\big],$$

resulting in:

$$\begin{aligned} &\lim_{\lambda_1 \to 0} \psi_2(\lambda;\tau) = 1, && \lim_{\lambda_1 \to \infty} \psi_2(\lambda;\tau) = 0, \\ &\lim_{\lambda_i \to 0} \psi_3(\lambda;\tau) = 0, && \lim_{\lambda_i \to \infty} \psi_3(\lambda;\tau) = 0, \qquad i = 1, 2, \end{aligned}$$

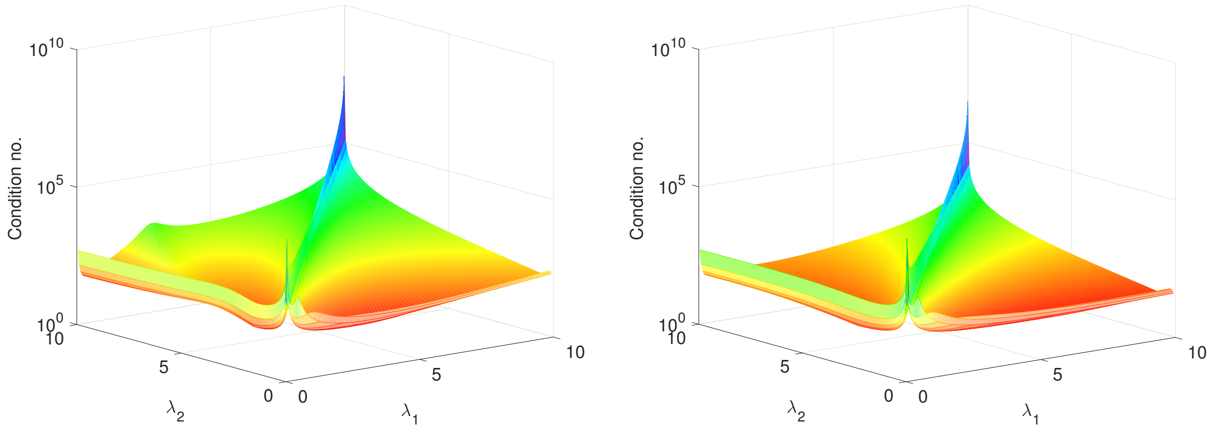
for $\tau > 0$, and by observing that $\psi_3(\lambda;\tau)$ and $\psi_4(\lambda;\tau)$ are about the same for $\lambda_1 \approx \lambda_2$. The severity of the ill-conditioning in the latter case is illustrated by the elevated diagonal in the surface plot of the condition number, see Subfigure A.1(b). Similar large condition numbers can be observed on the left hand side of the diagonal in form of a slightly bent curve. This curve arises from the fact that for the given maturity vector τ, the third and the fourth column of the matrix $\Psi(\lambda)$ become linearly dependent for a particular set of parameter combinations of λ_1 and λ_2, when most of the components of the exponential vectors $\mathrm{e}^{-\lambda_1\tau}$ and $\mathrm{e}^{-\lambda_2\tau}$ attain very small values.

Unlike the issues above, the latter linear dependence between the last two columns of the matrix $\Psi(\lambda)$ can be mitigated by additionally considering market rates of shorter maturities, if available. As an example, the impact of including short maturities into the maturity vector τ on the condition number of $\Psi(\lambda)$ is depicted in Figure A.2 for two different modifications. Hence, the inclusion may considerably improve the degree of ill-conditioning in the region to the left of the diagonal and hence enlarge the parameter space for which a solution may be acceptably stable. In contrast, incorporating maturities larger than 15 years into τ does not lead to a significant improvement of the condition of $\Psi(\lambda)$, as various tests have shown.

Let us point out that if the global optimal solution λ^* leads to an ill-conditioned $\Psi(\lambda^*)$, this means that the parameters of the model cannot be well identified – independent of the method used. Hence, ill-conditioning is an issue with the Nelson-Siegel and Svensson models themselves, which can occur for certain type of zero rate curves, i.e. certain shape parameters λ. Usually, simply shaped curves (e.g., flat curves, i.e. λ close to 0) lead to ill-conditioned solutions as in these cases the models are over-specified. Therefore, the condition number can act as an indicator for an over-specification of the model.

A.3.2.2 A Penalty Approach for Avoiding Ill-conditioned $\Psi(\lambda)$

The most obvious way of dealing with ill-conditioning in the fitting of the Nelson-Siegel and Svensson models is to restrict the parameter space according to the condition number of the



(a) Maturity vector $\tau^{(1)}$

(b) Maturity vector $\tau^{(2)}$

Figure A.2: The condition number of $\Psi(\lambda)$ as a function of λ for the Svensson model with maturity vectors $\tau^{(1)} = (1/4, 1/2, 1, 2, \dots, 15)^\top$ and $\tau^{(2)} = (1/12, 1/4, 1/2, 1, 2, \dots, 15)^\top$.

matrix $\Psi(\lambda)$. However, this approach is rather inconvenient as it bears several issues. Whereas the simple relation between condition number and nonlinear parameter may still allow for an adequate derivation of constraints for the Nelson-Siegel model, see Subfigure A.1(a), it is a fairly demanding task to constrain the parameter space for the Svensson model, see Subfigure A.1(b). Due to the irregularly distributed condition numbers over the parameter space, a suitable restriction only seems possible if the parameter space is modified accordingly, either through transformation or decomposition, or both. In any case, though, the derivation of constraints remains prone to inaccuracies as it depends on the visual amenability of the condition number in one or two dimensions. It thus also lacks theoretical foundation. Finally, the approach is somewhat inflexible since minor changes in the models, or even the use of other models that share the same separable structure, require the constraints to be readjusted.

