Efficient Global Surrogate Models for Responses of Expensive Simulations

Dissertation

zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat.)

Dem Fachbereich IV der Universität Trier
vorgelegt von

Benjamin Rosenbaum

Trier, 2013

Gutachter: Prof. Dr. Volker Schulz
Prof. Dr. Peter Benner

Zusammenfassung

Diese Arbeit beschäftigt sich mit der globalen Ersatzmodellierung bei teuren, computergestützen Simulationen. Die numerische Strömungssimulation hat sich zu einer entscheidenden Schlüsseltechnologie im Flugzeugbau entwickelt. Jedoch ist insbesondere die Rechenzeit für vollständige Simulationen ganzer Flugszenarien noch inakzeptabel hoch und kann nur durch den Einsatz von Ersatzmodellen beherrscht werden. So sollen beispielsweise die aerodynamischen Beiwerte einer bestimmten Flügelgeometrie nicht nur für eine einzige Konfiguration, sondern über eine ganze Spanne von verschiedenen Werten für Geschwindigkeit und Anstellwinkel berechnet werden. Selbst auf modernen Hochleistungsrechnern sind hierbei Rechenzeiten von mehreren Tagen für eine einzige Simulation keine Seltenheit. Anstatt nun eine Vielzahl an Simulationen für verschiedene Konfigurationen der Eingangsparameter durchzuführen, um das globale Verhalten einer Antwortfunktion unter Variation der Parameter zu beschreiben, werden in der Ersatzmodellierung nur wenige, geschickt gewählte Auswertungen in Stützststellen getätigt und interpoliert. Dadurch können hinreichend genaue Approximationen der unbekannten Antwortfunktion erzeugt werden, welche nur einen Bruchteil der herkömmlichen Rechenzeit benötigen.

Mit Kriging können hochgradig nichtlineare, deterministische Funktionen basierend auf beliebig verteilten Stützsstellen interpoliert werden. Durch die Modellierung über Korrelationsfunktionen können unterschiedliche Sensitivitäten bezüglich der Eingangsparameter bei der Antwortfunktion automatisch berücksichtigt werden. Sind über die reinen Auswertungen der Antwortfunktion hinaus auch Gradienteninformationen erhältlich, beispielsweise durch adjungierte Rechnungen, können diese in die Interpolation aufgenommen werden, um die globale Approximationsgüte zu verbessern. Mit adaptiven Versuchsplänen lassen sich effizientere Ersatzmodelle generieren. Im Gegensatz zu herkömmlichen nichtadaptiven, lediglich raumfüllenden Verteilungen der Stützstellen werden sowohl das Ersatzmodell als auch der Versuchsplan sequentiell generiert. Wissen über das Verhalten der Antwortfunktion kann somit in die Wahl der Stützstellen einfließen, um eine problemspezifische Interpolation zu generieren. Kritische Bereiche des Eingangsparameterbereichs, in denen beispielsweise große Steigungen oder Krümmungen der Antwortfunktion auftreten, werden automatisch identifiziert und dichter ausgewertet, um das Funktionsverhalten korrekt zu reproduzieren. Die Anzahl der benötigten teuren Simulationen wird somit erheblich verringert.

Oft sind solche Antwortfunktionen nicht beliebiger Natur, sondern gehören einer größeren Problemklasse an. So weisen die aerodynamischen Beiwertsfunktionen verschiedener Flügelgeometrien strukturelle Ähnlichkeiten auf. In dieser Arbeit wird der Frage nachgegangen, wie diese gemeinsamen Strukturen identifiziert und damit für neue Testfälle gewinnbringend genutzt werden können. Zu einer Datenbank von Antwortfunktionen verschiedener Flügelgeometrien wird mit Hilfe einer Hauptkomponentenanalyse ein generisches Ersatzmodell definiert. Somit können charakteristische Strukturen der gemeinsamen Problemklasse verwendet werden, um die Approximationsgüte für eine neue Flügelgeometrie zu verbessern.

Bei der herkömmlichen Ersatzmodellierung ist trotz Fortschritten, beispielweise durch die untersuchte adaptive Stützstellenauswahl, immer noch eine hohe Anzahl an Auswertungen notwendig, um über einen breiten Bereich der Eingangsparameter eine global hohe Approximationsgenauigkeit zu erreichen. Bei der Validierung des neuartigen Ansatzes an praxisrelevanten Testfällen zeigt sich, dass dieser deutlich effizientere Ersatzmodelle ermöglicht, deren Approximationsfehler signifikant geringer als bei herkömmlichen Interpolationen ausfällt. Dadurch sind praxistaugliche Ersatzmodelle bereits für eine geringe Anzahl an teuren Simulationen realisierbar.

Bisher wurden solche Interpolationsprobleme einzeln betrachtet. Der beschriebene Ansatz ist somit innovativ hinsichtlich der Nutzung struktureller Ähnlichkeiten einer gemeinsamen Problemklasse für die Ersatzmodellierung. Auf neuartige Weise werden Ansätze aus der mathematischen Ersatzmodellierung, der Methoden unterschiedlicher Eindringtiefe, der statistischen Versuchsplanung und insbesondere der Bildregistrierung und statistischer Bildanalyse interdisziplinär verbunden. Dabei ist die Anwendbarkeit nicht nur auf die aerodynamische Simulation beschränkt. Generische Ersatzmodelle können genutzt werden, wann immer teure Simulationen einer größeren Problemklasse zuzuordnen sind, in der hinreichende strukturelle Ähnlichkeiten vermutet werden.

Abstract

In this thesis, global surrogate models for responses of expensive simulations are investigated. Computational fluid dynamics (CFD) have become an indispensable tool in the aircraft industry. But simulations of realistic aircraft configurations remain challenging and computationally expensive despite the sustained advances in computing power. With the demand for numerous simulations to describe the behavior of an output quantity over a design space, the need for surrogate models arises. They are easy to evaluate and approximate quantities of interest of a computer code. Only a few number of evaluations of the simulation are stored for determining the behavior of the response over a whole range of the input parameter domain.

The Kriging method is capable of interpolating highly nonlinear, deterministic functions based on scattered datasets. Using correlation functions, distinct sensitivities of the response with respect to the input parameters can be considered automatically. Kriging can be extended to incorporate not only evaluations of the simulation, but also gradient information, which is called gradient-enhanced Kriging. Adaptive sampling strategies can generate more efficient surrogate models. Contrary to traditional one-stage approaches, the surrogate model is built step-by-step. In every stage of an adaptive process, the current surrogate is assessed in order to determine new sample locations, where the response is evaluated and the new samples are added to the existing set of samples. In this way, the sampling strategy learns about the behavior of the response and a problem-specific design is generated. Critical regions of the input parameter space are identified automatically and sampled more densely for reproducing the response's behavior correctly. The number of required expensive simulations is decreased considerably.

All these approaches treat the response itself more or less as an unknown output of a black-box. A new approach is motivated by the assumption that for a predefined problem class, the behavior of the response is not arbitrary, but rather related to other instances of the mutual problem class. In CFD, for example, responses of aerodynamic coefficients share structural similarities for different airfoil geometries. The goal is to identify the similarities in a database of responses via principal component analysis and to use them for a generic surrogate model. Characteristic structures of the problem class can be used for increasing the approximation quality in new test cases. Traditional approaches still require a large number of response evaluations, in order to achieve a globally high approximation quality. Validating the generic surrogate model for industrial relevant test cases shows that they generate efficient surrogates, which are more accurate than common interpolations. Thus practical, i.e. affordable surrogates are possible already for moderate sample sizes.

So far, interpolation problems were regarded as separate problems. The new approach uses the structural similarities of a mutual problem class innovatively for surrogate modeling. Concepts from response surface methods, variable-fidelity modeling, design of experiments, image registration and statistical shape analysis are connected in an interdisciplinary way.

Generic surrogate modeling is not restricted to aerodynamic simulation. It can be applied, whenever expensive simulations can be assigned to a larger problem class, in which structural similarities are expected.

Acknowledgements

First of all, I would like to express my sincerest gratitude to my advisor Prof. Dr. Volker Schulz. Without his excellent guidance and support, this dissertation would not have been possible. He gave me the opportunity to work in a group of inspiring researchers and also to participate in international conferences, and not only thereby helped me to see the bigger picture.

I also would like to thank Prof. Dr. Peter Benner for taking interest in my work and for serving as a second examiner.

This research has been supported by the German Federal Ministry of Economics and Technology within the collaborative project Computational Flight Testing (ComFliTe). Especially to the colleagues at the German Aerospace Center in Braunschweig, Dr. Stefan Görtz, Dr. Zhong-Hua Han and Dr. Ralf Zimmermann, I am grateful for the fruitful discussions during numerous project meetings.

To my colleagues at the Department of Mathematics of the University of Trier I am obliged for providing a pleasant and constructive atmosphere for conducting my research. For mathematical discussions, helping hands with implementation issues, proofreading and, last but not least, the less scientific conversations during coffee and lunch breaks, I would like to acknowledge Ulf Friedrich, Bastian Groß, Dr. Timo Hylla, Dr. Christina Jager, Dr. Christoph Käbe, Dr. Nils Langenberg, Dr. Andre Lörx, Dr. Claudia Schillings, Dr. Stephan Schmidt, Marina Schneider, Dr. Matthias Schu, Martin Siebenborn, Roland Stoffel, Dirk Thomas, Dr. Christian Wagner, Matthias Wagner, Dr. Xuancan Ye and Heinz Zorn.

I am deeply indebted to my family, especially to my mother Jutta, for their support during my studies. My special gratitude goes to Julia for her constant encouragement.

Benjamin Rosenbaum
Trier, 2013

Contents

1 Introduction

Multidisciplinary simulation and optimization have become key technologies in research and engineering. Advances in mathematics and in computing power during the last decades have allowed the numerical solution of increasingly complex systems. Computer experiments simulate for instance physical models, which are often governed by partial differential equations (PDEs). They can model input-output relationships, where some input parameters determine the response of a system. This can be for example the crashworthiness of a construction element depending on design parameters like its diameter, the output of a chemical system depending on conditions like the temperature, or aerodynamic coefficients of a wing depending on flight conditions like the velocity.

Realistic models are generally challenging and computationally intensive. Still, single simulations are manageable with an appropriate computational and human effort. But when the global behavior of an input-output relationship is sought, dense evaluations of a computer experiment over the whole input parameter space are out of reach. With the demand for numerous evaluations of the response, the need for surrogate models arises. Surrogate models, also known as metamodels or response surface models, approximate the response of a computer experiment based on the information contained in a set of previous computations for different input parameter configurations. They are significantly cheaper to evaluate than the original response and provide insight into the behavior of the system. Usually, an analytical closed-form representation is obtained, which enables the fast computation of approximate outputs for arbitrary inputs. Surrogate models are used for numerous purposes, such as global approximation, optimization, target contour approximation, visualization, uncertainty quantification, sensitivity analysis or statistical hypothesis testing. Methods for selecting samples in the input parameter space, for which the true response is evaluated, are called design of experiments (DOE). The choice of an appropriate surrogate model as well as a carefully determined design are crucial and have to be customized to the research purpose. Also, different types of outputs call for different surrogates and designs: the response can be stochastic or deterministic, erratic or smooth, linear or highly nonlinear, discrete or continuous.

This thesis is devoted to the global approximation of deterministic computer experiments. The output is assumed to be at least continuous with respect to the inputs, but the relationship may be highly nonlinear. Besides these minimal assumptions, nothing is known about the response up front and it is modeled as a black-box system. Two main objectives can be distinguished in global approximation methods: accuracy and efficiency. In order to describe the global behavior of a response correctly by the means of a surrogate model, generally a large number of expensive sample evaluations is required. Response surface methods typically suffer from the curse of dimensionality, i.e. the number of samples for covering the input parameter domain sufficiently dense grows exponentially with the dimension of the input space. For the sake of efficiency, surrogate models as well as design strategies have to

be developed, which use as few samples as possible without sacrificing accuracy.

Adaptive sampling strategies are state-of-the-art for approximating highly nonlinear and multimodal responses. The surrogate model is built sequentially. New samples for evaluating the original response are determined in each step of an adaptive process, thus incorporating information about the global behavior into the choice of the design. Error indicators like cross-validation-based methods can automatically identify critical regions, where a refinement of the design is required. This leads to an intelligent and customized distribution of the samples in the input parameter domain and prevents wasting expensive evaluations in areas where no additional information is needed.

Further enhancement of the approximation quality can be achieved by incorporating a secondary response into the model, which contains information about the primary one. If such auxiliary information is available at comparatively low costs, it can be used to predict the global behavior of the true response efficiently. The additional information can for instance consist of gradients obtained by adjoint computations, low-fidelity data by evaluating a simpler model, by using a coarser mesh for the numerical solution of a PDE, or even by using not fully converged solutions. Variable-fidelity modeling (VFM) is an active field of research, containing a large variety of methods.

The outline of this work is as follows. In **Chapter 2**, several surrogate modeling techniques are introduced. Starting with linear regression, which is a basis for understanding further methods, also radial basis functions are discussed briefly. The chapter's focus is on the presentation of the Kriging and the gradient-enhanced Kriging method. Kriging is capable of interpolating highly nonlinear, deterministic black-box functions by modeling correlations, therefore also providing basic statistical information about the response and an elementary error estimator. The gradient-enhanced version additionally uses derivative information for improving the approximation quality. A detailed investigation of both methods is given and their characteristic properties are deduced profoundly. An emphasis is put on suitable correlation models and on hyperparameter estimation, illustrated with examples for analytical test cases.

Chapter 3 is devoted to the study of design of experiments strategies. A broad overview on design concepts for globally accurate surrogates is presented. One-stage approaches as well as adaptive strategies are covered. They are discussed especially regarding the concepts of exploration and exploitation. In global surrogate modeling, exploration is understood as placing the samples such that they spread evenly in the input parameter domain, leaving no part of it uncovered. Exploitation on the other hand means in this context that samples are chosen where a currently predicted approximation error is highest. One-stage approaches determine all samples prior to evaluating the response, therefore they do not have access to any information on its true behavior. They are merely exploration methods aiming at space-filling designs. Since adaptive methods generate both the surrogate and the design sequentially, intermediate surrogate models can be used for rough error indicators and therefore to introduce an exploitation component. Three selected adaptive strategies are elaborated in detail and their functioning is illustrated with analytical examples.

The design strategies are validated for aerodynamical test cases using Kriging and gradient-enhanced Kriging in **Chapter 4**. The computational fluid dynamics (CFD) codes represent expensive black-box functions, yielding highly nonlinear and multimodal responses. A dense set of validation data allows the assessment of the global error performance. One-stage and

adaptive approaches are compared and it is shown that mixed exploration & exploitation strategies outperform pure exploration methods regarding global accuracy. They reduce the number of expensive response evaluations required for a target accuracy and use the samples efficiently. A comparison of the different distributions of the samples in the input parameter space is given to back up the results. Additionally, the adaptive strategies are compared to two theoretical sampling strategies, which make use of the validation data and can be regarded as a best-case study. Finally, an extension of adaptive sampling strategies for computer experiments with multiple responses is considered, providing an efficient framework for the investigated test cases.

In **Chapter 5**, a new framework called generic surrogate modeling is introduced, which represents the centerpiece of this work. The thorough derivation of this method includes a detailed mathematical presentation of the involved concepts. Often, a computer experiment is not treated as a single instance. Instead, it belongs to a larger problem class. This approach is driven by the assumption that for such a predefined problem class the behavior of a response is not arbitrary, but rather related to other instances of the mutual class. Motivated by the concept of statistical shape models, in a first step the most important features of a training database of responses are extracted by principal component analysis. Then, each new test case of the problem class can be approximated by a generic surrogate model, using only few evaluations of the response. Instead of directly interpolating the samples, first the principal components of the training database are fitted to the new data in a least squares sense. Based on this model, interpolation of the data is performed in a variable-fidelity modeling framework. This new approach aims at a globally accurate surrogate, which requires less samples than traditional methods. The global, generally highly nonlinear trend is captured by the information contained in the training database and not by massive evaluation of the response. It can be seen as a preprocessing, transferring online costs to offline costs. Less of the expensive evaluations, which dominate the online costs, are needed, while generating the database is seen as offline costs, since it can be stored and used for numerous new test cases.

A numerical validation of generic surrogate modeling is presented in **Chapter 6**. Aerodynamic coefficients depending on flight conditions have already been investigated in the previous chapters. Now it is demonstrated that a training database of responses of different airfoil geometries constitutes a mutual problem class. These responses feature strong structural similarities over the whole the input parameter domain and only few principal components are required to represent the whole problem class adequately. In three distinct test cases, i.e. for responses of three airfoil geometries, the generic surrogate models are compared to Kriging and gradient-enhanced Kriging using one-stage and also adaptive designs. It is shown that the proposed framework is a successful preprocessing for the interpolation. Kriging and, depending on the test case, even gradient-enhanced Kriging are outperformed regarding global accuracy and already a small number of samples is sufficient to reproduce the global trend of the responses. An error indicator is introduced and validated, which makes the new approach a reliable tool for global surrogate modeling.

This thesis concludes with a summary of the main results and an outlook to some open work in **Chapter 7**.

Unless stated otherwise, all numerical results in this thesis were produced by a C-code on a desktop computer with a 3.0 GHz Intel Core 2 Duo processor and 4 GB RAM. The methods were implemented using the *Surrogate-Modeling for Aero-Data Toolbox* (SMART) developed at DLR.

2 Surrogate modeling

Surrogate models approximate responses of a computer experiment. The evaluation of the computer experiment is assumed to be expensive and a surrogate model is a much simpler model which can be evaluated cheaply. In the literature, surrogate models are also referred to as approximation models, metamodels, response surface models, supervised learning or emulators (whereas the computer experiment is called simulator).