For these reasons, we suggest a different approach that deals with the ill-conditioning of the matrix $\Psi(\lambda)$ in a more general way, but still ensures the separability of the problem. The approach relies on penalising the objective function f^{sep} if the condition number of $\Psi(\lambda)$ exceeds a maximum allowed level, thus amounting to the objective function

$$f^{\text{pen}}(\lambda) := \left\| \Psi(\lambda)\Psi(\lambda)^\dagger y^{\text{mkt}} - y^{\text{mkt}} \right\|_2^2 + \eta^{\text{pen}} \big(\kappa_\Psi(\lambda) - \kappa^{\max} \big)^+, \tag{A.45}$$

where $\eta^{\text{pen}} > 0$ denotes the weight of the penalisation, $\kappa^{\max}$ the maximum condition number whose exceedance is penalised, and $(x)^+ = \max\{x, 0\}$.

Accordingly, adding a penalty term to the objective function avoids optimal solutions being situated in regions with relatively large condition numbers. Because of the direct relation between the nonlinear parameter λ and the condition number $\kappa_\Psi(\lambda)$ in the objective function f^{pen}, the impact of the condition number can be controlled more effectively than for any restriction of the parameter space. This is a particular advantage if there are no easy to identify regions of the parameters space in which the condition number is large, such as for the Svensson model. A further benefit of the approach lies in its flexibility as it only requires to set the weight parameter η^{pen} and the maximum unpenalised condition number $\kappa^{\max}$.

A potential drawback of this approach is that the function κ_Ψ evaluating the condition number as well as the maximum operator are not everywhere differentiable, such that an algorithm has to be used which is able to deal with nonsmooth objective functions.

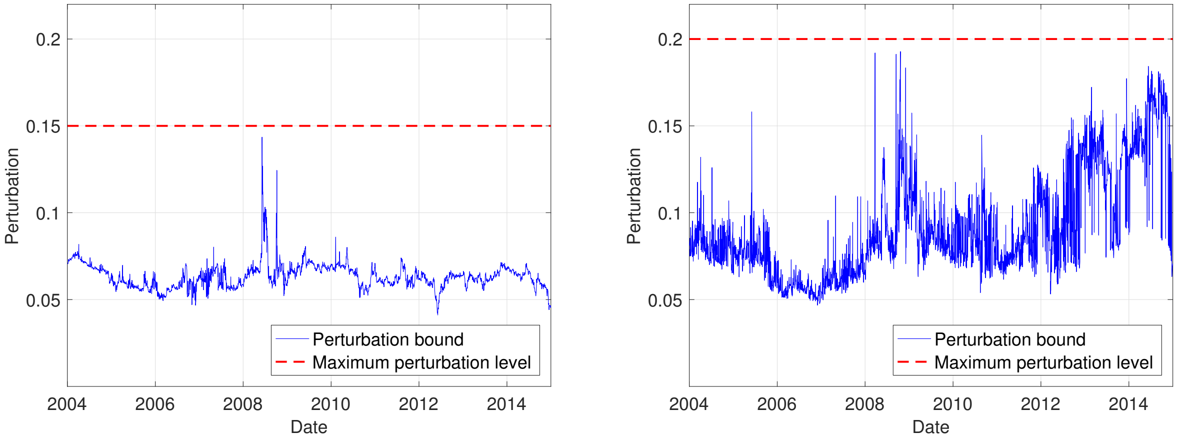
To determine the maximum unpenalised condition number $\kappa^{\max}$ in the penalisation term of the objective f^{pen}, we consider the stability of optimal linear solutions $\beta^*(\lambda^*)$ under perturbations of the matrix $\Psi(\lambda^*)$, according to Björck [1996], p. 29. Specifically, as Theorem A.10 ensures full rank of the matrix $\Psi(\lambda^*)$ for optimal λ^*, the absolute change in $\beta^*(\lambda^*)$ can be bounded under the assumption that $\|\delta\Psi(\lambda^*)\|_2 < \sigma_{d_\beta}(\lambda^*)$ by the first-order result[37]

$$\|\delta\beta^*(\lambda^*)\|_2 \leq \frac{\|\delta\Psi(\lambda^*)\|_2}{\|\Psi(\lambda^*)\|_2}\,\kappa_\Psi(\lambda^*)\left(\|\beta^*(\lambda^*)\|_2 + \frac{\|\Psi(\lambda^*)\beta^*(\lambda^*) - y^{\text{mkt}}\|_2}{\|\Psi(\lambda^*)\|_2}\,\kappa_\Psi(\lambda^*)\right), \qquad \text{(A.46)}$$

cf. Björck [1996], inequality (1.4.18). Now, since the perturbation bound on the right-hand side of inequality (A.46) is mainly influenced by the condition number of $\Psi(\lambda^*)$, it can be controlled to certain extent by the maximum allowed condition number $\kappa^{\max}$. Hence, to avoid situations in which optimal linear solutions become too sensitive to perturbations in $\Psi(\lambda^*)$, the value of $\kappa^{\max}$ should be chosen in such a way that the absolute change in $\beta^*(\lambda^*)$ does not exceed a reasonable level for all fittings at the worst.

For the data set used in Section 4.3.3, we fix the maximum acceptable perturbation in the optimal linear solutions at a level of 0.15 and 0.2, respectively. Thus, to exclude disproportionally large movements, $\beta^*(\lambda^*)$ is not allowed to change by more than 0.15 and 0.2, respectively, which is about five to ten times the average rate level or standard deviation of rates, respectively, cf. Figure 4.4 for historical zero rate levels and their oscillations. This implies that $\kappa^{\max}$ needs to be set to approximately 100 and 180, respectively, to guarantee reasonably stable and moderate parameters. For these values, the perturbation bounds resulting from the fitting of the models to given data are depicted in Figure A.3, along with the maximum perturbation levels. In contrast to $\kappa^{\max}$, the weight η^{pen} of the penalisation is less relevant for the minimisation of f^{pen}. To keep the resulting function values within a reasonable range in ill-conditioned regions of the parameter space, we set $\eta^{\text{pen}} = 10^{-6}$.

[37] Note that we have disregarded perturbations in the data vector y^{mkt} and therefore set $\delta y^{\text{mkt}} = 0$. If deemed relevant, the perturbation bound in (A.46) can easily be adjusted accordingly.



(a) Nelson-Siegel model

(b) Svensson model

Figure A.3: Perturbation bounds of inequality (A.46) with $\|\delta\Psi(\lambda^*)\|_2 = \sigma_{d_\beta}(\lambda^*)$ for fitting the Nelson-Siegel and Svensson models to the data set of Section 4.3.3, where $\kappa^{\max}$ is set to 100 and 180, corresponding to an acceptable perturbation level of 0.15 and 0.2, respectively.

List of Symbols

f deterministic objective function from $\mathbb{R}^d$ into $\mathbb{R}$.

$\mathcal{X}$ compact parameter space in $\mathbb{R}^d$.

g nonnegative function measuring the difference between model and market prices.

Π model pricing function of calibration instruments from $\mathcal{X}$ into $\mathbb{R}^l$; expected value of the discounted payoff function.

$h(x, Z)$ discounted payoff function for $x \in \mathcal{X}$ and random vector Z supported on $\mathcal{Z} \subset \mathbb{R}^{d_Z}$.

C^{mkt} vector of market prices in $\mathbb{R}^l$.

$\widehat{\Pi}_N$ Monte Carlo estimator of Π in the SAA strategy, based on the i.i.d. random vectors $Z_1, \dots, Z_N$, $N \in \mathbb{N}$.

$\hat{f}_N$ estimator of $f = g(\Pi - C^{\mathrm{mkt}})$, used for fixed N in the SAA strategy with Monte Carlo estimator $\widehat{\Pi}_N$.

$\widehat{\Pi}_{N_k}$ Monte Carlo estimator of Π in the VSAA strategy, based on the i.i.d. random vectors $Z_1^k, \dots, Z_{N_k}^k$, $N_k \in \mathbb{N}$.

$\hat{f}_{N_k}$ estimator of $f = g(\Pi - C^{\mathrm{mkt}})$, used at the k-th evaluation in the VSAA strategy with Monte Carlo estimator $\widehat{\Pi}_{N_k}$.

$\widetilde{X}$ $C(\mathcal{X}, \mathbb{R}^l)/C^1(\mathcal{X}, \mathbb{R}^l)/\mathbb{R}^l$-valued random variable $h(\,\cdot\,, Z) - \mathbb{E}^{\mathbb{Q}}[h(\,\cdot\,, Z)]$.

$\widetilde{X}_i$ $C(\mathcal{X}, \mathbb{R}^l)/C^1(\mathcal{X}, \mathbb{R}^l)$-valued random variable $h(\,\cdot\,, Z_i) - \mathbb{E}^{\mathbb{Q}}[h(\,\cdot\,, Z_i)]$ in the SAA strategy.

$\widetilde{X}_{ki}$ $C(\mathcal{X}, \mathbb{R}^l)/\mathbb{R}^l$-valued random variable $h(\,\cdot\,, Z_i^k) - \mathbb{E}^{\mathbb{Q}}[h(\,\cdot\,, Z_i^k)]$ in the VSAA strategy.

ϕ radial basis function.

$\mathcal{P}_m^d$ space of polynomials of total degree at most $m-1$ in $\mathbb{R}^d$.

$\mathcal{F}_\phi(\mathcal{D})$ space of linear combinations of radial basis functions $\phi(\|\cdot - x\|_2)$, $x \in \mathcal{D}$.

$\mathcal{A}_\phi(\mathcal{D})$ direct sum of $\mathcal{F}_\phi(\mathcal{D})$ and $\mathcal{P}_m^d$.

$\|\cdot\|_\phi$ semi-norm on $\mathcal{A}_\phi(\mathcal{D})$, induced by the semi-inner product $\langle\cdot,\cdot\rangle_\phi$.

$\mathcal{N}_\phi(\mathcal{D})$ native space of the radial basis function ϕ on $\mathcal{D}$.

$\|\cdot\|_{\mathcal{N}_\phi}$ semi-norm on $\mathcal{N}_\phi(\mathcal{D})$, induced by the semi-inner product $\langle\cdot,\cdot\rangle_{\mathcal{N}_\phi}$.

s_n interpolant from $\mathcal{A}_\phi(\mathcal{D})$ to the data $(x_1, f(x_1)), \ldots, (x_n, f(x_n))$.

f_n^* target value in the n-th iteration of the original RBF method and the RBF method for noisy objective functions.

$l_n(y,\cdot)$ interpolant from $\mathcal{A}_\phi(\mathcal{D})$ to the data $(x_1, 0), \ldots, (x_n, 0)$ and $(y, 1)$.

g_n utility function on $\mathcal{X}\backslash\{x_1, \ldots, x_n\}$ to be minimised for obtaining x_{n+1}.

$\Delta_n(y)$ function assigning the minimum Euclidean distance of $y \in \mathcal{X}$ to the set $\{x_1, \ldots, x_n\}$.

f_k residual function from $\mathbb{R}^d$ into $\mathbb{R}$, forming the objective function $f = \sum_{i=1}^{l} g(f_k)$.

$s_{n,k}$ interpolant from $\mathcal{A}_\phi(\mathcal{D})$ to the data $(x_1, f_k(x_1)), \ldots, (x_n, f_k(x_n))$.