The evaluation of the computer experiment yields a scalar response $y \in \mathbb{R}$, which depends on input parameters $x = (x_1, \ldots, x_d)^\top \in \Omega \subset \mathbb{R}^d$. It is assumed to be a deterministic black-box system, i.e. nothing is known about the global behavior of the response up front. Also, no random error is included and multiple evaluations with the same input parameters x will always produce the same response $y(x)$. Mathematically, the response is an unknown function

$$\begin{aligned} y : \Omega &\to \mathbb{R}, \\ x &\mapsto y(x). \end{aligned} \tag{2.1}$$

For this input-output relationship a surrogate $\widehat{y} : \Omega \to \mathbb{R}$ is sought on an input parameter domain $\Omega \subset \mathbb{R}^d$. The dimension of the input space is usually small ($d < 10$). In this thesis, only scalar responses are discussed. For multivariate responses $y \in \mathbb{R}^{\tilde{d}}$ ($\tilde{d} > 1$) it is assumed that a separate surrogate model is used for each dimension of the output parameter space.

The surrogate model is constructed from a set of evaluations of the computer experiment

$$\left\{(x^{(i)}, y_i)\right\}_{i=1}^{N}, \quad y_i = y(x^{(i)}) \quad (i = 1, \ldots, N). \tag{2.2}$$

The inputs $X := \{x^{(1)}, \ldots, x^{(N)}\} \subset \Omega$ are called samples and $Y := (y_1, \ldots, y_N)^\top \in \mathbb{R}^N$ are called observations. In the following, superscripts are used to denote samples, while subscripts are used to denote components of a vector, so $x_k^{(i)} \in \mathbb{R}$ is the k-th component of $x^{(i)} \in \mathbb{R}^d$. Because the scalar observations y_i, depending on the samples $x^{(i)}$ ($i = 1, \ldots, N$), are assembled in the vector $Y \in \mathbb{R}^N$, which is made use of frequently, subscripts will be used for them as well for convenience. It is assumed that $x^{(i)} \neq x^{(j)}$ ($\forall i \neq j$), since $y(x)$ is a deterministic function and identical data pairs $(x^{(i)}, y_i) = (x^{(j)}, y_j)$ do not add any supplementary information about $y(x)$. The choice of $X \subset \Omega$, also referred to as sampling, will be the subject of Chap. 3, where several designs of experiments (DOE) are discussed.

The aim of this thesis is to present methods for global approximation, in other words to construct surrogates $\widehat{y}(x)$ which are globally valid for $y(x)$ in $\Omega \subset \mathbb{R}^d$. Other methods like surrogate-based optimization methods on the contrary do not primarily aim at a good approximation quality in the whole domain Ω. The choice of the surrogate model and the design of experiments can strongly differ from global approximation methods. Optimization methods will not be the subject of this thesis. In surrogate modeling for global approximation

of expensive computer experiments, two main objectives can be distinguished: accuracy and efficiency. Clearly, $\widehat{y}(x)$ is supposed to be a good approximation to $y(x)$ for every $x \in \Omega$. But for the sake of efficiency, one is interested to keep the number of evaluations N as small as possible. Numerous evaluations that densely cover the input parameter domain Ω in order to precisely describe the response's behavior are computationally out of reach. Although with an increase of N also the computational complexity of model generation is intensified, the cost of constructing a surrogate from given data is negligible. The cost of one single evaluation of the true response is assumed to exceed the cost of model building.

The main purpose of this chapter is the introduction of the Kriging method in Sect. 2.3, which will be used for surrogate modeling throughout this thesis. An extension of this method is gradient-enhanced Kriging (GEK), which is able to include sensitivity information and is presented in Sect. 2.4. Two other techniques, linear regression and radial basis functions, are reviewed in Sects. 2.1 and 2.2. Their discussion offers a better insight into the functioning of the Kriging method.

2.1 Linear regression models

Linear regression models provide an easy-to-use method of constructing a surrogate. Their use dates back over 200 years to Carl Friedrich Gauß and their statistical properties are well-known. They can be applied in surrogate modeling, forecasting, uncertainty quantification and hypothesis testing. Detailed introductions can be found in [8, 101, 42, 28].

2.1.1 Ordinary least squares

In linear regression models, the observed response values $Y = (y_1, \ldots, y_N)^\top$ (2.2) are assumed to satisfy the equations

$$y_i = \sum_{j=1}^{K} \beta_j f_j(x^{(i)}) + z_i \quad (i = 1, \ldots, N). \tag{2.3}$$

$f_j : \mathbb{R}^d \to \mathbb{R}$ are known functions with yet unknown coefficients β_j $(j = 1, \ldots, K)$. z_i are normally distributed random variables, which are uncorrelated and have zero expectation and constant variance σ^2

$$\mathrm{E}[z_i] = 0, \quad \mathrm{Var}[z_i] = \sigma^2 \quad (i = 1, \ldots, N), \tag{2.4}$$

i.e. $z_i \sim \mathcal{N}(0, \sigma^2)$. The most popular choice for the functions f_j are polynomials, but the model is not restricted to them. The surrogate model is defined by

$$\begin{aligned} \widehat{y}(x) :=& \sum_{j=1}^{K} \beta_j f_j(x) \\ =& f(x)^\top \beta, \end{aligned} \tag{2.5}$$

using $\beta := (\beta_1, \ldots, \beta_K)^\top \in \mathbb{R}^K$ and $f(x) := (f_1(x), \ldots, f_K(x))^\top$, $f : \mathbb{R}^d \to \mathbb{R}^K$.

To be able to determine the coefficients β_j from the observed data Y, at least as many observations as unknowns are needed ($N \geq K$). A design matrix

$$F := \left[f_j(x^{(i)}) \right]_{i,j=1}^{N,K} \in \mathbb{R}^{N \times K} \tag{2.6}$$

and a sum of squared differences (or errors)

$$\begin{aligned} \mathrm{SSD}(\beta) :=& \sum_{i=1}^{N} \left(y_i - f(x^{(i)})^\top \beta \right)^2 \\ =& (Y - F\beta)^\top (Y - F\beta) \\ =& \|Y - F\beta\|_2^2 \end{aligned} \tag{2.7}$$

are defined. β is determined by minimizing the SSD, so the surrogate model (2.5) is a linear least squares fit to the observations Y.

Theorem 2.1. *If F has full column rank K, the linear least squares problem*

$$\min_{\beta \in \mathbb{R}^K} \ \mathrm{SSD}(\beta) \tag{2.8}$$

has a unique solution

$$\beta = \left(F^\top F \right)^{-1} F^\top Y. \tag{2.9}$$

Proof. The gradient and the Hessian of $\mathrm{SSD}(\beta)$ are

$$\nabla \mathrm{SSD}(\beta) = 2 \left(F^\top F \beta - F^\top Y \right), \tag{2.10}$$

$$\nabla^2 \mathrm{SSD}(\beta) = 2 \ F^\top F. \tag{2.11}$$

$F^\top F$ is a positive definite symmetric matrix: Otherwise there existed a $\beta \neq 0$ satisfying $0 \geq \beta^\top F^\top F \beta = \|F\beta\|_2^2$, which is only possible for $F\beta = 0$. Because of the full column rank of F this leads to $\beta = 0$, which is a contradiction. Necessary first order optimality conditions are applied and $\nabla \mathrm{SSD}(\beta) = 0$ is equivalent to $\beta = (F^\top F)^{-1} F^\top Y$. The Hessian is positive definite, so the second order sufficient condition is also fulfilled. □

The condition of F having full column rank K is crucial. Linear independent functions $\{f_1, \ldots, f_K\}$ (such as monomials) are only a necessary, but not sufficient condition. E.g. due to a degenerated sampling $X = \{x^{(1)}, \ldots, x^{(N)}\}$, the rank of F can be smaller than K. In the remainder of this thesis, linear independent functions and a full column rank of F are assumed. For a discussion of problems with the rank of F see [8, 28] and for numerical methods see [6].

Another interpretation of (2.9) is given. The matrix $F^+ := (F^\top F)^{-1} F^\top \in \mathbb{R}^{K \times N}$ is the Moore-Penrose pseudo-inverse of F and satisfies $F^+ F = I_K$, see [41]. Generally, the system $F\beta = Y$ is overdetermined, because $N > K$ and a solution does not exist. $\beta = F^+ Y$ then minimizes the sum of squared errors $\|F\beta - Y\|_2^2$ among all $\beta \in \mathbb{R}^K$. Note that the statistical assumption about the errors z_i (2.4) is not needed for the construction of the ordinary least

squares estimator $\widehat{y}(x)$ (2.5). But it is associated with the choice of the norm $\|\cdot\|_2$, which was used to measure the error. The next section will illustrate how other assumptions about the errors z_i can lead to other norms.

2.1.2 Generalized least squares

Analogously to the previous section, the generalized least squares model is defined by

$$y_i = \sum_{j=1}^{K} \beta_j f_j(x^{(i)}) + z_i \quad (i = 1, \ldots, N) \tag{2.12}$$

with the f_j as in (2.3). Suppose now that the z_i are random variables with zero expectation, but the assumption of uncorrelatedness or constant variance no longer holds. $Z := (z_1, \ldots, z_N)^\top$ is a multivariate normally distributed random vector satisfying

$$\mathrm{E}[Z] = 0 \in \mathbb{R}^N, \quad \mathrm{Cov}[Z, Z] = \Sigma \in \mathbb{R}^{N \times N}, \tag{2.13}$$

i.e. $Z \sim \mathcal{N}_N(0, \Sigma)$, where Σ is an arbitrary symmetric positive definite covariance matrix. This general setting includes cases like $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_N^2)$ (uncorrelated but different variances, "heteroscedasticity"), $\Sigma = \sigma^2 R$ with a correlation matrix R (constant variance but correlated) or even the previous section's setting $\Sigma = \sigma^2 I$. For the time being suppose Σ to be known. The surrogate model is again defined by

$$\widehat{y}(x) := f(x)^\top \beta. \tag{2.14}$$

Under the assumptions (2.12)–(2.13), $Y \sim \mathcal{N}_N(F\beta, \Sigma)$ is a multivariate normally distributed vector. Maximum likelihood estimation is applied for determining β in (2.14), see [8]. For a given vector of observations Y the likelihood function of β is

$$L(\beta|Y) = (2\pi)^{-\frac{N}{2}} (\det \Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(Y - F\beta)^\top \Sigma^{-1} (Y - F\beta)\right). \tag{2.15}$$

It will now be shown that for the general case (2.13) the ordinary linear least squares fit is not an appropriate surrogate model.

Theorem 2.2. *If F has full column rank K and the covariance matrix Σ is nonsingular, the maximum likelihood problem*

$$\max_{\beta \in \mathbb{R}^K} L(\beta|Y) \tag{2.16}$$

has a unique solution

$$\beta = \left(F^\top \Sigma^{-1} F\right)^{-1} F^\top \Sigma^{-1} Y. \tag{2.17}$$

Proof. Problem (2.16) is equivalent to minimizing the negative logarithm of the likelihood function $-\log L(\beta|Y)$

$$\min_{\beta \in \mathbb{R}^K} \left\{ \frac{N}{2} \log(2\pi) + \frac{1}{2} \log(\det \Sigma) + \frac{1}{2} (Y - F\beta)^\top \Sigma^{-1} (Y - F\beta) \right\}. \tag{2.18}$$

The gradient and the Hessian of $-\log L(\beta|Y)$ with respect to β are

$$\nabla\left(-\log L(\beta|Y)\right) = F^\top\Sigma^{-1}F\beta - F^\top\Sigma^{-1}Y, \tag{2.19}$$

$$\nabla^2\left(-\log L(\beta|Y)\right) = F^\top\Sigma^{-1}F. \tag{2.20}$$

$F^\top\Sigma^{-1}F$ is a positive definite symmetric matrix: Otherwise there existed a $\beta \neq 0$ satisfying $\beta^\top F^\top\Sigma^{-1}F\beta \leq 0$. Since Σ is positive definite and symmetric, it has a Cholesky decomposition $\Sigma = LL^\top$, where L is a lower triangular matrix with strictly positive diagonal entries. Using this in the inequality yields $0 \geq \beta^\top F^\top L^{-\top}L^{-1}F\beta = \|F\beta\|^2_{L^{-1}}$, where $\|\cdot\|_{L^{-1}} = \left\|L^{-1}\cdot\right\|_2$. This is only possible for $F\beta = 0$ and because of the full column rank of F this leads to $\beta = 0$, which is a contradiction. Now first and second order optimality conditions can be applied: $\nabla\left(-\log L(\beta|Y)\right) = 0$ is equivalent to $\beta = (F^\top\Sigma^{-1}F)^{-1}F^\top\Sigma^{-1}Y$ and the Hessian is positive definite, from which the claim follows. □

While the ordinary least squares method estimated β by minimizing the $\|\cdot\|_2$-norm of the errors $Z = Y - F^\top\beta$, the generalized least squares estimation for β minimizes the $\|\cdot\|_{L^{-1}}$-norm of Z

$$\|Y - F\beta\|^2_{L^{-1}} = (Y - F\beta)^\top\,\Sigma^{-1}\,(Y - F\beta)\,, \tag{2.21}$$

where $LL^\top = \Sigma$. (2.21) is also called Mahalanobis distance of Y and $F\beta$, cf. [4]. This formulation illustrates how ordinary least squares is a special case of generalized least squares. If the assumption $Z \sim \mathcal{N}_N(0, \sigma^2 I)$ holds, minimizing (2.21) is equivalent to minimizing the ordinary SSD(β) (2.7) and β in (2.17) is the same as in (2.9), as σ^2 is canceled out. Another instance is $Z \sim \mathcal{N}_N(0, \operatorname{diag}(\sigma_1^2, \ldots, \sigma_N^2))$, where the z_i are uncorrelated but have different variances σ_i^2 ("heteroscedasticity"). In this case, (2.17) is the solution of a weighted least squares problem $\min_{\beta\in\mathbb{R}^K}\sum_{i=1}^N \sigma_i^{-1}(y_i - f(x^{(i)})^\top\beta)^2$.

2.1.3 Discussion

Linear regression models are now briefly reviewed for their appropriateness in global approximations of expensive deterministic computer experiments. They are applied in many different fields of research, model generation is simple and computational costs are low. They are capable of dealing with a large number of samples, as the cost of evaluating the surrogate $\widehat{y}(x)$ is just $\mathcal{O}(N)$ and the cost for determining β is dominated by $\mathcal{O}(NK^2)$ for ordinary least squares (computation of $F^\top F$) and by $\mathcal{O}(N^3)$ for generalized least squares (inversion of Σ). But besides the fact that large sample sizes N are not discussed within the scope of this thesis due to the huge computational cost of one single evaluation of the response $y(x)$, some points which limit the capability of regression models deserve attention.

- For $K < N$, regression models are an approximation to the observed data Y and not an interpolation. Supposing $y(x)$ is deterministic and continuous in x, a surrogate model should be accurate at least in the neighborhood of any sample $x^{(i)} \in X$, which can be achieved by interpolation rather than approximation.

- In deterministic computer experiments, no noise is included in the evaluation of $y(x)$. Thus, the surrogate model's deviation $\widehat{y}(x) - y(x)$ from the true response has to be

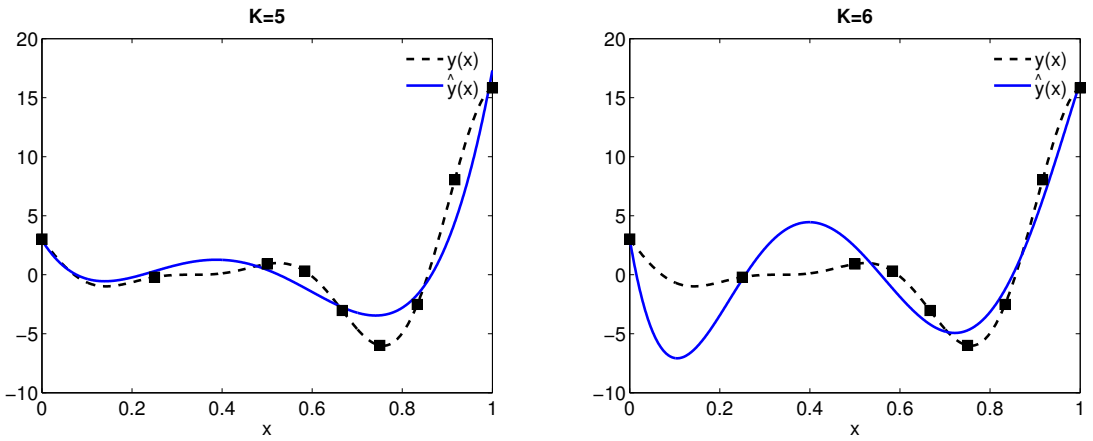

Figure 2.1: Ordinary least squares approximation $\widehat{y}(x) = \sum_{j=1}^{K} \beta_j x^{j-1}$ for $K = 5$ (underfitted) and $K = 6$ (overfitted), $y(x) = (6x - 2)^2 \sin(12x - 4)$.

seen as a systematic deviation rather than a random error and statistics based on the random errors z_i become meaningless or even impossible to compute, cf. [122, 85].

- In fact, regression models are only easy to use if the set of functions $\{f_1, \ldots, f_K\}$ is already known. Since the surrogate model is solely defined by $f(x)^\top \beta$, ideally the user would already have a strong hint about the behavior of the true response $y(x)$ and therefore about the choice of f, which would be contradictory to the concept of a black-box model. Selecting a set of suitable functions from a larger admissible set is a nontrivial task and just choosing the whole admissible set might not produce the best result, as explained in the following argument.