$f_{n,k}^*$ target value for the k-th residual in the n-th iteration of the modified RBF method.

$g_{n,k}$ utility function on $\mathcal{X}\backslash\{x_1, \ldots, x_n\}$ for the k-th residual in the weighted sum scalarisation to be minimised for obtaining x_{n+1}.

$\hat{f}$ noisy objective function from $\mathbb{R}^d$ into $\mathbb{R}$.

$s_n^{\gamma_n}$ regularised least-squares approximant with parameter γ_n from $\mathcal{A}_\phi(\mathcal{D})$ to the data $(x_1, \hat{f}(x_1)), \ldots, (x_n, \hat{f}(x_n))$.

$l_n^{\gamma_n}(y,\cdot)$ regularised least-squares approximant with parameter γ_n from $\mathcal{A}_\phi(\mathcal{D})$ to the data $(x_1, 0), \ldots, (x_n, 0)$, subject to interpolating $(y, 1)$.

$g_n^{\gamma_n}$ utility function with parameter γ_n on $\mathcal{X}\backslash\{x_1, \ldots, x_n\}$ to be minimised for obtaining x_{n+1}.

$\widetilde{w}_n(y)$ function assigning the weight w_i of the sample point x_i, $i \in \{1, \ldots, n\}$, that is closest to $y \in \mathcal{X}$.

Bibliography

Aliprantis, C. D. and Border, K. (2006). *Infinite Dimensional Analysis - A Hitchhiker's Guide*. Springer, Berlin Heidelberg.

Angelini, F. and Herzel, S. (2002). Consistent Initial Curves for Interest Rate Models. *The Journal of Derivatives*, 9(4):8–17.

Annaert, J., Claes, A. G. P., De Ceuster, M. J. K., and Zhang, H. (2013). Estimating the spot rate curve using the Nelson-Siegel model: A ridge regression approach. *International Review of Economics and Finance*, 27:482–496.

Araujo, A. and Giné, E. (1980). *The Central Limit Theorem for Real and Banach Valued Random Variables*. Wiley, New York, NY.

Atkins, R. (2014). Europe shows negative interest rates not absurd – and might work. The Financial Times Limited, 18 September 2014, Available at: `http://www.ft.com/cms/s/0/db1f5da4-3e89-11e4-a620-00144feabdc0.html#axzz41Cvbt4Ux`.

Banholzer, D., Fliege, J., and Werner, R. (2017a). A Modified RBF Method with Extended Local Search. Working paper.

Banholzer, D., Fliege, J., and Werner, R. (2017b). A RBF Method for Noisy Objective Functions. Working paper.

Banholzer, D., Fliege, J., and Werner, R. (2017c). On the inherent instability of the Nelson-Siegel and the Svensson models and potential remedies. Working paper.

Banholzer, D., Fliege, J., and Werner, R. (2018a). On rates of convergence for sample average approximations in the almost sure sense and in mean. *Optimization Online*, paper 5834.

Banholzer, D., Fliege, J., and Werner, R. (2018b). Uniform strong consistency and sample path bounds for variable-sample average approximations. Working paper.

Bates, C. and White, H. (1985). A unified theory of consistent estimation for parametric models. *Econometric Theory*, 1(2):151–178.

Bauer, H. (2001). *Measure and Integration Theory*. De Gruyter, Berlin.

Björck, A. (1996). *Numerical Methods for Least Squares Problems*. SIAM, Philadelphia, PA.

Björk, T. (2009). *Arbitrage Theory in Continuous Time*. Oxford University Press, New York, NY.

Björkman, M. and Holmström, K. (2000). Global Optimization of Costly Nonconvex Functions Using Radial Basis Functions. *Optimization and Engineering*, 1(4):373–397.

Black, F. (1976). The pricing of commodity contracts. *Journal of Financial Economics*, 3(1–2):167–179.

Bonnans, J. F. and Shapiro, A. (2000). *Perturbation Analysis of Optimization Problems*. Springer, New York, NY.

Brigo, D. and Mercurio, F. (2006). *Interest Rate Models – Theory and Practice (With Smile, Inflation and Credit)*. Springer, Berlin Heidelberg.

Buhmann, M. D. (1998). Radial functions on compact support. *Proceedings of the Edinburgh Mathematical Society*, 41(1):33–46.

Buhmann, M. D. (2003). *Radial Basis Functions: Theory and Implementations*. Cambridge University Press, Cambridge.

Cairns, A. J. G. and Pritchard, D. J. (2001). Stability of descriptive models for the term structure of interest rates with application to German market data. *British Actuarial Journal*, 7(3):467–507.

Cassioli, A. and Schoen, F. (2013). Global optimization of expensive black box problems with a known lower bound. *Journal of Global Optimization*, 57(1):177–190.

Costa, A. and Nannicini, G. (2014). RBFOpt: an open-source library for black-box optimization with costly function evaluations. *Optimization Online*, paper 4538.

Cox, D. D. and John, S. (1997). SDO: A Statistical Method for Global Optimization. In Alexandrov, N. M. and Hussaini, M. Y., editors, *Multidisciplinary Design Optimization: State-of-the-Art*, pages 315–329. SIAM, Philadelphia, PA.

Cressie, N. A. C. (1991). *Statistics for Spatial Data*. Wiley, New York, NY.

Dai, L., Chen, C. H., and Birge, J. R. (2000). Convergence properties of two-stage stochastic programming. *Journal of Optimization Theory and Applications*, 106(3):489–509.

Danskin, J. M. (1966). The theory of max-min, with applications. *SIAM Journal on Applied Mathematics*, 14(4):641–664.

De Marchi, S., Schaback, R., and Wendland, H. (2005). Near-optimal data-independent point locations for radial basis function interpolation. *Advances in Computational Mathematics*, 23(3):317–330.

De Pooter, M. (2007). Examining the Nelson-Siegel Class of Term Structure Models: In-Sample Fit versus Out-of-Sample Forecasting Performance. *Discussion Paper 2007-043/4, Tinbergen Institute, Erasmus University.*

De Rezende, R. B. (2011). Giving Flexibility to the Nelson-Siegel Class of Term Structure Models. *Revista Brasileira de Finanças*, 9(1):27–49.

Dembo, A. and Zeitouni, O. (1998). *Large Deviations Techniques and Applications*. Springer, New York, NY.

Diebold, F. X. and Li, C. (2006). Forecasting the Term Structure of Government Bond Yields. *Journal of Econometrics*, 130(2):337–364.

Diebold, F. X. and Rudebusch, G. D. (2013). *Yield Curve Modeling and Forecasting: The Dynamic Nelson-Siegel Approach*. Princeton University Press, Princeton, NJ.

Diestel, J. and Uhl, J. J. (1977). *Vector Measures*. American Mathematical Society, Providence, RI.

Dieudonné, J. (1960). *Foundations of Modern Analysis*. Academic Press, New York, NY.

Dixon, L. C. W. and Szegö, G. P. (1978). The global optimisation problem: an introduction. In Dixon, L. C. W. and G.P., S., editors, *Towards Global Optimization 2*, pages 1–15. North-Holland, Amsterdam.

Domowitz, I. and White, H. (1982). Misspecified models with dependent observations. *Journal of Econometrics*, 20(1):35–58.

Dupačová, J. and Wets, R. J.-B. (1988). Asymptotic Behavior of Statistical Estimators and of Optimal Solutions of Stochastic Optimization Problems. *The Annals of Statistics*, 16(4):1517–1549.

Edman, C. (2016). *Black box optimization with exact subsolvers – A radial basis function algorithm for problems with convex constraints*. PhD thesis, Universität Trier.

Ehrgott, M. (2005). *Multicriteria Optimization*. Springer, Berlin Heidelberg.

Einmahl, U. and Li, D. (2008). Characterization of LIL behavior in Banach space. *Transactions of the American Mathematical Society*, 360(12):6677–6693.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 226–231. AAAI Press, Portland, OR.

Evgeniou, T., Pontil, M., and Poggio, T. (2000). Regularization Networks and Support Vector Machines. *Advances in Computational Mathematics*, 13(1):1–50.

Fabozzi, F. J., Martellini, L., and Priaulet, P. (2005). Predictability in the Shape of the Term Structure of Interest Rates. *Journal of Fixed Income*, 15(1):40–53.

Fasshauer, G. E. (2007). *Meshfree Approximation Methods with MATLAB*. World Scientific, Singapore.

Fernique, X. (1970). Intégrabilité des vecteurs gaussiens. *Comptes rendus de l'Académie des Sciences Paris*, 270(25):1698–1699.

Finkel, D. E. (2004). DIRECT – A Global Optimization Algorithm. Available at: `https://ctk.math.ncsu.edu/Finkel_Direct/`.

Föllmer, H. and Schied, A. (2004). *Stochastic Finance: An Introduction in Discrete Time*. De Gruyter, Berlin.

Forrester, A. I. J. and Keane, A. J. (2009). Recent advances in surrogate-based optimization. *Progress in Aerospace Sciences*, 45(1–3):50–79.

Forrester, A. I. J., Sobester, A., and Keane, A. J. (2008). *Engineering Design via Surrogate Modelling: A Practical Guide*. Wiley, Chichester.

Fowkes, J. (2011). *Bayesian Numerical Analysis: Global Optimization and Other Applications*. PhD thesis, University of Oxford.

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1):1–67.

Gallant, A. R. and White, H. (1988). *Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. Blackwell, Oxford.

Gauthier, G. and Simonato, J.-G. (2012). Linearized Nelson-Siegel and Svensson models for the estimation of spot interest rates. *European Journal of Operational Research*, 219(2):442–451.

Geletu, A. (2006). Introduction to Topological Spaces and Set-Valued Maps (Lecture Notes). Available at: `http://www.tu-ilmenau.de/en/simulation-and-optimal-processes-group/staff/dr-rer-nat-abebe-geletu-w-selassie/`.

Gilli, M., Grosse, S., and Schumann, E. (2010). Calibrating the Nelson-Siegel-Svensson model. *COMISEF Working Paper Series No. 31.*

Girosi, F. (1992). Some extensions of radial basis functions and their applications in artificial intelligence. *Computers & Mathematics with Applications*, 24(12):61–80.

Glasserman, P. (2003). *Monte Carlo Methods in Financial Engineering*. Springer, New York, NY.

Gnedenko, B. V. and Kolmogorov, A. N. (1954). *Limit Distributions for Sums of Independent Random Variables.* Addison-Wesley, Reading, MA. Translated from Russian by Chung, K. L. with appendices by Doob, J. L. and Hsu, P. L. (revised edition 1968).

Golub, G. H. and Pereyra, V. (1973). The Differentiation of Pseudo-Inverses and Nonlinear Least Squares Problems Whose Variables Separate. *SIAM Journal on Numerical Analysis*, 10(2):413–432.

Goodman, V., Kuelbs, J., and Zinn, J. (1981). Some Results on the LIL in Banach Space with Applications to Weighted Empirical Processes. *The Annals of Probability*, 9(5):713–752.

Gramacy, R. B. and Lee, H. K. H. (2011). Optimization Under Unknown Constraints. In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M., editors, *Bayesian Statistics 9*. Oxford University Press, Oxford.