- The authors of [41] pointed out that polynomial models are not suitable for modeling the highly nonlinear responses, which often arise in engineering problems. To capture a highly nonlinear behavior, many degrees of freedom K of the surrogate model are needed. This often is related to the problem of underfitting vs. overfitting. Choosing K too small results in a poor fit in terms of the SSD (2.7)/(2.21), while choosing K too large reduces the SSD, but can produce unfavorable variations in undersampled regions of $\Omega \setminus X$. The problem is highlighted with a one-dimensional example in Fig. 2.1. The ordinary least squares approximation $\widehat{y}(x) = \sum_{j=1}^{K} \beta_j x^{j-1}$ to the $N = 9$ data points is a poor approximation for $K = 5$, e.g. in the region around the global minimum. Increasing the degrees of freedom by one to $K = 6$ reduces the SSD by about 50% from 25.0 to 12.1, but doubles the total average error $\int_0^1 |y(x) - \widehat{y}(x)| dx$ from 1.1 to 2.3.

2.2 Radial basis functions

Before the Kriging method is introduced in the next section, radial basis functions (RBF) are presented briefly. While the last section covered the topic of approximation of functions

with a known or suspected structure, radial basis functions offer a method of interpolating scattered data without prior information about the true response $y(x)$. They are also used in a broad field of applications like multivariate interpolation, image registration or meshfree solution of PDEs. General introductions as well as theoretical discussions can be found in [106, 9, 62, 35]. They can also be interpreted as a single-layer artificial neural network (ANN), cf. [41].

The surrogate

$$\widehat{y}(x) := \sum_{i=1}^{N} w_i R(\|x - x^{(i)}\|) \tag{2.22}$$

is defined by a weighted sum of radially symmetric functions

$$\begin{aligned} R &: \mathbb{R}^d \to \mathbb{R}, \\ x - x^{(i)} &\mapsto R(\|x - x^{(i)}\|) \end{aligned} \tag{2.23}$$

with $w := (w_1, \ldots, w_N)^\top \in \mathbb{R}^N$. The centers of these functions coincide with the samples $x^{(i)}$ $(i = 1, \ldots, N)$, but generally other centers $c^{(i)} \neq x^{(i)}$ are possible. Usually, the Euclidean distance $\|\cdot\|_2$ is used. Similarly to the previous section, the surrogate is a linear combination of basis functions, but instead of defining an optimization problem (2.8)/(2.18), an interpolation condition

$$\widehat{y}(x^{(i)}) = y_i \quad (i = 1, \ldots, N) \tag{2.24}$$

is defined which determines the unknowns w_i. The condition is equivalent to

$$Rw = Y, \tag{2.25}$$

where $R := [R(\|x^{(j)} - x^{(i)}\|)]_{i,j=1}^{N,N} \in \mathbb{R}^{N \times N}$ is assumed to be nonsingular. Clearly, R becomes singular for any $x^{(i)} = x^{(j)}$ $(i \neq j)$ due to equal lines and columns, so this case is excluded by default. Conditions under which a radial basis function always generates a nonsingular system matrix R and examples of such functions can e.g. be found in [9, 62, 147, 35], see also Sect. 2.3.3.

Gaussian	e^{-h^2}
inverse multiquadric	$(1 + h^2)^{-\frac{1}{2}}$
multiquadric	$(1 + h^2)^{\frac{1}{2}}$
polyharmonic splines ($k \in \mathbb{N}$)	h^{2k-1} $h^{2k} \log h$

Table 2.1: Radial basis functions, $h := \|x - x^{(i)}\|$

Popular choices for radial basis functions are listed in Table 2.1. They can be subdivided in local and global basis functions. Gaussian and inverse multiquadric basis functions are local in the sense of $R(h) \to 0$ $(h \to \infty)$, while multiquadric basis functions and polyharmonic splines are global, i.e. $R(h) \to \infty$ $(h \to \infty)$. This is not to be mistaken for global and local surrogate models, as also a surrogate model based on local basis functions can be a globally

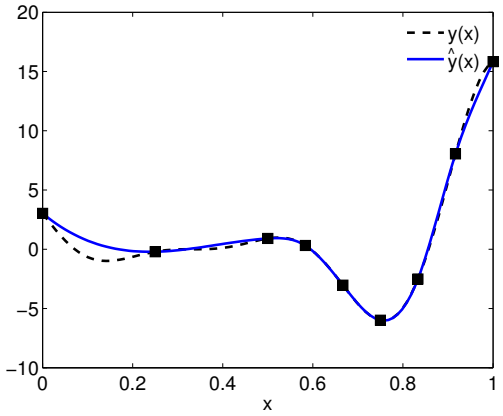

Figure 2.2: RBF interpolation $\widehat{y}(x)$ using $R(h) = |h|^3$, $y(x) = (6x-2)^2 \sin(12x-4)$.

valid one. Figure 2.2 depicts an RBF interpolation of the data from the previous section's example (Fig. 2.1) with polyharmonic splines $R(|x - x^{(i)}|) = |x - x^{(i)}|^3$. Contrary to the ordinary least squares approximation, the RBF interpolation is able to capture the nonlinear behavior of the response accurately.

There exist numerous enhancements of the basic method. h can be replaced by θh, where $\theta > 0$ is a scaling factor, adding more freedom to the model. An additional regression term $f(x)^\top \beta$, e.g. polynomials of low order, can be included in the model (2.22), if a global trend of $y(x)$ is assumed. Finally, radial basis functions can also be used for approximation, either by using less basis functions than samples and solving (2.25) only in a least squares sense, or by regularizing R by $R + \epsilon I$, which can be used if the data is noisy. The surrogate then acts as a smoothing approximation of Y instead of an interpolation.

A short presentation of radial basis functions was given. They are capable of interpolating arbitrary observations Y of pairwise disjunct samples X without any statistical assumptions. However, the fact that the basis functions are assumed radial may be inappropriate for the use in surrogate modeling for computer experiments, because the response $y(x)$ may have a different sensitivity with respect to each input parameter x_k $(k = 1, \ldots, d)$. This problem can be avoided by using a different norm instead of the Euclidean distance $\|\cdot\|_2$. In Kriging, these sensitivities can be obtained automatically from the data. It will also be shown that, under certain assumptions, interpolations with local radial basis functions are a special case of the Kriging method.

2.3 Kriging

The Kriging method is named after the mining engineer Daniel Krige, who used this technique first in the 1950s for the prediction of ore concentration based on core samples [74], and was further developed by Matheron [94]. Since then it has been widely applied in spatial prediction and geostatistics, see [21] for an overview. The method can interpolate scattered data by modeling spatial correlations. Nowadays, Kriging is a topic in many textbooks on statistics for spatial data [61, 22, 146, 126, 15]. Since the 1980s it has also been applied in

the interpolation of responses generated by deterministic computer experiments, introduced by Sacks et al. [122], see also [26], and [73] for a more detailed discussion. A thorough mathematical investigation can be found in the monograph [134]. It gained popularity in surrogate modeling for expensive simulations and engineering problems and since the 2000s it is discussed in textbooks in this field of research [123, 33, 112, 70, 41]. Recently, a review on Kriging for computer experiments was given in [83].

2.3.1 Model assumptions

The main idea for a Kriging prediction $\widehat{y}(x)$ for $y(x)$ is to use a weighted sum of the previous evaluations $Y = (y_1, \ldots, y_N)^\top$

$$\widehat{y}(x) = \sum_{i=1}^{N} \lambda_i(x) y_i \tag{2.26}$$

with weights $\lambda(x) := (\lambda_1(x), \ldots, \lambda_N(x))^\top \in \mathbb{R}^N$ depending on x. Intuitively, the influence of y_i on $\widehat{y}(x)$ should be large if x is close to $x^{(i)}$. But instead of using a rather simple approach like inverse distance weighting, the Kriging method makes use of the underlying statistical assumptions on the unknown response $y(x)$. It is treated as a sum of a linear regression part and a lack-of-fit term

$$y(x) = \sum_{j=1}^{K} \beta_j f_j(x) + z(x), \quad x \in \Omega \subset \mathbb{R}^d, \tag{2.27}$$

where $f_j(x) : \mathbb{R}^d \to \mathbb{R}$ are known functions with coefficients $\beta_j \in \mathbb{R}$ $(j = 1, \ldots, K)$ similarly to (2.3) and (2.12). But unlike linear regression approximation in Sect. 2.1, the error $z(x) : \mathbb{R}^d \to \mathbb{R}$ is not assumed to be a random error or white noise. It is rather a systematic deviation from the linear model and captures the nonlinear behavior of $y(x)$, which is not described by $\sum_{j=1}^{K} \beta_j f_j(x)$. $z(x)$ is modeled as the realization of a Gaussian stochastic process, i.e. any vector of a finite number of evaluations $Z := (z(x^{(1)}), \ldots, z(x^{(N)}))^\top \in \mathbb{R}^N$ is multivariate normally distributed. Furthermore, the Gaussian process is assumed to be second-order stationary, i.e. it has a constant mean function and a translational symmetric autocovariance function, additionally satisfying

$$\mathrm{E}\left[z(x)\right] = 0, \quad \mathrm{Cov}\left[z(x), z(\tilde{x})\right] = \sigma^2 R(\tilde{x} - x) \quad \forall x, \tilde{x} \in \Omega. \tag{2.28}$$

$\sigma^2 > 0$ is the process variance and $\mathrm{Corr}[z(x), z(\tilde{x})] := R(\tilde{x} - x)$ is a spatial autocorrelation function. Any valid autocorrelation function must satisfy the conditions of symmetry and nonnegative definiteness

$$R(\tilde{x} - x) = R(x - \tilde{x}) \quad \forall x, \tilde{x} \in \Omega, \tag{2.29}$$

$$\sum_{i,j=1}^{N} w_i w_j R(x^{(j)} - x^{(i)}) \geq 0 \quad \forall \left\{x^{(1)}, \ldots, x^{(N)}\right\} \subset \Omega,\ w \in \mathbb{R}^N. \tag{2.30}$$

Additionally, for the stationarity argument, the correlation function only depends on the distance vector $\tilde{x} - x$ and not on the locations x and $\tilde{x}$ themselves. Assuming that $z(x)$

and $z(\tilde{x})$ are strongly correlated when x and $\tilde{x}$ are close and just weakly correlated for a large distance $\|\tilde{x} - x\|$, usually a correlation function is chosen that satisfies $R(0) = 1$ and $R(\tilde{x} - x) \longrightarrow 0$ for $\|\tilde{x} - x\| \longrightarrow \infty$ for arbitrary norms $\|\cdot\|$ and therefore acts merely locally. For the theory of stochastic processes the reader is referred to e.g. [109] and for Gaussian process modeling to [112].

With $z(x)$ being a stationary Gaussian process, $y(x)$ also has the property of a Gaussian process (generally nonstationary due to the non-constant mean) with

$$\mathrm{E}\left[y(x)\right] = f(x)^\top \beta, \quad \mathrm{Cov}\left[y(x), y(\tilde{x})\right] = \sigma^2 R(\tilde{x} - x) \quad \forall x, \tilde{x} \in \Omega, \tag{2.31}$$

using $\beta := (\beta_1, \ldots, \beta_K)^\top \in \mathbb{R}^K$ and $f(x) := (f_1(x), \ldots, f_K(x))^\top$, $f : \mathbb{R}^d \to \mathbb{R}^K$ as in Sect. 2.1. Hence, Z and $Y = F\beta + Z$ are both multivariate normally distributed

$$Y \sim \mathcal{N}_N(F\beta, \sigma^2 R), \tag{2.32}$$

$$Z \sim \mathcal{N}_N(0, \sigma^2 R), \tag{2.33}$$

where $F := [f_j(x^{(i)})]_{i,j=1}^{N,K} \in \mathbb{R}^{N \times K}$ is the regression matrix and $R := [R(x^{(j)} - x^{(i)})]_{i,j=1}^{N,N} \in \mathbb{R}^{N \times N}$ is the symmetric and positive definite correlation matrix. The symmetry is guaranteed by (2.29), but for the positive definiteness formulation, (2.30) is replaced by

$$\sum_{i,j=1}^{N} w_i w_j R(x^{(j)} - x^{(i)}) > 0 \quad \forall x^{(i)} \neq x^{(j)} \ (i \neq j), \ w \in \mathbb{R}^N \setminus \{0\}, \tag{2.34}$$

such that the existence of R^{-1} is ensured. Possible choices for valid correlation functions and a continued discussion will be presented in Sect. 2.3.3.

2.3.2 The Kriging predictor

For the regression models in Sect. 2.1, the statistical assumptions (2.3) or (2.12) were rather strong, because the surrogate model $\widehat{y}(x)$ was merely defined by the linear model $\sum_{j=1}^{K} \beta_j f_j(x)$. In Kriging however, surrogate modeling is not performed by choosing a sophisticated linear model, but rather by exploiting the capability of the Gaussian process to interpolate arbitrary data. Therefore, Kriging also goes under the name of Gaussian process prediction, see e.g. [112]. Based on the observed data Y, the properties of $z(x)$ are used to derive a best linear unbiased predictor (BLUP) $\widehat{y}(x)$ for the response $y(x)$.

Definition 2.3 (Best linear unbiased prediction). *A prediction $\widehat{y}(x)$ is a BLUP for $y(x)$ based on $Y = (y_1, \ldots, y_N)^\top$ if it is a linear combination of the entries*

$$\widehat{y}(x) = \sum_{i=1}^{N} \lambda_i(x) y_i, \tag{2.35}$$

which minimizes the mean squared error

$$\mathrm{MSE}\left[\widehat{y}(x)\right] := \mathrm{E}\left[\left(\widehat{y}(x) - y(x)\right)^2\right] \tag{2.36}$$

subject to the unbiasedness condition

$$\mathrm{E}\left[\widehat{y}(x) - y(x)\right] = 0 \tag{2.37}$$

for all $x \in \Omega$.

In the following theorem, a necessary and sufficient condition is derived for which the Kriging predictor (2.26) is a BLUP and also an interpolator. Let $r(x) := (R(x^{(i)} - x))_{i=1}^{N} \in \mathbb{R}^N$ identify a correlation vector for an arbitrary $x \in \Omega$ and all $x^{(1)}, \ldots, x^{(N)}$. $\mu(x) \in \mathbb{R}^K$ is a Lagrange variable and R, F, $f(x)$, $\lambda(x)$ are defined as in the previous section.

Theorem 2.4. *Let* $x^{(i)} \neq x^{(j)}$ *for* $i \neq j$ *and* F *have full column rank* K. *Provided* $\binom{\lambda(x)}{\mu(x)} \in \mathbb{R}^{N+K}$ *is the solution of the linear system*

$$\begin{bmatrix} R & F \\ F^\top & 0 \end{bmatrix} \begin{pmatrix} \lambda(x) \\ \mu(x) \end{pmatrix} = \begin{pmatrix} r(x) \\ f(x) \end{pmatrix}, \tag{2.38}$$

the Kriging predictor $\widehat{y}(x) = \lambda(x)^\top Y$ *is a BLUP for* $y(x)$, $x \in \Omega$.

Proof. The proof mainly follows the argumentation of [123]. For an arbitrary $x \in \Omega$, a solution of the constrained optimization problem

$$\min_{\lambda(x) \in \mathbb{R}^N} \ \mathrm{E}\left[\left(\lambda(x)^\top Y - y(x)\right)^2\right] \tag{2.39}$$

$$\text{s. t. } \mathrm{E}\left[\lambda(x)^\top Y - y(x)\right] = 0 \tag{2.40}$$

is sought. Exploiting $\mathrm{E}[z(x)] = 0$, the unbiasedness condition (2.40) can be rewritten as

$$\begin{aligned} 0 &= \mathrm{E}\left[\lambda(x)^\top Y - y(x)\right] \\ &= \mathrm{E}\left[\lambda(x)^\top (F\beta + Z) - f(x)^\top \beta - z(x)\right] \\ &= \lambda(x)^\top F\beta + \lambda(x)^\top \mathrm{E}[Z] - f(x)^\top \beta - \mathrm{E}[z(x)] \\ &= \left(F^\top \lambda(x) - f(x)\right)^\top \beta. \end{aligned} \tag{2.41}$$

This must hold for any β, so the unbiasedness condition holds for

$$F^\top \lambda(x) - f(x) = 0. \tag{2.42}$$

With this identity, the error of the surrogate model can be rewritten as

$$\begin{aligned} \lambda(x)^\top Y - y(x) &= \lambda(x)^\top (F\beta + Z) - f(x)^\top \beta - z(x) \\ &= \left(F^\top \lambda(x) - f(x)\right)^\top \beta + \lambda(x)^\top Z - z(x) \\ &= \lambda(x)^\top Z - z(x). \end{aligned} \tag{2.43}$$

Hence, the MSE$[\widehat{y}(x)]$ reads

$$\begin{aligned}\mathrm{E}\left[\left(\lambda(x)^\top Y - y(x)\right)^2\right] &= \mathrm{E}\left[\left(\lambda(x)^\top Z - z(x)\right)^2\right] \\ &= \mathrm{E}\left[\lambda(x)^\top Z Z^\top \lambda(x) - 2\lambda(x)^\top Z z(x) + z(x)^2\right] \\ &= \sigma^2\left(\lambda(x)^\top R\lambda(x) - 2\lambda(x)^\top r(x) + 1\right). \end{aligned} \tag{2.44}$$

Now, a new optimization problem can be posed using (2.44) and (2.42), whose solution is sufficient for (2.39)–(2.40):

$$\min_{\lambda(x)\in\mathbb{R}^N} \frac{1}{2}\lambda(x)^\top R\lambda(x) - r(x)^\top\lambda(x) \tag{2.45}$$

$$\text{s.t. } F^\top\lambda(x) - f(x) = 0. \tag{2.46}$$

Introducing a Lagrange variable $\mu(x) \in \mathbb{R}^K$, the corresponding Karush-Kuhn-Tucker system (cf. [103]) reads

$$\begin{bmatrix} R & F \\ F^\top & 0 \end{bmatrix}\begin{pmatrix}\lambda(x)\\ \mu(x)\end{pmatrix} = \begin{pmatrix} r(x)\\ f(x)\end{pmatrix}. \tag{2.47}$$

Since R is positive definite (for $x^{(i)} \neq x^{(j)}$, $i \neq j$) and F has full column rank K, the system matrix is nonsingular and (2.47) is a necessary and sufficient optimality condition for the convex quadratic problem with linear constraints, again see [103]. □

From the condition (2.38) not only the BLUP property can be derived, but also the interpolation property. Inserting any $x^{(i)}$ for x ($i = 1, \ldots, N$), the right-hand side $\left(\begin{smallmatrix} r(x^{(i)})\\ f(x^{(i)})\end{smallmatrix}\right)$ becomes the i-th column of the system matrix $\left[\begin{smallmatrix} R & F\\ F^\top & 0\end{smallmatrix}\right]$ and $\left(\begin{smallmatrix}\lambda(x^{(i)})\\ \mu(x^{(i)})\end{smallmatrix}\right) = \left(\begin{smallmatrix} e_i\\ 0\end{smallmatrix}\right)$ is the solution, where e_i denotes the i-th unit vector. Then evaluating the surrogate in $x^{(i)}$

$$\widehat{y}(x^{(i)}) = e_i^\top Y = y_i \tag{2.48}$$

yields the correct response $y_i = y(x^{(i)})$.