Guo, K., Hu, S., and Sun, X. (1993). Conditionally Positive Definite Functions and Laplace-Stieltjes Integrals. *Journal of Approximation Theory*, 74(3):249–265.

Gutmann, H.-M. (2001a). A Radial Basis Function Method for Global Optimization. *Journal of Global Optimization*, 19(3):201–227.

Gutmann, H.-M. (2001b). *Radial Basis Function Methods for Global Optimization.* PhD thesis, University of Cambridge.

Hansen, P. C. (2010). *Discrete Inverse Problems: Insight and Algorithms.* SIAM, Philadelphia, PA.

Hartman, P. and Wintner, A. (1941). On the Law of the Iterated Logarithm. *American Journal of Mathematics*, 63(1):169–176.

Hildebrandt, S. (2003). *Analysis 2*. Springer, Berlin Heidelberg.

Hirsa, A. (2012). *Computational Methods in Finance.* Chapman & Hall/CRC, Boca Raton, FL.

Holmström, K. (2008). An adaptive radial basis algorithm (ARBF) for expensive black-box global optimization. *Journal of Global Optimization*, 41(3):447–464.

Homem-de-Mello, T. (2003). Variable-sample methods for stochastic optimization. *ACM Transactions on Modeling and Computer Simulation*, 13(2):108–133.

Homem-de-Mello, T. (2008). On Rates of Convergence for Stochastic Optimization Problems Under Non-Independent and Identically Distributed Sampling. *SIAM Journal on Optimization*, 19(2):524–551.

Homem-de-Mello, T. and Bayraksan, G. (2014). Monte Carlo sampling-based methods for stochastic optimization. *Surveys in Operations Research and Management Science*, 19(1):56–85.

Horst, R. and Pardalos, P. M. (1995). *Handbook of Global Optimization.* Kluwer Academic Publishers, Dordrecht.

Hu, T.-C., Móricz, F., and Taylor, R. (1989). Strong laws of large numbers for arrays of rowwise independent random variables. *Acta Mathematica Hungarica*, 54(1–2):153–162.

Hu, T.-C., Rosalsky, A., Szynal, D., and Volodin, A. I. (1999). On complete convergence for arrays of rowwise independent random elements in Banach spaces. *Stochastic Analysis and Applications*, 17(6):963–992.

Huang, D., Allen, T. T., Notz, W. I., and Zeng, N. (2006). Global Optimization of Stochastic Black-Box Systems via Sequential Kriging Meta-Models. *Journal of Global Optimization*, 34(3):441–466.

Hull, J. and White, A. (1990). Pricing interest-rate-derivative securities. *The Review of Financial Studies*, 3(4):573–592.

Hull, J. C. (2017). *Options, Future, and Other Derivatives.* Pearson, London.

Huyer, W. and Neumaier, A. (1999). Global Optimization by Multilevel Coordinate Search. *Journal of Global Optimization*, 14(4):331–355.

Huyer, W. and Neumaier, A. (2000). MCS – Global Optimization by Multilevel Coordinate Search. Available at: `https://www.mat.univie.ac.at/~neum/software/mcs/`.

Iske, A. (2000). Optimal Distribution of Centers for Radial Basis Function Methods. Technical report, Technische Universität München.

Iske, A. (2004). *Multiresolution Methods in Scattered Data Modelling.* Springer, Berlin Heidelberg.

Jakobsson, S., Patriksson, M., Rudholm, J., and Wojciechowski, A. (2010). A method for simulation based optimization using radial basis function. *Optimization and Engineering*, 11(4):501–532.

Jamshidian, F. (1989). An Exact Bond Option Formula. *The Journal of Finance*, 44(1):205–209.

Ji, Y., Kim, S., and Lu, W. X. (2013). A new framework for combining global and local methods in black box optimization. *Optimization Online*, paper 3977.

Jones, D. R. (1996). Global Optimization with Response Surfaces. Fifth SIAM Conference on Optimization, Victoria, Canada.

Jones, D. R. (2001). A Taxonomy of Global Optimization Methods Based on Response Surfaces. *Journal of Global Optimization*, 21(4):345–383.

Jones, D. R., Perttunen, C. D., and Stuckman, B. E. (1993). Lipschitz optimization without the Lipschitz constant. *Journal of Optimization Theory and Applications*, 78(1):157–181.

Jones, D. R., Schonlau, M., and Welch, W. J. (1998). Efficient Global Optimization of Expensive Black-Box Functions. *Journal of Global Optimization*, 13(4):455–492.

Kaniovski, Y. M., King, A. J., and Wets, R. J.-B. (1995). Probabilistic bounds (via large deviations) for the solutions of stochastic programming problems. *Annals of Operations Research*, 56(1):189–208.

Kelley, C. T. (1999). *Iterative Methods for Optimization.* SIAM, Philadelphia, PA.

Kelley, C. T. (2001). *Implicit Filtering.* SIAM, Philadelphia, PA.

Kienitz, J. (2017). Negative Rates: New Market Practice. In Ehrhardt, M., Günther, M., and ter Maten, E. J. W., editors, *Novel Methods in Computational Finance*, Mathematics in Industry, pages 47–63. Springer, Cham.

Kienitz, J. and Wetterau, D. (2012). *Financial Modelling: Theory, Implementation and Practice (with MATLAB Source).* Wiley, Chichester.

Kim, S., Pasupathy, R., and Henderson, S. G. (2015). A Guide to Sample Average Approximation. In Fu, M. C., editor, *Handbook of Simulation Optimization*, volume 216 of *International Series in Operations Research & Management Science*, pages 207–243. Springer, New York, NY.

King, A. J. and Rockafellar, R. T. (1993). Asymptotic Theory for Solutions in Statistical Estimation and Stochastic Programming. *Mathematics of Operations Research*, 18(1):148–162.