Also, the MSE $[\widehat{y}(x)]$ can be written as

$$\mathrm{MSE}\,[\widehat{y}(x)] = \sigma^2\left\{1 - \begin{pmatrix} r(x)\\ f(x)\end{pmatrix}^\top \begin{bmatrix} R & F\\ F^\top & 0\end{bmatrix}^{-1}\begin{pmatrix} r(x)\\ f(x)\end{pmatrix}\right\}, \tag{2.49}$$

cf. [122]. From this representation it is easy to derive MSE$[\widehat{y}(x^{(i)})] = 0$ ($i = 1, \ldots, N$): the right matrix vector product again yields e_i, the i-th entry in the left vector is $R(x^{(i)} - x^{(i)}) = 1$ and the difference yields zero.

A short remark is given on the possibility to weaken the interpolation condition (2.48). When the observed data is assumed to be noisy or even stochastic, interpolation might not produce a suitable surrogate model. The Kriging method can be extended in order to approximate the data instead of interpolating it. In geostatistics, this is known as "nugget-effect", see e.g. [146]. A smoothing approximation for the response evaluations Y can for

instance be generated by regularizing the correlation matrix $R + \epsilon I$ ($\epsilon > 0$) [39, 110] or, with a similar effect, by scaling the correlation function $\widetilde{R}(h) := (1-\epsilon)R(h)$ for $h \neq 0$ while still satisfying $\widetilde{R}(0) = 1$ ($\epsilon \in (0,1)$) [124]. Since the responses investigated in this thesis are assumed to be deterministic and sufficiently smooth, interpolation is the method of choice and regularization is not applied.

Connection to radial basis functions

A more descriptive interpretation of the surrogate model is now presented. The expression (2.26) can be expanded in the following way:

$$\begin{aligned} \lambda(x)^\top Y &= \begin{pmatrix} \lambda(x) \\ \mu(x) \end{pmatrix}^\top \begin{pmatrix} Y \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} r(x) \\ f(x) \end{pmatrix}^\top \begin{bmatrix} R & F \\ F^\top & 0 \end{bmatrix}^{-1} \begin{pmatrix} Y \\ 0 \end{pmatrix}. \end{aligned} \tag{2.50}$$

Instead of solving the linear equation (2.38) for $\lambda(x)$ each time when the surrogate model $\widehat{y}(x) = \lambda(x)^\top Y$ is evaluated, one only has to solve the linear equation

$$\begin{bmatrix} R & F \\ F^\top & 0 \end{bmatrix} \begin{pmatrix} w^{(r)} \\ w^{(f)} \end{pmatrix} = \begin{pmatrix} Y \\ 0 \end{pmatrix} \tag{2.51}$$

once with a numerical effort of $\mathcal{O}\left((N+K)^3\right)$. The solution $\begin{pmatrix} w^{(r)} \\ w^{(f)} \end{pmatrix} \in \mathbb{R}^{N+K}$ is independent of $x \in \Omega$ and depends only on the given data set X, Y. The surrogate model can be evaluated in an arbitrary $x \in \Omega$

$$\widehat{y}(x) = \begin{pmatrix} w^{(r)} \\ w^{(f)} \end{pmatrix}^\top \begin{pmatrix} r(x) \\ f(x) \end{pmatrix} \tag{2.52}$$

with the numerical effort of a scalar product $\mathcal{O}\left(N+K\right)$.

In the last formulation (2.52) the connection between the Kriging method and radial basis functions method reveals. This property is sometimes referred to as duality of Kriging, e.g. in [100]. The surrogate is a weighted sum of the correlation functions $R(x^{(i)} - x)$ ($i = 1, \ldots, N$) and the regression functions $f_j(x)$ ($j = 1, \ldots, K$), where the weights $\begin{pmatrix} w^{(r)} \\ w^{(f)} \end{pmatrix}$ satisfy an interpolation condition (and an additional unbiasedness condition) (2.51). The correlation functions can be interpreted as local basis functions with centers $x^{(i)}$, satisfying $R(x^{(i)} - x^{(i)}) = 1$ and $R(x - x^{(i)}) \longrightarrow 0$ for $\|x^{(i)} - x\| \longrightarrow \infty$. But contrary to radial basis functions, $R(x^{(i)} - x)$ does not have to be radially symmetric around $x^{(i)}$, leaving more freedom to the model, cf. Sect. 2.3.3.

Connection to the model assumptions

In a second interpretation of the optimal weights $\lambda(x)$ in (2.38), a connection between the surrogate model $\widehat{y}(x) = \lambda(x)^\top Y$ (2.26) and the model assumption (2.27) is established. Using a partitioned inversion formula (cf. [123])

$$\begin{bmatrix} R & F \\ F^\top & 0 \end{bmatrix}^{-1} = \begin{bmatrix} R^{-1} - R^{-1}F(F^\top R^{-1}F)^{-1}F^\top R^{-1} & R^{-1}F(F^\top R^{-1}F)^{-1} \\ (F^\top R^{-1}F)^{-1}F^\top R^{-1} & -(F^\top R^{-1}F)^{-1} \end{bmatrix}, \tag{2.53}$$

the solution $\lambda(x)$ and $\mu(x)$ of (2.38) can be read separately

$$\mu(x) = (F^\top R^{-1}F)^{-1}\left(F^\top R^{-1}r(x) - f(x)\right), \tag{2.54}$$

$$\lambda(x) = R^{-1}\left(r(x) - F\mu(x)\right). \tag{2.55}$$

Hence, the Kriging predictor can be rewritten in the following form:

$$\begin{aligned} \widehat{y}(x) &= \lambda(x)^\top Y \\ &= \left(r(x) - F\mu(x)\right)^\top R^{-1}Y \\ &= r(x)^\top R^{-1}Y - \left(F^\top R^{-1}r(x) - f(x)\right)^\top \underbrace{(F^\top R^{-1}F)^{-1}F^\top R^{-1}Y}_{=:\beta} \\ &= f(x)^\top\beta + r(x)^\top R^{-1}(Y - F\beta) \\ &= f(x)^\top\beta + r(x)^\top R^{-1}Z, \end{aligned} \tag{2.56}$$

where $R^{-1}Z = w^{(r)}$ and $\beta = w^{(f)}$ correspond to the weights in (2.52). Thus, according to the model assumption (2.27), the surrogate is a sum of a regression model $f(x)^\top\beta$ with a generalized least squares estimate $\beta = (F^\top R^{-1}F)^{-1}F^\top R^{-1}Y$ as in (2.17) and a second term $r(x)^\top R^{-1}Z$. This second term is a weighted sum of correlation functions $R(x^{(i)} - x)$ and the weights are defined by an interpolation condition for the deviations Z from the regression model. While the regression model (usually of low order) approximates the drift or the global trend of the unknown response $y(x)$, the second term captures the nonlinear behavior of $y(x)$ and "pulls" the surrogate $\widehat{y}(x)$ through the observed data Y based on the statistical assumptions on the Gaussian process $z(x)$, cf. [73].

Commonly, the linear model is kept small on purpose and often only a constant regression term or monomials of order 1 are used. The simplest case $K = 1$, $f_1 \equiv 1$, $y(x) = \beta_1 + z(x)$ is called *ordinary Kriging*, whereas the more general case using low-order polynomials is called *universal Kriging*. The author of [124] demonstrated that the influence of the regression model on the quality of the Kriging prediction is negligible when interpolating an unknown nonlinear function. If no global trend is known, ordinary Kriging suffices. In the remainder of this thesis ordinary Kriging will be used for interpolation, but the universal Kriging notation is maintained to emphasize that all methods can be applied to a general $\sum_{j=1}^{K} f_j(x)\beta_j$ as well.

2.3.3 Correlation functions

Before a Kriging predictor is generated, a correlation function $R(\tilde{x} - x)$ for $z(x)$ has to be specified. As identified in the previous section, the surrogate model $\widehat{y}(x) = \lambda(x)^\top Y = \begin{pmatrix} w^{(r)} \\ w^{(f)} \end{pmatrix}^\top \begin{pmatrix} r(x) \\ f(x) \end{pmatrix}$ is a weighted sum of correlation functions $r_i(x) = R(x^{(i)} - x)$ with centers $x^{(i)}$ $(i = 1, \ldots, N)$ and the regression functions $f_j(x)$ $(j = 1, \ldots, K)$. Hence, the analytical properties of $\widehat{y}(x)$ are directly defined by $R(\tilde{x} - x)$ and $f_j(x)$ $(j = 1, \ldots, K)$. Therefore, the choice of an appropriate correlation function becomes crucial.

In the following, let a family of correlation functions be parametrized by so-called *hyperparameters* $\theta = (\theta_1, \ldots, \theta_d)^\top \in \mathbb{R}^d$. A common restriction is to model the multivariate $R : \mathbb{R}^d \to \mathbb{R}$ as a product of univariate correlation functions

$$R(\tilde{x} - x, \theta) = \prod_{k=1}^{d} R_k \left(|\tilde{x}_k - x_k|, \theta_k \right). \tag{2.57}$$

$\theta_k > 0$ is usually a scaling factor, which controls the influence of an evaluation of the response $y_i = y(x^{(i)})$ along the x_k-axis, cf. Fig. 2.3. The θ_k can be chosen different for each dimension k of the input parameter space Ω, so different sensitivities of the response y with respect to each variable x_k can be taken into account $(k = 1, \ldots, d)$, see Fig. 2.4. This leaves more freedom to the surrogate model than radial basis functions do, where the basis functions depend only on $\|\tilde{x} - x\|_2$.

It is also possible to model *anisotropy*. Not only scaling in each direction is permitted, but also a rotation in the input space $\mathbb{R}^d$ by additional parameters $\bar{\theta}$. A rotation matrix $Q_{\bar{\theta}}$ is multiplied to the argument $\tilde{x} - x$ and the correlation function becomes $R(Q_{\bar{\theta}}(\tilde{x} - x), \theta)$. For $d = 2$, a rotation angle $\bar{\theta}_1$ has to be specified for the rotation matrix

$$Q_{\bar{\theta}_1} = \begin{bmatrix} \cos\bar{\theta}_1 & -\sin\bar{\theta}_1 \\ \sin\bar{\theta}_1 & \cos\bar{\theta}_1 \end{bmatrix}. \tag{2.58}$$

For $d = 3$, the rotation matrix is obtained by multiplying the rotation matrices for the $x_{1,2}$-, $x_{1,3}$- and $x_{2,3}$-plane

$$Q_{\bar{\theta}_1,\bar{\theta}_2,\bar{\theta}_3} = \begin{bmatrix} \cos\bar{\theta}_3 & -\sin\bar{\theta}_3 & 0 \\ \sin\bar{\theta}_3 & \cos\bar{\theta}_3 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\bar{\theta}_2 & 0 & -\sin\bar{\theta}_2 \\ 0 & 1 & 0 \\ \sin\bar{\theta}_2 & 0 & \cos\bar{\theta}_2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\bar{\theta}_1 & -\sin\bar{\theta}_1 \\ 0 & \sin\bar{\theta}_1 & \cos\bar{\theta}_1 \end{bmatrix}. \tag{2.59}$$

In geostatistics for instance, there is often little reason for the assumption that the direction of a strong correlation coincides with the north-south-axis or the east-west-axis in spatial modeling, cf. [146]. Then, rotation of the correlation axis can improve the approximation quality. In computer experiments however, the input parameters x_k $(k = 1, \ldots, d)$ generally correspond to different physical inputs. The global influence of each x_k is modeled by the scaling factor θ_k and rotation is not applied. Also, for $d > 3$, modeling anisotropy becomes challenging because of the growing number of the rotation parameters $\bar{\theta}$ yet to identify.

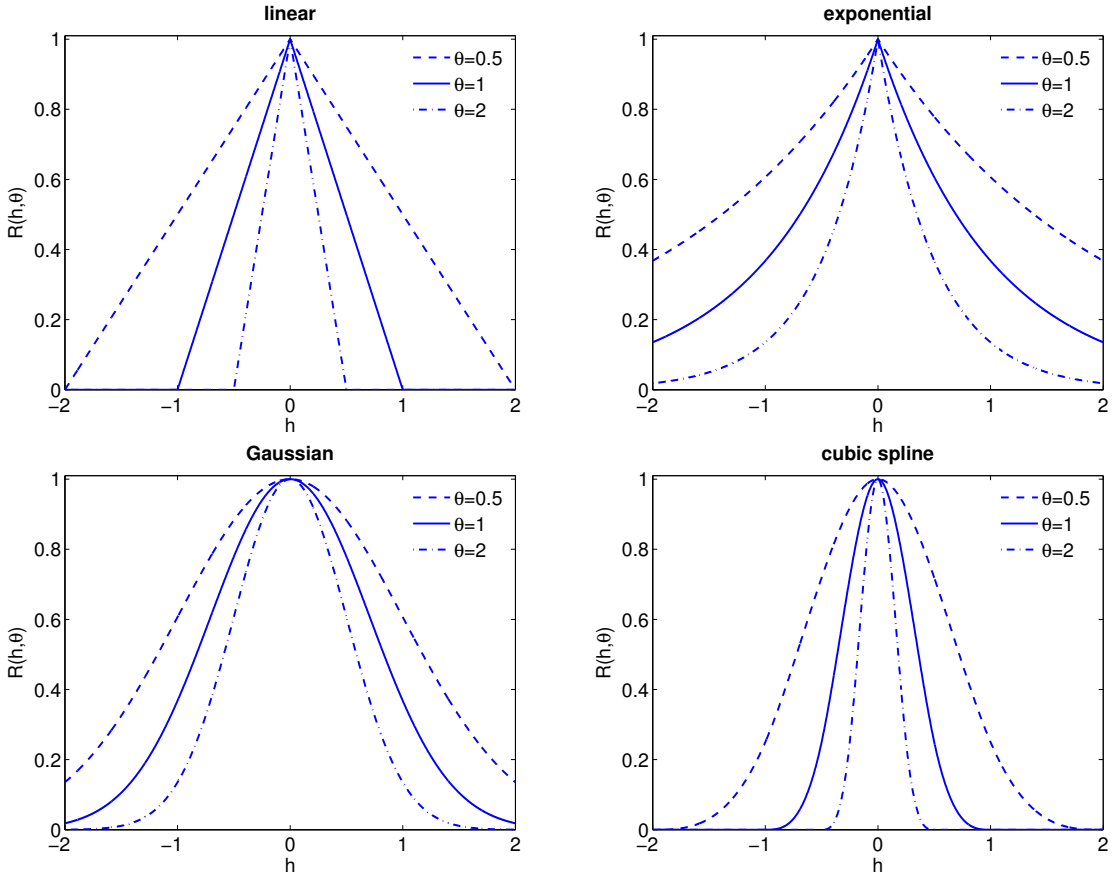

Figure 2.3: Univariate correlation functions for $\theta = 0.5,\ 1.0,\ 2.0$ ($h := \tilde{x} - x$).

Examples

Several popular choices for families of univariate correlation functions are now presented and depicted in Fig. 2.3. For simplicity, a distance vector $h := \tilde{x} - x$ is defined. The *linear correlation functions*

$$R(\tilde{x} - x, \theta) = \max\left\{1 - \theta\left|\tilde{x} - x\right|, 0\right\} \tag{2.60}$$

are $C^0(\mathbb{R})$-functions, which are not differentiable in the origin and in $|h| = \frac{1}{\theta}$. Also, for $|h| \geq \frac{1}{\theta}$, $z(x)$ and $z(\tilde{x})$ have zero correlation. This can be favored in applications where for a sufficient distance of x and $\tilde{x}$, the responses $y(x)$ and $y(\tilde{x})$ are assumed to behave distinctly. In terms of the basis functions $R(x^{(i)} - x)$ this means a true local behavior, since they have a finite support $\mathrm{supp}(R(x^{(i)} - x)) = \{x : |x - x^{(i)}| < \frac{1}{\theta}\}$ $(i = 1, \ldots, N)$.

A second example are the *exponential correlation functions*

$$R(\tilde{x} - x, \theta) = \exp\left\{-\theta\left|\tilde{x} - x\right|\right\}. \tag{2.61}$$

They are also $C^0(\mathbb{R})$, not differentiable in the origin and $R(h)$ decreases fast for small h. In contrast to the previous example, $R(h) > 0$ holds for all $h \in \mathbb{R}$.