Klenke, A. (2008). *Probability Theory: A Comprehensive Course.* Springer, London.

Kleywegt, A., Shapiro, A., and Homem-de Mello, T. (2001). The Sample Average Approximation Method for Stochastic Discrete Optimization. *SIAM Journal on Optimization*, 12(2):479–502.

Knott, G. D. (2000). *Interpolating Cubic Splines.* Birkhäuser, New York, NY.

Kuelbs, J. (1976a). A Strong Convergence Theorem for Banach Space Valued Random Variables. *The Annals of Probability*, 4(5):744–771.

Kuelbs, J. (1976b). A Counterexample for Banach Space Valued Random Variables. *The Annals of Probability*, 4(4):684–689.

Kuelbs, J. (1977). Kolmogorov's law of the iterated logarithm for Banach space valued random variables. *Illinois Journal of Mathematics*, 21(4):784–800.

Kushner, H. J. (1962). A versatile stochastic model of a function of unknown and time varying form. *Journal of Mathematical Analysis and Applications*, 5(1):150–167.

Kushner, H. J. (1964). A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise. *Journal of Basic Engineering*, 86(1):97–106.

Ledoux, M. and Talagrand, M. (1988). Characterization of the Law of the Iterated Logarithm in Banach Spaces. *The Annals of Probability*, 16(3):1242–1264.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes.* Springer, Berlin Heidelberg.

Levin, D. (1998). The approximation power of moving least-squares. *Mathematics of Computation*, 67(224):1517–1531.

Li, D., Rao, M., and Tomkins, R. (1995). A Strong Law for B-Valued Arrays. *Proceedings of the American Mathematical Society*, 123(10):3205–3212.

Locatelli, M. and Schoen, F. (2013). *Global Optimization: Theory, Algorithms, and Applications.* SIAM, Philadelphia, PA.

Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, 58(8):1246–1266.

Micchelli, C. A. (1986). Interpolation of scattered data: Distance matrices and conditionally positive definite functions. *Constructive Approximation*, 2(1):11–22.

Mockus, J., Tiesis, V., and Žilinskas, A. (1978). The Application of Bayesian Methods for Seeking the Extremum. In Dixon, L. C. W. and G.P., S., editors, *Towards Global Optimization 2*, pages 117–128. North-Holland, Amsterdam.

Moré, J. J., Garbow, B. S., and Hillstrom, K. E. (1981). Testing Unconstrained Optimization Software. *ACM Transactions on Mathematical Software*, 7(1):17–41.

Musiela, M. and Rutkowski, M. (2005). *Martingale Methods in Financial Modelling.* Springer, Berlin Heidelberg.

Myers, R. H. and Montgomery, D. C. (1995). *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. Wiley, New York, NY.

Nelson, C. R. and Siegel, A. F. (1987). Parsimonious Modeling of Yield Curves. *Journal of Business*, 60(4):473–489.

Neumaier, A. (2004). Complete search in continuous global optimization and constraint satisfaction. *Acta Numerica*, 13:271–369.

Nocedal, J. and Wright, S. (2006). *Numerical Optimization*. Springer, New York, NY.

Pardalos, P. M. and Romeijn, H. E. (2002). *Handbook of Global Optimization: Volume 2*. Kluwer Academic Publishers, Dordrecht.

Pisier, G. (1975). Le théorème de la limite centrale et la loi du logarithme itérée dans les espaces de Banach. *Séminaire Maurey-Schwartz 1975-1976, and exposés III et IV.*

Pisier, G. and Zinn, J. (1978). On the limit theorems for random variables with values in the spaces L_p $(2 \leq p < \infty)$. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 41(4):289–304.

Plambeck, E. L., Fu, B. R., Robinson, S. M., and Suri, R. (1996). Sample-path optimization of convex stochastic performance functions. *Mathematical Programming*, 75(2):137–176.

Powell, M. J. D. (1992). The theory of radial basis function approximation in 1990. In Light, W. A., editor, *Advances in Numerical Analysis II: Wavelets, Subdivision Algorithms, and Radial Basis Functions*, pages 105–210. Oxford University Press, Oxford.

Powell, M. J. D. (1996). A review of algorithms for thin plate spline interpolation. In Fontanella, F., Jetter, K., and Laurent, P. J., editors, *Advanced Topics in Multivariate Approximation*, pages 303–322. World Scientific, Singapore.

Powell, M. J. D. (2002). UOBYQA: unconstrained optimization by quadratic approximation. *Mathematical Programming*, 92(3):555–582.

Pukelsheim, F. (2006). *Optimal Design of Experiments*. SIAM, Philadelphia, PA.

Regis, R. G. and Shoemaker, C. A. (2005). Constrained Global Optimization of Expensive Black Box Functions Using Radial Basis Functions. *Journal of Global Optimization*, 31(1):153–171.

Regis, R. G. and Shoemaker, C. A. (2007a). A Stochastic Radial Basis Function Method for the Global Optimization of Expensive Functions. *INFORMS Journal on Computing*, 19(4):497–509.

Regis, R. G. and Shoemaker, C. A. (2007b). Improved strategies for radial basis function methods for global optimization. *Journal of Global Optimization*, 37(1):113–135.

Regis, R. G. and Shoemaker, C. A. (2007c). Parallel radial basis function methods for the global optimization of expensive functions. *European Journal of Operational Research*, 182(2):514–535.

Regis, R. G. and Shoemaker, C. A. (2013). A quasi-multistart framework for global optimization of expensive functions using response surface models. *Journal of Global Optimization*, 56(4):1719–1753.

Robinson, S. M. (1996). Analysis of Sample-Path Optimization. *Mathematics of Operations Research*, 21(3):513—528.