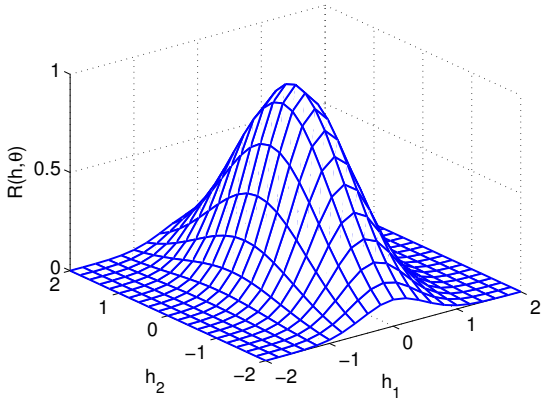

Figure 2.4: A two-dimensional Gaussian correlation function for $\theta_1 = 2$, $\theta_2 = 0.5$.

A family of widely used functions are the *Gaussian correlation functions*

$$R(\tilde{x} - x, \theta) = \exp\left\{-\theta \left(\tilde{x} - x\right)^2\right\}, \tag{2.62}$$

which are $C^\infty(\mathbb{R})$ and therefore produce very smooth interpolations $\widehat{y}(x)$. Although this is a desired quality in many applications, for responses with highly nonlinear (e.g. by large gradients and curvature) or even erratic behavior, this can produce inappropriate interpolations. Also, the correlation matrices based on the smooth Gaussian correlation functions are known to be ill-conditioned in many cases, e.g. if θ is small or if any $x^{(i)}, x^{(j)}$ are close $(i, j = 1, \ldots, N)$, see Fig. 2.5 and also [88, 30].

Cubic splines whose shapes resemble the Gaussian correlation functions are defined by

$$R(\tilde{x} - x, \theta) = \begin{cases} 1 - 6\left(\theta|h|\right)^2 + 6\left(\theta|h|\right)^3, & 0 \le \theta|h| < 0.5 \\ 2\left(1 - \theta|h|\right)^3, & 0.5 \le \theta|h| < 1 \\ 0, & 1 \le \theta|h| \end{cases}. \tag{2.63}$$

They are twice continuously differentiable in $\mathbb{R}$ and have a finite support $\mathrm{supp}(R(h)) = \{h : |h| < \frac{1}{\theta}\}$. Unlike the Gaussian correlation functions, the correlation matrices based on these cubic splines are well-conditioned, cf. Fig. 2.5.

The *generalized exponential correlation functions*

$$R(\tilde{x} - x, \theta, p) = \exp\left\{-\theta \left|\tilde{x} - x\right|^p\right\} \tag{2.64}$$

have a second hyperparameter $p \in (0, 2]$ which can also vary for the different dimensions $k = 1, \ldots, d$ of multivariate correlation functions $R(h, \theta, p) = \prod_{k=1}^{d} \exp\left\{-\theta_k |h_k|^{p_k}\right\}$. It defines the smoothness of the correlation function and the model contains exponential correlation functions $(p = 1)$ as well as Gaussian correlation functions $(p = 2)$.

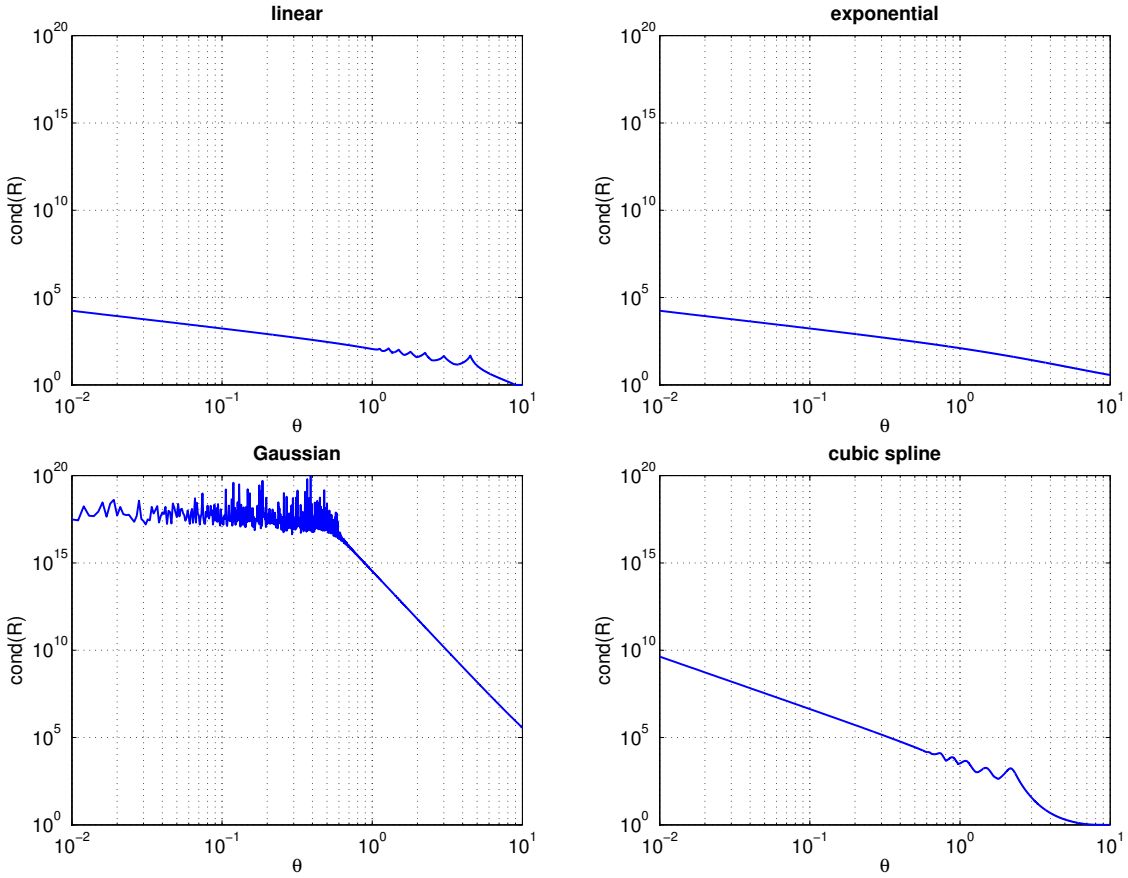

Figure 2.5: Condition number of a correlation matrix $R \in \mathbb{R}^{10\times 10}$ depending on θ for $X = \{x^{(i)} = \frac{i-1}{10-1},\ i = 1,\ldots,10\}$.

A necessary and sufficient condition for validness

Not any function $R : \mathbb{R}^d \to \mathbb{R}$ can be used as a correlation function. In this section, a criterion for validness based on Fourier transformations is discussed.

Definition 2.5 ((strictly) positive definite functions). *A function $R : \mathbb{R}^d \to \mathbb{R}$ is called positive definite, if for arbitrary $N \in \mathbb{N}$*

$$\sum_{i,j=1}^{N} w_i w_j R(x^{(j)} - x^{(i)}) \geq 0 \quad \forall x^{(1)},\ldots,x^{(N)} \in \mathbb{R}^d,\ w \in \mathbb{R}^N. \tag{2.65}$$

It is called strictly positive definite, if

$$\sum_{i,j=1}^{N} w_i w_j R(x^{(j)} - x^{(i)}) > 0 \quad \forall x^{(i)} \neq x^{(j)}\ (i \neq j),\ w \in \mathbb{R}^N \setminus \{0\}. \tag{2.66}$$

This is the classical terminology of definiteness for functions, which is not consistent with

the terminology of definiteness for matrices. Therefore, in the literature sometimes "positive semidefinite" is used for positive definite functions and "positive definite" for strictly positive functions, e.g. in [35, 147]. Now any correlation function $R(\tilde{x}-x)$ is required to be a positive definite function, because any correlation matrix $R = [R(x^{(j)} - x^{(i)})]_{i,j=1}^{N,N} \in \mathbb{R}^{N\times N}$ has to be positive semidefinite. Furthermore, any positive definite function $R(h)$ satisfying $R(0) = 1$ is a valid correlation function, because positive definiteness already implies e.g. $R(h) \leq R(0)$, see [123]. A necessary and sufficient characterization of positive definite functions is given by Bochner's theorem, dating back to 1932, which is here presented in a real-valued version of [134].

Theorem 2.6 (Bochner). *A continuous function $R : \mathbb{R}^d \to \mathbb{R}$ is positive definite if and only if it can be represented as the characteristic function of a positive finite measure $S(\cdot)$:*

$$R(h) = \int_{\mathbb{R}^d} \exp\{i\omega^\top h\}\ dS(\omega). \tag{2.67}$$

Proof. A proof can be found in [44], Chap. IV, Sect. 2. □

Since $R(h)$ is real-valued and $\exp\{i\omega^\top h\} = \cos(\omega^\top h)+i\sin(\omega^\top h)$, $S(\cdot)$ has to be symmetric such that the imaginary sine term under the integral vanishes. (2.67) then reduces to

$$R(h) = \int_{\mathbb{R}^d} \cos(\omega^\top h)\ dS(\omega). \tag{2.68}$$

If $S(\cdot)$ has a density $s(\cdot)$, this reads

$$R(h) = \int_{\mathbb{R}^d} \cos(\omega^\top h)s(\omega)\ d\omega \tag{2.69}$$

and the density function is the Fourier transform of $R(h)$, cf. [134]:

$$\begin{aligned} s(\omega) &= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \exp\{-i\omega^\top h\}R(h)\ dh \\ &= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \cos(\omega^\top h)R(h)\ dh. \end{aligned} \tag{2.70}$$

Thus Bochner's theorem contains both a method for constructing correlation functions based on symmetric density functions as well as a method of checking if a given function is a valid correlation function. If for a given $R(h)$ the Fourier transform $s(\omega)$ exists and $s(\omega) \geq 0$, $\int_{\mathbb{R}^d} s(\omega)\ d\omega < \infty$, then $R(h)$ can be represented as the characteristic function a positive finite measure and therefore is a valid correlation function. For a product model correlation function (2.57)

$$R(h) = \prod_{k=1}^{d} R_k(h_k) \tag{2.71}$$

it is sufficient that all one-dimensional $R_k(h_k)$ are positive definite ($k = 1, \ldots, d$), since products of positive definite functions are positive definite, see Sect. 2.3 in [134].

For Kriging (and also for local radial basis functions) positive definite functions are not sufficient. For practical use, the correlation matrix $R = [R(x^{(j)} - x^{(i)})]_{i,j=1}^{N,N} \in \mathbb{R}^{N\times N}$ has

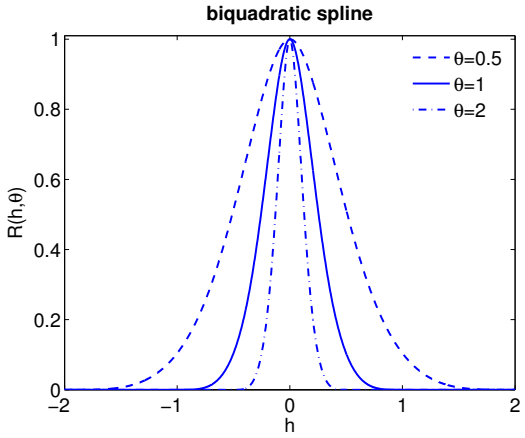

Figure 2.6: Biquadratic spline correlation function for $\theta = 0.5,\ 1.0,\ 2.0$.

to be nonsingular and therefore any correlation function is required to be strictly positive definite. In [147], an extension of Bochner's characterization on strictly positive definite functions is given:

Theorem 2.7. *A continuous function $R : \mathbb{R}^d \to \mathbb{R}$, $R \in L^1(\mathbb{R}^d)$ is strictly positive definite if and only if $R(h)$ is bounded and its Fourier transform is nonnegative and not equal to zero.*

Proof. A proof is given in [147], Theorem 6.11. □

Also, for the product model correlation function (2.57), it is sufficient that all one-dimensional $R_k(h_k)$ are strictly positive definite ($k = 1, \ldots, d$), since products of strictly positive definite functions are strictly positive definite, again see [147] or a recent discussion in [59]. The one-dimensional examples for correlation functions from the previous section are all strictly positive definite functions, cf. [35, 123]

The criterion is now illustrated with an example. A cubic spline correlation function was already introduced, cf. (2.63). It can be used as an approximation to the Gaussian correlation function (2.62), but with a compact support and well-conditioned correlation matrices. It is only twice continuously differentiable in $\mathbb{R}$. Now suppose $R(h) \in C^3(\mathbb{R} \setminus \{0\})$ is required (cf. Sect. 2.3.4). The *biquadratic spline* (Fig. 2.6)

$$R(\tilde{x} - x, \theta) = \begin{cases} 1 - 15(\theta|h|)^2 + 35(\theta|h|)^3 - \frac{195}{8}(\theta|h|)^4, & 0 \le \theta|h| < 0.4 \\ \frac{5}{3} - \frac{20}{3}(\theta|h|) + 10(\theta|h|)^2 - \frac{20}{3}(\theta|h|)^3 + \frac{5}{3}(\theta|h|)^4, & 0.4 \le \theta|h| < 1 \\ 0, & 1 \le \theta|h| \end{cases} \tag{2.72}$$

is also an approximation to the Gaussian correlation function and it is twice continuously differentiable in $\mathbb{R}$ and three times in $\mathbb{R} \setminus \{0\}$. It is bounded, has a compact support and $\int_{\mathbb{R}} R(h, \theta)\, dh < \infty$ for all $\theta > 0$. Also, the Fourier transform $s(\omega) = \frac{1}{2\pi} \int_{\mathbb{R}} \cos(\omega h) R(h, \theta)\, dh$ exists for all $\theta > 0$ and $\omega \in \mathbb{R}$. Since both θ and ω act as a linear scaling parameter in h, it is sufficient to show that $s(\omega) \ge 0\ \forall \omega \in \mathbb{R}$ for a single $\theta > 0$. Then $s(\omega) \ge 0\ \forall \omega \in \mathbb{R}$ holds for all $\theta > 0$. Figure 2.7 shows the Fourier transform $s(\omega)$ of $R(h, \theta)$ for $\theta = 1$, computed

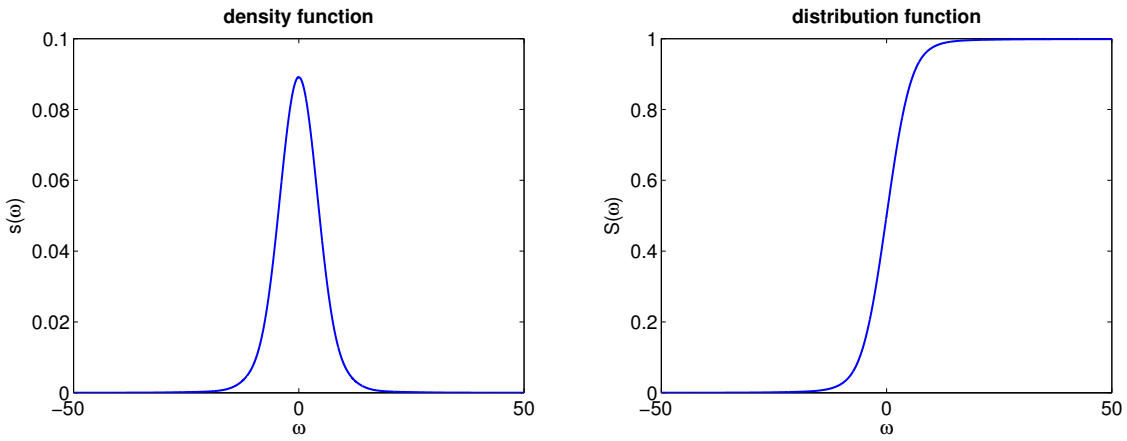

Figure 2.7: Fourier transform $s(\omega)$ of the biquadratic spline for $\theta = 1$ and distribution function $S(\omega) := \int_{-\infty}^{\omega} s(\xi)\, d\xi$.

with symbolic integration using Mathematica. It is nonnegative, therefore the biquadratic spline is a strictly positive definite function and also a valid correlation function.

2.3.4 Hyperparameter estimation

When a suitable correlation function $R(h, \theta)$ is chosen, the hyperparameters $\theta = (\theta_1, \ldots, \theta_d)^\top$ still have to be defined. For the correlation functions in the previous sections, the hyperparameters act as a scaling parameter in h. For the introduced functions, the scaling in h_k is proportional to $\frac{1}{\theta_k}$, except for the Gaussian correlation function where it is proportional to $\sqrt{\frac{1}{\theta_k}}$ and the generalized exponential function where it is proportional to $\sqrt[p]{\frac{1}{\theta_k}}$ $(k = 1, \ldots, d)$. The Kriging predictor is defined by a weighted sum of the correlation functions $R(x^{(i)} - x, \theta)$ $(i = 1, \ldots, N)$ and the regression functions $f_j(x)$ $(j = 1, \ldots, K)$. So not only the choice of the correlation model, but also the hyperparameters θ have a great influence on the approximation quality.

Again, the statistical properties of $y(x)$ and therefore $Y = (y_1, \ldots, y_N)^\top$ can be exploited to determine optimal hyperparameters θ, which suit the given data set X, Y best. The process of adapting the hyperparameters to the data is often referred to as model fitting. Note that for every $\theta \in \mathbb{R}_+^d$, the Kriging predictor is an interpolator and the value of θ affects $\widehat{y}(x)$ only in $\Omega \setminus X$. The N-dimensional vector of evaluations is multivariate normally distributed $Y \sim \mathcal{N}_N(F\beta, \sigma^2 R(\theta))$, and the correlation matrix's entries are depending on θ. Maximum likelihood estimation is applied for finding the unknowns $\beta \in \mathbb{R}^K$, $\sigma^2 \in \mathbb{R}_+$ and $\theta \in \mathbb{R}_+^d$, cf. also Sect. 2.1.2. The likelihood function reads

$$L(\beta, \sigma^2, \theta | Y) = (2\pi\sigma^2)^{-\frac{N}{2}} (\det R(\theta))^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(Y - F\beta)^\top R(\theta)^{-1}(Y - F\beta)\right). \quad (2.73)$$

Maximizing the likelihood function $L(\beta, \sigma^2, \theta | Y)$ is equivalent to minimizing the negative

logarithm of $L(\beta, \sigma^2, \theta|Y)$

$$-\log L(\beta, \sigma^2, \theta|Y) = \frac{N}{2}\log(2\pi\sigma^2) + \frac{1}{2}\log(\det R(\theta)) + \frac{1}{2\sigma^2}(Y - F\beta)^\top R(\theta)^{-1}(Y - F\beta). \quad (2.74)$$

Its partial derivatives with respect to β_j, σ^2 and θ_k ($j = 1, \ldots, K$, $k = 1, \ldots, d$) are

$$\frac{\partial}{\partial \beta_j}(-\log L) = -\frac{1}{\sigma^2}\left(F^\top R(\theta)^{-1}(Y - F\beta)\right)_j, \quad (2.75)$$

$$\frac{\partial}{\partial \sigma^2}(-\log L) = \frac{N}{2\sigma^2} - \frac{1}{2\sigma^4}(Y - F\beta)^\top R(\theta)^{-1}(Y - F\beta), \quad (2.76)$$

$$\frac{\partial}{\partial \theta_k}(-\log L) = \frac{1}{2}\operatorname{trace}\left(R(\theta)^{-1}\frac{\partial R(\theta)}{\partial \theta_k}\right) - \frac{1}{2\sigma^2}(Y - F\beta)^\top R(\theta)^{-1}\frac{\partial R(\theta)}{\partial \theta_k}R(\theta)^{-1}(Y - F\beta), \quad (2.77)$$

where $\frac{\partial R(\theta)}{\partial \theta_k} \in \mathbb{R}^{N\times N}$ is the component-wise derivative of the correlation matrix $R(\theta)$, cf. [73]. First order optimality conditions are then applied. The partial derivatives with respect to β are zero if and only if

$$\beta^* := \beta(\theta) = \left(F^\top R(\theta)^{-1}F\right)^{-1} F^\top R(\theta)^{-1}Y, \quad (2.78)$$

where $F^\top R(\theta)^{-1}F$ is a positive definite matrix (see Theorem 2.2). So, for any $\theta \in \mathbb{R}^d_+$, $\beta^* = \beta(\theta)$ is the maximum likelihood estimate. It is equal to the generalized least squares estimate in (2.56). The derivative with respect to σ^2 is zero if and only if

$$\sigma^{2*} := \sigma^2(\theta, \beta(\theta)) = \frac{1}{N}(Y - F\beta(\theta))^\top R(\theta)^{-1}(Y - F\beta(\theta)). \quad (2.79)$$

So, for any $\theta \in \mathbb{R}^d_+$, the maximum likelihood estimate $\sigma^{2*} = \sigma^2(\theta, \beta(\theta))$ is also defined. The target function $-\log L$ is highly nonlinear in θ, since it involves the correlation functions and the inverse of the correlation matrix as well as the logarithm of its determinant. Therefore, it is not possible to derive the maximum likelihood estimate θ^* in an easy fashion like β^* and σ^{2*}. Instead, numerical optimization is used to solve the problem

$$\min_{\beta\in\mathbb{R}^K, \sigma^2\in\mathbb{R}_+, \theta\in\mathbb{R}^d_+} \left\{-\log L(\beta, \sigma^2, \theta|Y)\right\}. \quad (2.80)$$

The maximum likelihood problem can be reduced by substituting β and σ^2 by β^* and σ^{2*} in (2.74). The last summand is canceled down to a constant, leaving a reduced optimization problem

$$\min_{\theta\in\mathbb{R}^d_+} \left\{N\log\left(\sigma^2(\theta, \beta(\theta))\right) + \log\left(\det R(\theta)\right)\right\}, \quad (2.81)$$

which is only depending on θ. Note that though β and σ^2 are not needed explicitly for the Kriging predictor $\widehat{y}(x)$, they have to be computed for the maximum likelihood estimation of θ, either directly in (2.80) or implicitly in (2.81).

When using a gradient-based optimization algorithm, formulation (2.81) can cause difficulties because of the highly nonlinear implicit dependencies of θ in $\beta(\theta)$ and $\sigma^2(\theta, \beta(\theta))$. The nonlinearity of the target function in θ is even increased in comparison to the original problem (2.80). This results in a complex computation of the derivatives due to an exhaustive exploitation of the chain rule, see [140]. For this reason, it can be of advantage to solve the original problem (2.80) instead. The additional variables σ^2 and β cause no severe increase in the complexity of a gradient-based numerical optimization method. σ^2 is always scalar and β has the size of the regression model K, which is usually kept small and for *ordinary Kriging* $K = 1$. The target function in (2.80) has nice properties in β and σ^2. Not only do its partial derivatives with respect to β and σ^2 have unique zeros. Examining also the second order partial derivatives

$$\frac{\partial^2}{\partial \beta^2}(-\log L) = \frac{1}{\sigma^2} F^\top R(\theta)^{-1} F, \tag{2.82}$$

$$\frac{\partial^2}{\partial (\sigma^2)^2}(-\log L) = -\frac{N}{2\sigma^4} + \frac{1}{\sigma^6}(Y - F\beta)^\top R(\theta)^{-1}(Y - F\beta), \tag{2.83}$$

it can easily be shown that $-\log L$ is quadratic and convex in $\beta \in \mathbb{R}^K$ for any $\theta \in \mathbb{R}_+^d$, $\sigma^2 \in \mathbb{R}_+$, and convex in $\sigma^2 \in (0, \frac{2}{N}(Y - F\beta)^\top R(\theta)^{-1}(Y - F\beta))$ for any $\theta \in \mathbb{R}_+^d$, $\beta \in \mathbb{R}^K$. The upper bound of this interval is close to $2\sigma^{2^*}$ if β is close to β^*.

Note that whichever approach is pursued, the partial derivatives of the objective function with respect to θ_k contain partial derivatives of the correlation functions $\frac{\partial}{\partial \theta_k} R_k(h_k, \theta_k)$. To ensure a smooth objective function, a family of correlation functions which is differentiable with respect to $\theta > 0$ should be chosen. This is e.g. not the case for the linear correlation function $R(h, \theta) = \max\{1 - \theta|h|, 0\}$, which is not differentiable in $\theta = \frac{1}{|h|}$. In this case the objective function is not differentiable in

$$\left\{\theta \in \mathbb{R}_+^d : \theta_k = |x_k^{(j)} - x_k^{(i)}|^{-1},\ k = 1, \ldots, d;\ i, j = 1, \ldots, N\right\},$$

a set which consists of up to $d \cdot \binom{N}{2}$ hyperplanes of dimension $(d - 1)$. However, all other correlation functions presented in the previous section are differentiable with respect to $\theta > 0$.

A theoretical and numerical discussion of the maximum likelihood problem for spatial covariance functions was given in [92], where several numerical procedures were suggested. Since the introductory publication on computer experiments by Sacks et al. [122], the reduced problem (2.81) is found often in the literature, e.g. [22, 134, 89, 123]. An alternative method is to update $\beta(\theta)$ and $\sigma^2(\theta, \beta(\theta))$ in every step of an iterative process and then do a descent step in θ using the partial derivative (2.77), see [73]. Two main difficulties may arise when solving the maximum likelihood problem numerically: possible monotonic behavior of the target function and possible multiple local optima. The former was e.g. discussed in [124], where for a one-dimensional test case no correlation between the $y(x^{(i)})$ $(i = 1, \ldots, N)$ was detected, i.e. the target function decreased monotonically in θ. This results in a large solution θ^* and a poor approximation quality of the surrogate model $\widehat{y}(x)$. A similar behavior has recently been observed in [152], where the target function reached a constant plateau and numerically no local optimum existed. However, these were cases with $d = 1$ and only

a small number of samples $N < 10$. The authors of [112] emphasized that this behavior is typical for few data points and for a larger N the target function is less shallow. The latter difficulty of multiple local optima was e.g. investigated in [88, 112, 30]. Contrary to the partial derivatives with respect to β and σ^2, there is no guarantee that the partial derivatives with respect to θ_k (2.77) have unique zeros or that the reduced likelihood problem (2.81) has only one local minimum. Again, the authors of [112] pointed out that in practice, multiple local optima do not pose a severe threat and for larger sample sizes N the global optimum has a target value which differs vastly from the other local optimas' target values.

The approach of using correlation functions and obtaining parameters by maximum likelihood estimation generally differs from the approach pursued in geostatistics. Classically, in statistics for spatial data a variogram function

$$2\gamma(h) := \mathrm{Var}\left[y(x+h) - y(x)\right], \tag{2.84}$$

satisfying $\gamma(0) = 0$, $\gamma(h) \geq 0$ and $\gamma(-h) = \gamma(h)$, is utilized instead of a correlation function. It describes the average deviation in the response values depending on the distance vector $h \in \mathbb{R}^d$. Typically, it is assumed to be radially symmetric around the origin and monotonically increasing in $\|h\|$. A variogram cloud is a point cloud in $\mathbb{R}^2$ of size $\binom{N}{2}$ defined by

$$\left\{\left(\|x^{(j)} - x^{(i)}\|, (y(x^{(j)}) - y(x^{(i)}))^2\right),\ i < j\right\}. \tag{2.85}$$

A variogram function is then obtained by fitting a function from a suitable family of variogram functions to the variogram cloud. For this purpose, a least squares problem has to be solved to determine the right parametrization of the variogram function. The connection to correlation functions is given by the fact that by

$$\gamma(h) = \sigma^2\left(1 - R(h)\right) \tag{2.86}$$

a variogram function can be defined. More information can be found in geostatistical textbooks [22, 146, 126].

Examples

The subject of hyperparameter estimation is now illustrated with two examples. The first one is a one-dimensional test function taken from [41]

$$y(x) = (6x-2)^2 \sin(12x-4), \tag{2.87}$$

which is interpolated in $\Omega = [0, 1]$. The influence of the hyperparameter θ on the approximation quality of $\widehat{y}(x)$ is now discussed. Figure 2.8 depicts interpolations for different values of θ using the cubic correlation function. The $N = 9$ samples $x^{(i)}$ are distributed more densely in the upper half of the interval in order to resolve the more erratic behavior of the function more precisely. Recall that θ is a scaling factor in h for the correlation function. The larger θ, the faster $R(h, \theta)$ decreases in h. For the cubic correlation function, $R(h, \theta) = 0$ for all $h \geq \frac{1}{\theta}$ holds. Since $\widehat{y}(x)$ is a weighted sum of the correlation functions $R(x^{(i)} - x, \theta)$ with centers in $x^{(i)}$ and a constant regression term (2.52), the response value $y(x^{(i)})$ affects the

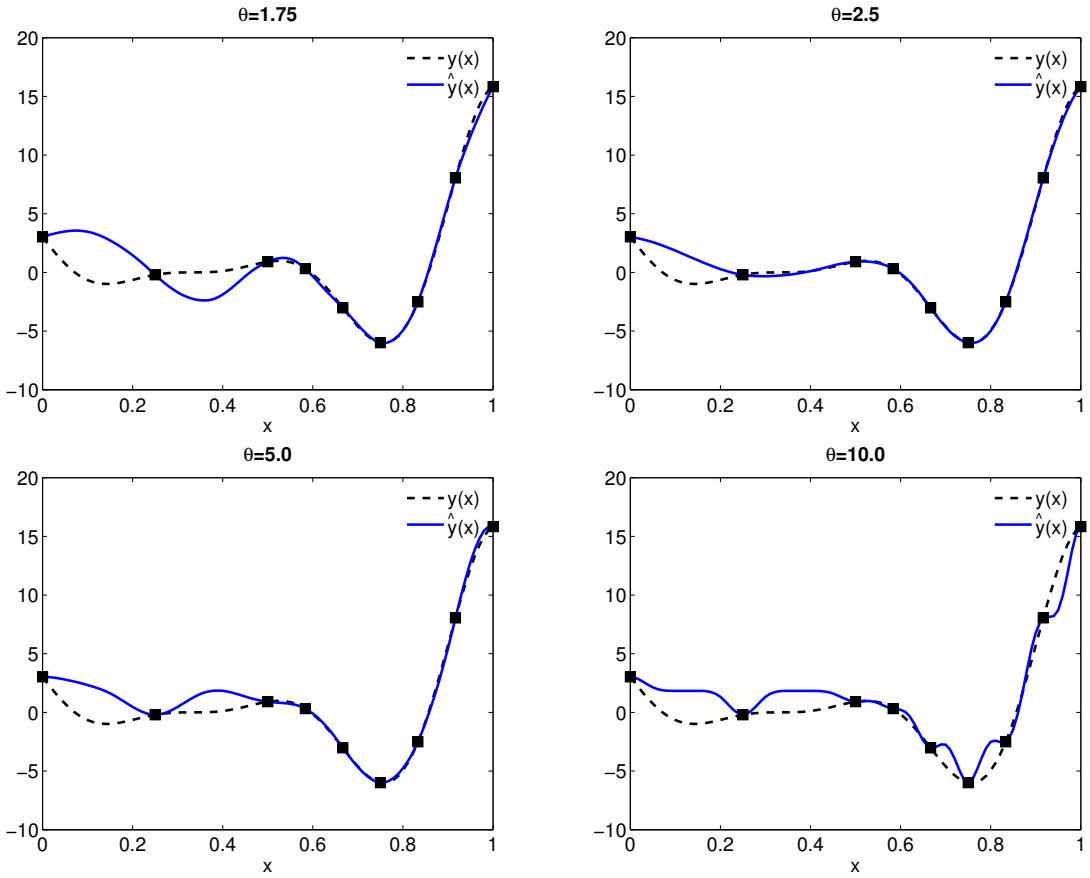

Figure 2.8: Kriging interpolations $\widehat{y}(x)$ using the cubic spline correlation for different θ values, $y(x) = (6x-2)^2 \sin(12x-4)$.

surrogate model $\widehat{y}(x)$ only in the interval $(x^{(i)} - \frac{1}{\theta}, x^{(i)} + \frac{1}{\theta})$. $\theta^* = 2.5$ is the approximate solution of the maximum likelihood problem (cf. Fig. 2.9). The corresponding surrogate model is a good approximation for the original response, except for $x \in (0, 0.25)$ where the data X, Y does not contain enough information about $y(x)$. In this region, no choice of θ can produce a suitable approximation. Now on the one hand, if θ is chosen too small (e.g. $\theta = 1.75$), the correlation between the $y(x^{(i)})$ $(i = 1, \ldots, N)$ is overestimated. This leads to a large influence of the $y(x^{(i)})$ on the behavior of $\widehat{y}(x)$, even for the $x^{(i)}$ which are far away from a location x. In the region with few samples $x \in [0, 0.5]$, unwanted variations of $\widehat{y}(x)$ occur and the approximation quality suffers. On the other hand, if θ is chosen too large ($\theta = 5.0$), $R(h, \theta)$ decays too fast towards zero. Again, in regions with few samples the approximation quality suffers. With an even larger $\theta = 10.0$, also in the denser sampled upper half interval, $\widehat{y}(x)$ is not a suitable approximation to $y(x)$ anymore. Generally, for any correlation function, which has θ as a scaling parameter in h, $R(x^{(i)} - x, \theta) \to \mathbb{1}_{\{0\}}(x)$ holds for $\theta \to \infty$, where $\mathbb{1}_{\{0\}}$ is the indicator function ($\mathbb{1}_{\{0\}}(0) = 1$, $\mathbb{1}_{\{0\}}(x) = 0$ elsewhere). Therefore, the correlation matrix becomes the identity matrix ($R \to I$). Furthermore, for any $x \neq x^{(i)}$ $(i = 1, \ldots, N)$, the surrogate converges to the regression term $\widehat{y}(x) \to f(x)^\top \beta$,

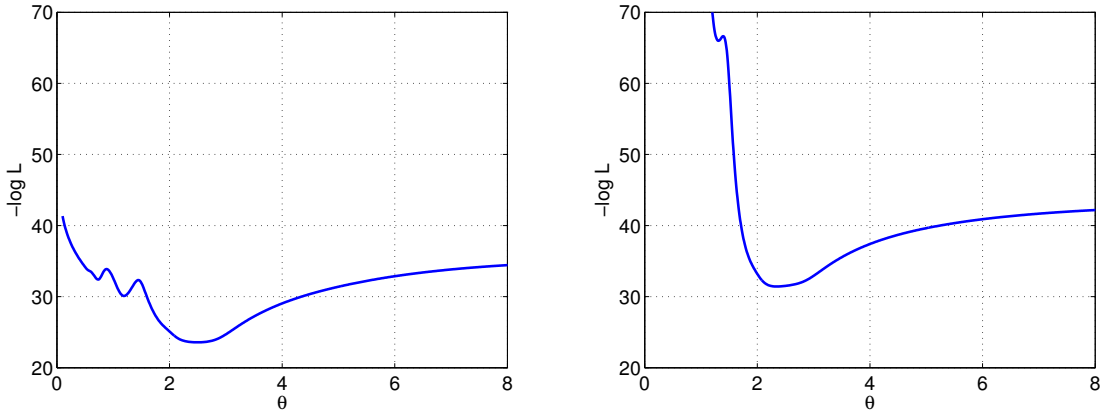

Figure 2.9: $-\log L$ for the data as in Fig. 2.8. Cubic spline correlation function. Left: reduced problem. Right: original problem (fixed β, σ^2).