Royset, J. O. (2013). On sample size control in sample average approximations for solving smooth stochastic programs. *Computational Optimization and Applications*, 55(2):265–309.

Rubinstein, R. Y. and Shapiro, A. (1990). Optimization of static simulation models by the score function method. *Mathematics and Computers in Simulation*, 32(4):373–392.

Rubinstein, R. Y. and Shapiro, A. (1993). *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method*. Wiley, Chichester.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989). Design and Analysis of Computer Experiments. *Statistical Science*, 4(4):409–423.

Schaback, R. (1993). Comparison of Radial Basis Function Interpolants. In Jetter, K. and Utreras, F. I., editors, *Multivariate Approximation: From CAGD to Wavelets*, pages 293–305. World Scientific, Singapore.

Schaback, R. (1999). Native spaces for radial basis functions I. In Müller, M. W., Buhmann, M. D., Mache, D. H., and Felten, M., editors, *New Developments in Approximation Theory*, pages 255–282. Birkhäuser, Basel.

Schaback, R. and Wendland, H. (2001). Characterization and construction of radial basis functions. In Dyn, N., Leviatan, D., Levin, D., and Pinkus, A., editors, *Multivariate Approximation and Applications*. Cambridge University Press, Cambridge.

Schaback, R. and Wendland, H. (2006). Kernel techniques: From machine learning to meshless methods. *Acta Numerica*, 15:543–639.

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA.

Schonlau, M. (1997). *Computer Experiments and Global Optimization*. PhD thesis, University of Waterloo.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York, NY.

Shapiro, A. (1989). Asymptotic Properties of Statistical Estimators in Stochastic Programming. *Annals of Statistics*, 17(2):841–858.

Shapiro, A. (1990). On concepts of directional differentiability. *Journal of Optimization Theory and Applications*, 66(3):477–487.

Shapiro, A. (1991). Asymptotic analysis of stochastic programs. *Annals of Operations Research*, 30(1):169–186.

Shapiro, A. (1993). Asymptotic Behavior of Optimal Solutions in Stochastic Programming. *Mathematics of Operations Research*, 18(4):829–845.

Shapiro, A. (2000). Statistical inference of stochastic optimization problems. In P., U. S., editor, *Probabilistic Constrained Optimization*, volume 49 of *Nonconvex Optimization and Its Applications*, pages 282–304. Springer, Boston, MA.

Shapiro, A. (2003). Monte Carlo Sampling Methods. In Ruszczyński, A. and Shapiro, A., editors, *Stochastic Programming*, volume 10 of *Handbooks in Operations Research and Management Science*, pages 353–425. Elsevier, Amsterdam.

Shapiro, A., Dentcheva, D., and Ruszczyński, A. (2014). *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, Philadelphia, PA.

Shapiro, A. and Homem-de-Mello, T. (2000). On the Rate of Convergence of Optimal Solutions of Monte Carlo Approximations of Stochastic Programs. *SIAM Journal on Optimization*, 11(1):70—86.

Singer, I. (1970). *Best Approximation in Normed Linear Spaces by Elements of Linear Subspaces*, volume 171 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin Heidelberg.

Sóbester, A., Leary, S. J., and Keane, A. J. (2005). On the Design of Optimization Strategies Based on Global Response Surface Approximation Models. *Journal of Global Optimization*, 33(1):31–59.

Strassen, V. (1964). An invariance principle for the law of the iterated logarithm. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 3(3):211–226.

Strassen, V. (1966). A converse to the law of the iterated logarithm. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4(4):265–268.

Svensson, L. E. O. (1995). Estimating Forward Interest Rates with the Extended Nelson and Siegel Method. *Sveriges Riksbank Quarterly Review*, 3:13–26.

Törn, A. and Žilinskas, A. (1989). *Global Optimization*. Springer, Berlin Heidelberg.

Vasicek, O. (1977). An equilibrium characterization of the term structure. *Journal of Financial Economics*, 5(2):177–188.

Villemonteix, J., Vazquez, E., and Walter, E. (2009). An informational approach to the global optimization of expensive-to-evaluate functions. *Journal of Global Optimization*, 44(4):509–534.

Virmani, V. (2012). On estimability of parsimonious term structure models: an experiment with the Nelson-Siegel specification. *Applied Economics Letters*, 19(17):1703–1706.

Vu, K. K., D'Ambrosio, C., Hamadi, Y., and Liberti, L. (2017). Surrogate-based methods for black-box optimization. *International Transactions in Operational Research*, 24(3):393–424.

Žilinskas, A. (2010). On similarities between two models of global optimization: statistical models and radial basis functions. *Journal of Global Optimization*, 48(1):173–182.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia, PA.

Wahba, G. (1999). Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. In Schölkopf, B., Burges, C. J. C., and Smola, A. J., editors, *Advances in Kernel Methods*, pages 69–88. MIT Press, Cambridge, MA.

Wendland, H. (1995). Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Advances in Computational Mathematics*, 4(1):389–396.

Wendland, H. (2005a). Computational aspects of radial basis function approximation. In Jetter, K., Buhmann, M. D., Haussmann, W., Schaback, R., and Stöckler, J., editors, *Topics in Multivariate Approximation and Interpolation*, volume 12, pages 231–256. Elsevier, Amsterdam.

Wendland, H. (2005b). *Scattered Data Approximation*. Cambridge University Press, Cambridge.

Wendland, H. and Rieger, C. (2005). Approximate Interpolation with Applications to Selecting Smoothing Parameters. *Numerische Mathematik*, 101(4):729–748.

Wu, Z. (1995). Compactly supported positive definite radial functions. *Advances in Computational Mathematics*, 4:283–292.