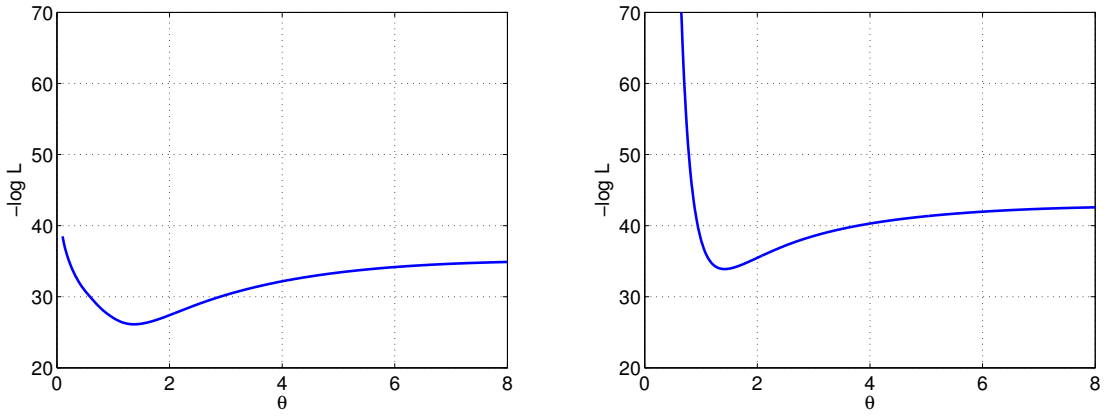

Figure 2.10: $-\log L$ for the data as in Fig. 2.8. Biquadratic spline correlation function. Left: reduced problem. Right: original problem (fixed β, σ^2).

cf. representation (2.56).

The maximum likelihood problem is also discussed. In Fig. 2.9, the target function of the reduced likelihood problem (2.81) depending only on θ is plotted for $\theta \in [0.05, 8.0]$ as well as the target function of the original problem $-\log L(\theta, \beta, \sigma^2|Y)$, where β and σ^2 are kept fixed in $\beta(\theta^*) = 2.95$ and $\sigma^2(\theta^*, \beta(\theta^*)) = 35.35$. Clearly, both have the same global optimal point $\theta^* = 2.50$. Both target functions tend to infinity for $\theta \to 0$ and converge to a (not coinciding) constant for $\theta \to \infty$, see [152] for a detailed discussion of the asymptotic behavior. The reduced target function has two local minima left of the global optimum to which a gradient based optimization algorithm will presumably converge if a starting point is chosen too small. The target function of the full problem (2.80) for fixed β and σ^2 also has a local optimum left of the global one, but combined with a very steep ascend left of θ^*, so numerical optimization methods will less likely converge to the local optimum instead to the global one. Note that for the computation of this target function, the optimal solutions β^*

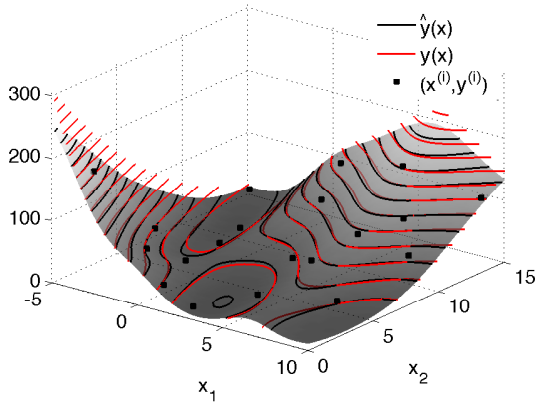

Figure 2.11: Kriging interpolation of the Branin test function.

and σ^{2*} have already to be known, but a similar steep ascend left of the global optimum can also be observed for other β and σ^2. Contrary to the cubic correlation function, the target functions for the biquadratic correlation function have only one optimum, see Fig. 2.10.

In a second example, the Branin test function

$$y(x_1, x_2) = \left(x_2 - \frac{5}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6\right)^2 + 10\left(1 - \frac{1}{8\pi}\right)\cos(x_1) + 10 \tag{2.88}$$

is interpolated in $\Omega = [-5, 10] \times [0, 15]$ using the cubic spline correlation function and a Latin hypercube sampling of size $N = 20$ (see Sect. 3.1.3). For the interpolation and the maximum likelihood problem, the x-data is transformed to $\widetilde{\Omega} = [0, 1]^2$ and retransformed afterwards. In Fig. 2.11, the interpolation $\widehat{y}(x)$ and the original response $y(x)$ are plotted. The target functions of the reduced likelihood problem (2.81) and the original problem (2.80) are both plotted for $\theta \in [0.05, 2.0]^2$ in Fig. 2.12. For the original problem, $\beta(\theta^*) = 1.27 \cdot 10^2$ and $\sigma^2(\theta^*, \beta(\theta^*)) = 1.21 \cdot 10^4$ are kept fixed again. The optimal solution is $\theta^* = (0.86, 0.49)$, which indicates that the $y(x^{(i)})$ are weaker correlated along the x_1-axis and stronger correlated along the x_2-axis.

The reduced problem's target function has altogether three local optima, one with $\theta_1 > {\theta_1}^*$ and one with $\theta_1 < {\theta_1}^*$. Numerical optimization methods starting in a large θ_1 will most likely converge to the first one. Starting in a θ close to zero, one will probably end up in the second one. The original problem with fixed β and σ^2 has only one optimum. Also, a very steep ascend of the target function for $\theta_k \to 0$ $(k = 1, 2)$ can be observed. However, there is no guarantee that the full problem will not have multiple local optima (like observed with the one-dimensional test function, cf. Fig. 2.9). But since nonlinearity is reduced and the target function behaves less erratic, it seems beneficial to keep β and σ^2 as variables in the optimization problem. Furthermore, the numerical effort for any approach is dominated by $\mathcal{O}(N^3)$ for $R(\theta)^{-1}$ in every iteration, so the increase in the number of sought variables from d to $d+K+1$ (here $K = 1$) is not a severe drawback. Further discussions on hyperparameter tuning strategies can for instance be found in [139, 140]. A brief description of the algorithm for maximum likelihood estimation used in this thesis is given in the numerical test cases

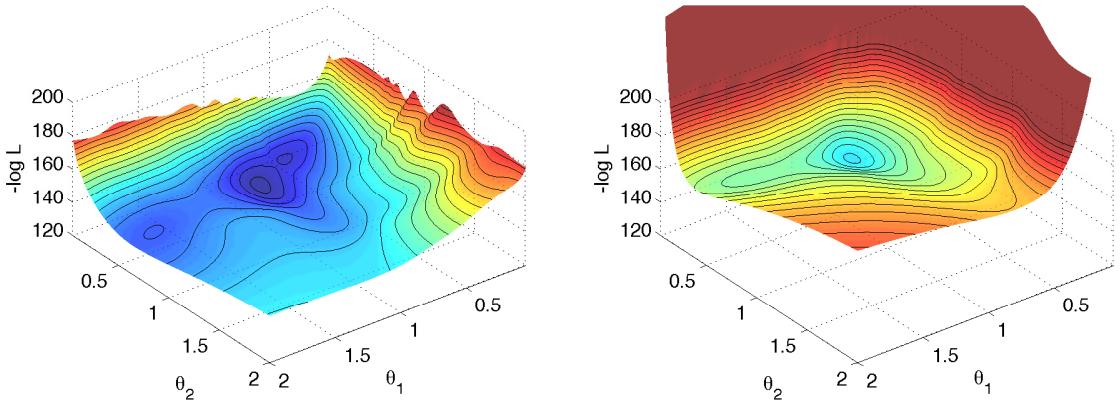

Figure 2.12: $-\log L$ for the data as in Fig. 2.11. Cubic spline correlation function. Left: reduced problem. Right: original problem (fixed β, σ^2).

section 4.1.

2.4 Gradient-enhanced Kriging

When not only the evaluations of the response but also its derivatives with respect to the input parameters are available, these can be used to improve the surrogate model. In Kriging, the values of the surrogate model $\widehat{y}(x^{(i)})$ are defined by the evaluations of the response in the samples $y(x^{(i)}) = y_i$ $(i = 1, \ldots, N)$. If additionally the partial derivatives $\frac{\partial}{\partial x_l}\widehat{y}(x^{(i)})$ are prescribed by the gradient information of the true response $\frac{\partial}{\partial x_l}y(x^{(i)})$ $(i = 1, \ldots, N,$ $l = 1, \ldots, d)$, an enhancement of the approximation quality is expected. Gradient information can always be approximated by finite differences. But especially when the response is based on the output of a computer experiment like a PDE solver, often the gradient can be computed more efficiently using automatic differentiation [50] or adjoint computations, see [91] for example. A comparison of Kriging and gradient-enhanced Kriging regarding efficiency (i.e. the amount of effort invested for generating information) is given in Sect. 4.1.

The gradient-enhanced Kriging (GEK) method was introduced in [99]. First estimations of the benefit of incorporating gradient information and of the computational costs of adjoint computations were given. Early reviews can be found in [73] and [84]. The authors of [17] interpreted the approach as a special case of Cokriging and used it for optimization in an aerodynamic test case, see e.g. the geostatistical textbooks [22, 146] for Cokriging. They also introduced a new indirect approach, which was also pursued in [86] and will be explained in the following. In [81], gradients were generated by a discrete adjoint CFD solver and adaptive sampling strategies for global approximation were compared for the gradient-enhanced Kriging method. Recently, the method was used for uncertainty quantification in aerodynamics [31] and in nuclear engineering [87], both exploiting the benefits of adjoint computations. Applications in aerodynamic design optimization using adjoints can be found in [79, 80, 150], while in the latter publication an extension to Hessian-enhanced Kriging based on second-order derivatives was developed. Most recently, GEK was utilized in a

variable fidelity-modeling framework for aero-loads prediction [54].

There are two main approaches of adding gradient information to the Kriging system: a direct one and an indirect one. In the direct formulation, the partial derivatives are incorporated by defining the surrogate model as a weighted sum of the response evaluations and the derivative evaluations. The weights are determined by modeling correlations between the response evaluations as well as the derivatives and solving an augmented linear system instead of the conventional Kriging system (2.38). In the indirect formulation, a new data pair

$$\left(x^{(i)} + \Delta x_l \cdot e_l,\ y(x^{(i)}) + \Delta x_l \cdot \frac{\partial y(x^{(i)})}{\partial x_l} \right) \tag{2.89}$$

for every partial derivative can be generated by using a Taylor series expansion as an approximation to the response value $y(x^{(i)} + \Delta x_l e_l)$, provided the evaluations $y(x^{(i)})$ and $\frac{\partial}{\partial x_l} y(x^{(i)})$ in $x^{(i)}$ are given ($i = 1, \ldots, N$, $l = 1, \ldots, d$). In contrast to the direct formulation, no modification of the Kriging framework is necessary. Hence, the surrogate model can be computed by any existing Kriging software which is at hand, once the additional samples (2.89) are generated. The indirect formulation demands additional parameters Δx_l which have to be tuned. Choosing Δx_l too large can produce poor approximations for $y(x^{(i)} + \Delta x_l e_l)$, while a small Δx_l yields very close samples $x^{(i)}, x^{(i)} + \Delta x_l e_l$, causing an ill-conditioned correlation matrix R. In direct GEK, no additional parameters have to be determined. Both approaches require an equal numerical effort and it has been observed that they often yield similar results, while direct GEK can be more accurate in some cases, cf. [17, 81, 150, 54]. Therefore, in this thesis the direct approach will be pursued.

The gradient-enhanced Kriging predictor is now deduced in the same way as the Kriging predictor in the last section by exploiting the properties of Gaussian processes. A detailed derivation is presented for the case that for each evaluation of the response $y_i = y(x^{(i)})$ the whole gradient $\nabla y(x^{(i)}) = (\frac{\partial}{\partial x_1} y(x^{(i)}), \ldots, \frac{\partial}{\partial x_d} y(x^{(i)}))^\top$ is available ($i = 1, \ldots, N$). The methodology can be extended to the case that arbitrary response and derivative evaluations are available in samples which do not have to coincide, see e.g. [54]. But in the most practical applications one does not have access to a partial derivative $\frac{\partial}{\partial x_l} y(x^{(i)})$ without knowing the response $y(x^{(i)})$. Its computation is needed for finite difference evaluations, automatic differentiation or discrete adjoint computation. However, the approach is straightforward from the one presented here and if not all derivatives are available for all $x^{(i)}$, the GEK predictor can be evaluated by simply deleting the corresponding vector entries, rows and columns in the involved matrices.

2.4.1 Model assumptions

Suppose for all samples $x^{(i)}$ ($i = 1, \ldots, N$) the evaluations of the response $y(x^{(i)}) = y_i$ and also all partial derivatives $\frac{\partial}{\partial x_l} y(x^{(i)})$ ($l = 1, \ldots, d$) are given. All data is assembled in a

vector Y^∂ which is defined as follows:

$$\begin{aligned} Y^\partial &:= \left(Y^{0\top}, Y^{\partial_1\top}, \ldots, Y^{\partial_d\top}\right)^\top \in \mathbb{R}^{(d+1)N}, \\ \text{where} \quad Y^0 &:= (y_1, \ldots, y_N)^\top \in \mathbb{R}^N \\ \text{and} \quad Y^{\partial_l} &:= \left(\frac{\partial y(x^{(1)})}{\partial x_l}, \ldots, \frac{\partial y(x^{(N)})}{\partial x_l}\right)^\top \in \mathbb{R}^N \quad (l = 1, \ldots, d). \end{aligned} \tag{2.90}$$

The GEK prediction $\widehat{y}(x)$ for $y(x)$ is defined by a weighted sum of both the evaluations of the response and the partial derivatives

$$\begin{aligned} \widehat{y}(x) &= \sum_{i=1}^{N} \lambda_i^0(x) y_i + \sum_{i=1}^{N} \sum_{l=1}^{d} \lambda_i^{\partial_l}(x) \frac{\partial y(x^{(i)})}{\partial x_l} \\ &= \lambda^\partial(x)^\top Y^\partial \end{aligned} \tag{2.91}$$

with weights

$$\begin{aligned} \lambda^\partial(x) &:= \left(\lambda^0(x)^\top, \lambda^{\partial_1}(x)^\top, \ldots, \lambda^{\partial_d}(x)^\top\right)^\top \in \mathbb{R}^{(d+1)N}, \\ \text{where} \quad \lambda^0(x) &\in \mathbb{R}^N \\ \text{and} \quad \lambda^{\partial_l}(x) &\in \mathbb{R}^N \quad (l = 1, \ldots, d), \end{aligned} \tag{2.92}$$

depending on $x \in \Omega$. Now statistical assumptions on the unknown response $y(x)$ and its derivatives $\frac{\partial}{\partial x_l} y(x)$ $(l = 1, \ldots, d)$ are used to determine the weights $\lambda^\partial(x) \in \mathbb{R}^{(d+1)N}$, for which the linear predictor $\widehat{y}(x)$ is unbiased and the mean squared error $\mathrm{MSE}[\widehat{y}(x)] = \mathrm{E}[(\widehat{y}(x) - y(x))^2]$ is minimal (BLUP).

As in Kriging, the response is modeled by a linear regression term and a Gaussian process $z(x)$

$$y(x) = \sum_{j=1}^{K} \beta_j f_j(x) + z(x), \quad x \in \Omega \subset \mathbb{R}^d. \tag{2.93}$$

Its partial derivatives then read

$$\frac{\partial y(x)}{\partial x_l} = \sum_{j=1}^{K} \beta_j \frac{\partial f_j(x)}{\partial x_l} + \frac{\partial z(x)}{\partial x_l}, \quad x \in \Omega \subset \mathbb{R}^d, \ l = 1, \ldots, d. \tag{2.94}$$

The properties of $\frac{\partial}{\partial x_l} z(x)$ are now discussed. Generally, any derivative $z^{(a_1,\ldots,a_d)}(x) := \frac{\partial^{\Sigma a_k}}{\partial x_1^{a_1} \ldots \partial x_d^{a_d}} z(x)$ defines a new Gaussian process, cf. [99]. The expectation and covariance of $z(x)$ were defined by $\mathrm{E}[z(x)] = 0$, $\mathrm{Cov}[z(x), z(\tilde{x})] = \sigma^2 R(h)$, $h = \tilde{x} - x$, cf. (2.28). Since $\mathrm{E}[z(x)]$ is constant in Ω (stationarity of $z(x)$),

$$\mathrm{E}\left[\frac{\partial z(x)}{\partial x_l}\right] = 0 \quad \forall x \in \Omega, \ l = 1, \ldots, d \tag{2.95}$$

is deduced. If a product model (2.71) is used for the correlation function $R(h) = \prod_{k=1}^{d} R_k(h_k)$, the covariance reads

$$\mathrm{Cov}\left[z^{(a_1,\ldots,a_d)}(x), z^{(b_1,\ldots,b_d)}(\tilde{x})\right] = \sigma^2(-1)^{\Sigma a_k} \prod_{k=1}^{d} R_k^{(a_k+b_k)}(h_k) \quad \forall x, \tilde{x} \in \Omega, \tag{2.96}$$

see [99]. From this representation, the covariances of the zeroth and first order partial derivatives can be deduced:

$$\mathrm{Cov}\left[z(x), z(\tilde{x})\right] = \sigma^2 R(h) \tag{2.97}$$

$$\mathrm{Cov}\left[z(x), \frac{\partial z(\tilde{x})}{\partial x_l}\right] = \sigma^2 \frac{\partial}{\partial h_l} R(h) \tag{2.98}$$

$$\mathrm{Cov}\left[\frac{\partial z(x)}{\partial x_l}, z(\tilde{x}),\right] = -\sigma^2 \frac{\partial}{\partial h_l} R(h) \tag{2.99}$$

$$\mathrm{Cov}\left[\frac{\partial z(x)}{\partial x_k}, \frac{\partial z(\tilde{x})}{\partial x_l}\right] = -\sigma^2 \frac{\partial^2}{\partial h_k \partial h_l} R(h), \tag{2.100}$$

see [73] for example. The covariances (2.98)–(2.100) are also called cross-covariances, because they denote the covariance of two different stochastic processes. All covariances are symmetric again, which can easily be shown using the symmetry properties of the one-dimensional correlation function $R_k(h_k) = R(-h_k)$, $\frac{\partial}{\partial h_k} R(h_k) = -\frac{\partial}{\partial h_k} R(-h_k)$ and $\frac{\partial^2}{\partial h_k^2} R(h_k) = \frac{\partial^2}{\partial h_k^2} R(-h_k)$, $h_k = \tilde{x}_k - x_k$ in the product model:

$$\begin{aligned}
\mathrm{Cov}\left[z(x), \frac{\partial z(\tilde{x})}{\partial x_l}\right] &= \sigma^2 \frac{\partial}{\partial h_l} R(h) \\
&= -\sigma^2 \frac{\partial}{\partial h_l} R(-h) \\
&= \mathrm{Cov}\left[\frac{\partial z(\tilde{x})}{\partial x_l}, z(x)\right] \quad \forall l, \qquad (2.101) \\
\mathrm{Cov}\left[\frac{\partial z(x)}{\partial x_k}, \frac{\partial z(\tilde{x})}{\partial x_l}\right] &= -\sigma^2 \frac{\partial^2}{\partial h_k \partial h_l} R(h) \\
&= -\sigma^2 \frac{\partial^2}{\partial h_l \partial h_k} R(-h)(-1)^2 \\
&= \mathrm{Cov}\left[\frac{\partial z(\tilde{x})}{\partial x_l}, \frac{\partial z(x)}{\partial x_k}\right] \quad \forall l \neq k, \qquad (2.102) \\
\mathrm{Cov}\left[\frac{\partial z(x)}{\partial x_l}, \frac{\partial z(\tilde{x})}{\partial x_l}\right] &= -\sigma^2 \frac{\partial^2}{\partial h_l^2} R(h) \\
&= -\sigma^2 \frac{\partial^2}{\partial h_l^2} R(-h) \\
&= \mathrm{Cov}\left[\frac{\partial z(\tilde{x})}{\partial x_l}, \frac{\partial z(x)}{\partial x_l}\right] \quad \forall l. \qquad (2.103)
\end{aligned}$$

Having discussed that both $z(x)$ and all $\frac{\partial}{\partial x_l} z(x)$ $(l = 1, \ldots, d)$ are stationary Gaussian

processes which are correlated with each other, yields now that both $y(x)$ and all $\frac{\partial}{\partial x_l} y(x)$ $(l = 1, \ldots, d)$ are (nonstationary) Gaussian processes. From equation (2.94) one concludes

$$\mathrm{E}\left[\frac{\partial y(x)}{\partial x_l}\right] = \sum_{j=1}^{K} \beta_j \frac{\partial f_j(x)}{\partial x_l} \quad \forall x \in \Omega. \tag{2.104}$$

Since the addition of the first (deterministic) summand in (2.94) does not affect the covariance, the covariance of $\frac{\partial}{\partial x_l} y(x)$ is the same as of $\frac{\partial}{\partial x_l} z(x)$ in (2.97)–(2.100). For given samples $X = \{x^{(1)}, \ldots, x^{(N)}\}$, a vector Z^{∂} is defined analogously to Y^{∂} in (2.90):

$$\begin{aligned} Z^{\partial} &:= \left(Z^{0\top}, Z^{\partial_1\top}, \ldots, Z^{\partial_d\top}\right)^{\top} \in \mathbb{R}^{(d+1)N}, \\ \text{where} \quad Z^0 &:= \left(z(x^{(1)}), \ldots, z(x^{(N)})\right)^{\top} \in \mathbb{R}^N \\ \text{and} \quad Z^{\partial_l} &:= \left(\frac{\partial z(x^{(1)})}{\partial x_l}, \ldots, \frac{\partial z(x^{(N)})}{\partial x_l}\right)^{\top} \in \mathbb{R}^N \quad (l = 1, \ldots, d). \end{aligned} \tag{2.105}$$

Also, similarly

$$\begin{aligned} F^{\partial} &:= \left[F^{0\top} \; F^{\partial_1\top} \; \cdots \; F^{\partial_d\top}\right]^{\top} \in \mathbb{R}^{(d+1)N \times K}, \\ \text{where} \quad F^0 &:= \left[f_j(x^{(i)})\right]_{i,j=1}^{N,K} \in \mathbb{R}^{N \times K} \\ \text{and} \quad F^{\partial_l} &:= \left[\frac{\partial}{\partial x_l} f_j(x^{(i)})\right]_{i,j=1}^{N,K} \in \mathbb{R}^{N \times K} \quad (l = 1, \ldots, K). \end{aligned} \tag{2.106}$$

Furthermore, the symmetric correlation matrix of Z^{∂} is given by

$$R^{\partial} := \begin{bmatrix} R^{00} & R^{0\partial_1} & \cdots & R^{0\partial_d} \\ R^{0\partial_1\top} & R^{\partial_1\partial_1} & \cdots & R^{\partial_1\partial_d} \\ \vdots & \vdots & \ddots & \vdots \\ R^{0\partial_d\top} & R^{\partial_d\partial_1} & \cdots & R^{\partial_d\partial_d} \end{bmatrix} \in \mathbb{R}^{(d+1)N \times (d+1)N} \tag{2.107}$$

and the submatrices are defined as follows:

$$\begin{aligned} R^{00} &:= \left[R(x^{(j)} - x^{(i)})\right]_{i,j=1}^{N,N} \in \mathbb{R}^{N \times N}, \\ R^{0\partial_l} &:= \left[\frac{\partial}{\partial h_l} R(x^{(j)} - x^{(i)})\right]_{i,j=1}^{N,N} \in \mathbb{R}^{N \times N} \quad (l = 1, \ldots, d), \\ R^{\partial_l \partial_k} &:= \left[-\frac{\partial^2}{\partial h_l \partial h_k} R(x^{(j)} - x^{(i)})\right]_{i,j=1}^{N,N} \in \mathbb{R}^{N \times N} \quad (k, l = 1, \ldots, d). \end{aligned}$$

Eventually, after having defined all required expressions, one can write $Y^{\partial} = F^{\partial}\beta + Z^{\partial}$ and

both Y^∂ and Z^∂ are multivariate normally distributed

$$Y^\partial \sim \mathcal{N}_{(d+1)N}(F^\partial \beta, \sigma^2 R^\partial), \tag{2.108}$$
$$Z^\partial \sim \mathcal{N}_{(d+1)N}(0, \sigma^2 R^\partial). \tag{2.109}$$

Also, the correlation matrix R^∂ is assumed to be positive definite

$$w^\top R^\partial w > 0 \quad \forall x^{(i)} \neq x^{(j)} \ (i \neq j), \ w \in \mathbb{R}^{(d+1)N} \setminus \{0\}, \tag{2.110}$$

see Sect. 2.4.3 for a discussion on suitable correlation functions.

2.4.2 The GEK predictor

The following theorem provides a necessary and sufficient condition for the weights λ^∂, for which the GEK predictor (2.91) is a best linear unbiased predictor (BLUP). It is deduced analogously to the condition for the Kriging predictor in Theorem 2.4. Let $r^\partial(x) \in \mathbb{R}^{(d+1)N}$ be the correlation vector containing the correlations of $y(x)$ and all $y(x^{(i)})$ and $\frac{\partial}{\partial x_l} y(x^{(i)})$ $(i = 1, \ldots, N, \ l = 1, \ldots, d)$:

$$r^\partial(x) := \left(r^0(x)^\top, r^{\partial_1}(x)^\top, \ldots, r^{\partial_d}(x)^\top\right)^\top \in \mathbb{R}^{(d+1)N}, \tag{2.111}$$
$$\text{where} \quad r^0(x) := \left(R(x^{(i)} - x)\right)_{i=1}^N \in \mathbb{R}^N$$
$$\text{and} \quad r^{\partial_l}(x) := \left(\frac{\partial}{\partial h_l} R(x^{(i)} - x)\right)_{i=1}^N \in \mathbb{R}^N \quad (l = 1, \ldots, d).$$

All other expressions are defined as before.

Theorem 2.8. *Let $x^{(i)} \neq x^{(j)}$ for $i \neq j$ and F^∂ have full column rank K. Provided $\binom{\lambda^\partial(x)}{\mu(x)} \in \mathbb{R}^{(d+1)N+K}$ is the solution of the linear system*

$$\begin{bmatrix} R^\partial & F^\partial \\ F^{\partial\top} & 0 \end{bmatrix} \begin{pmatrix} \lambda^\partial(x) \\ \mu(x) \end{pmatrix} = \begin{pmatrix} r^\partial(x) \\ f(x) \end{pmatrix}, \tag{2.112}$$

the GEK predictor $\widehat{y}(x) = \lambda^\partial(x)^\top Y^\partial$ is a BLUP for $y(x)$, $x \in \Omega$.

Proof. The proof is straightforward from the proof in the Kriging case. The GEK predictor is BLUP for $y(x)$ for arbitrary $x \in \Omega$, if $\lambda^\partial(x)$ is the solution of the constrained optimization problem

$$\min_{\lambda^\partial(x) \in \mathbb{R}^{(d+1)N}} \ \mathrm{E}\left[\left(\lambda^\partial(x)^\top Y^\partial - y(x)\right)^2\right] \tag{2.113}$$
$$\text{s.t.} \ \mathrm{E}\left[\lambda^\partial(x)^\top Y^\partial - y(x)\right] = 0. \tag{2.114}$$

Since $\mathrm{E}[z(x)] = 0$ and $\mathrm{E}[Z^\partial] = 0$, the unbiasedness condition (2.114) reads

$$\begin{aligned} 0 &= \mathrm{E}\left[\lambda^\partial(x)^\top Y^\partial - y(x)\right] \\ &= \mathrm{E}\left[\lambda^\partial(x)^\top (F^\partial \beta + Z^\partial) - f(x)^\top \beta - z(x)\right] \\ &= \lambda^\partial(x)^\top F^\partial \beta + \lambda^\partial(x)^\top \mathrm{E}[Z^\partial] - f(x)^\top \beta - \mathrm{E}[z(x)] \\ &= \left(F^{\partial^\top} \lambda^\partial(x) - f(x)\right)^\top \beta \end{aligned} \tag{2.115}$$

for all $\beta \in \mathbb{R}^K$, which yields

$$F^{\partial^\top} \lambda^\partial(x) - f(x) = 0. \tag{2.116}$$

Then, the error is rewritten as

$$\begin{aligned} \widehat{y}(x) - y(x) &= \lambda^\partial(x)^\top (F^\partial \beta + Z^\partial) - f(x)^\top \beta - z(x) \\ &= \left(F^{\partial^\top} \lambda^\partial(x) - f(x)\right)^\top \beta + \lambda^\partial(x)^\top Z^\partial - z(x) \\ &= \lambda^\partial(x)^\top Z^\partial - z(x). \end{aligned} \tag{2.117}$$

This identity and the covariance formulas (2.97)–(2.100) are exploited for the calculation of the $\mathrm{MSE}[\widehat{y}(x)]$:

$$\begin{aligned} \mathrm{E}\left[\left(\lambda^\partial(x)^\top Y^\partial - y(x)\right)^2\right] &= \mathrm{E}\left[\left(\lambda^\partial(x)^\top Z^\partial - z(x)\right)^2\right] \\ &= \mathrm{E}\left[\lambda^\partial(x)^\top Z^\partial Z^{\partial^\top} \lambda^\partial(x) - 2\lambda^\partial(x)^\top Z^\partial z(x) + z(x)^2\right] \\ &= \sigma^2 \left(\lambda^\partial(x)^\top R^\partial \lambda^\partial(x) - 2\lambda^\partial(x)^\top r^\partial(x) + 1\right). \end{aligned} \tag{2.118}$$

With the rewritten $\mathrm{MSE}[\widehat{y}(x)]$ (2.118) and unbiasedness condition (2.116), the original problem (2.113)–(2.114) is equivalent to

$$\min_{\lambda^\partial(x) \in \mathbb{R}^{(d+1)N}} \frac{1}{2} \lambda^\partial(x)^\top R^\partial \lambda^\partial(x) - r^\partial(x)^\top \lambda^\partial(x) \tag{2.119}$$

$$\text{s.t.} \quad F^{\partial^\top} \lambda^\partial(x) - f(x) = 0. \tag{2.120}$$

With a Lagrange variable $\mu(x) \in \mathbb{R}^K$, the Karush-Kuhn-Tucker system for this convex and quadratic optimization problem with linear constraints reads

$$\begin{bmatrix} R^\partial & F^\partial \\ F^{\partial^\top} & 0 \end{bmatrix} \begin{pmatrix} \lambda^\partial(x) \\ \mu(x) \end{pmatrix} = \begin{pmatrix} r^\partial(x) \\ f(x) \end{pmatrix}. \tag{2.121}$$

By positive definiteness of R^∂ and full column rank of F^∂ the nonsingularity of the system matrix is guaranteed and (2.121) is a necessary and sufficient optimality condition, see [103]. □

Properties

Just like the Kriging predictor in (2.50), the GEK predictor can be expanded

$$\lambda^{\partial}(x)^{\top} Y^{\partial} = \begin{pmatrix} \lambda^{\partial}(x) \\ \mu(x) \end{pmatrix}^{\top} \begin{pmatrix} Y^{\partial} \\ 0 \end{pmatrix} = \begin{pmatrix} r^{\partial}(x) \\ f(x) \end{pmatrix}^{\top} \begin{bmatrix} R^{\partial} & F^{\partial} \\ F^{\partial\top} & 0 \end{bmatrix}^{-1} \begin{pmatrix} Y^{\partial} \\ 0 \end{pmatrix} \tag{2.122}$$

and the solution $\begin{pmatrix} w^{(r^{\partial})} \\ w^{(f)} \end{pmatrix} \in \mathbb{R}^{(d+1)N+K}$ of the linear equation

$$\begin{bmatrix} R^{\partial} & F^{\partial} \\ F^{\partial\top} & 0 \end{bmatrix} \begin{pmatrix} w^{(r^{\partial})} \\ w^{(f)} \end{pmatrix} = \begin{pmatrix} Y^{\partial} \\ 0 \end{pmatrix} \tag{2.123}$$

is independent of x. For arbitrary $x \in \Omega$, the GEK predictor can be evaluated at the cost of a scalar product $\mathcal{O}\left((d+1)N + K\right)$

$$\widehat{y}(x) = \begin{pmatrix} w^{(r^{\partial})} \\ w^{(f)} \end{pmatrix}^{\top} \begin{pmatrix} r^{\partial}(x) \\ f(x) \end{pmatrix}. \tag{2.124}$$

This is a weighted sum of $R(x^{(i)} - x)$, $\frac{\partial}{\partial h_l} R(x^{(i)} - x)$ $(l = 1, \ldots, d;\ i = 1, \ldots, N)$ and $f_j(x)$ $(j = 1, \ldots, K)$.

The GEK system (2.112) serves also as an interpolation condition both for all $y(x^{(i)})$ and all $\frac{\partial}{\partial x_l} y(x^{(i)})$ which is now shown. For the former, inserting any $x^{(i)}$ $(i = 1, \ldots, N)$ for x into the equation yields

$$\left(r^{\partial}(x^{(i)})\right)_j = R(x^{(j)} - x^{(i)}) = \left[R^{00}\right]_{i,j} = \left[R^{00}\right]_{j,i} = \left[R^{\partial}\right]_{j,i}$$

$$\text{and} \quad \left(r^{\partial}(x^{(i)})\right)_{Nl+j} = \frac{\partial}{\partial h_l} R(x^{(j)} - x^{(i)}) = \left[R^{0\partial_l}\right]_{i,j} = \left[R^{0\partial_l\top}\right]_{j,i} = \left[R^{\partial}\right]_{Nl+j,i}$$

for $j = 1, \ldots, N$ and $l = 1, \ldots, d$, cf. the definition of the correlation matrix (2.107) and the correlation vector (2.111). Also,

$$f_j(x^{(i)}) = \left[F^{0}\right]_{i,j} = \left[F^{\partial}\right]_{i,j} = \left[F^{\partial\top}\right]_{j,i}$$

holds for $j = 1, \ldots, K$. Thus, the right-hand side $\begin{pmatrix} r^{\partial}(x^{(j)}) \\ f(x^{(j)}) \end{pmatrix}$ in (2.112) is identical to the i-th column (and row) of $\begin{bmatrix} R^{\partial} & F^{\partial} \\ F^{\partial\top} & 0 \end{bmatrix}$ and $\lambda^{\partial}(x^{(i)}) = e_i$, the i-th unit vector. Hence, the true response y_i is correctly interpolated in $x^{(i)}$

$$\widehat{y}(x^{(i)}) = e_i^{\top} Y^{\partial} = y_i. \tag{2.125}$$

For the verification of the interpolation property of the gradients, the partial derivatives of the surrogate model

$$\frac{\partial}{\partial x_l} \widehat{y}(x) = \begin{pmatrix} w^{(r^{\partial})} \\ w^{(f)} \end{pmatrix}^{\top} \begin{pmatrix} \frac{\partial}{\partial x_l} r^{\partial}(x) \\ \frac{\partial}{\partial x_l} f(x) \end{pmatrix} = \begin{pmatrix} Y^{\partial} \\ 0 \end{pmatrix}^{\top} \begin{bmatrix} R^{\partial} & F^{\partial} \\ F^{\partial\top} & 0 \end{bmatrix}^{-1} \begin{pmatrix} \frac{\partial}{\partial x_l} r^{\partial}(x) \\ \frac{\partial}{\partial x_l} f(x) \end{pmatrix} \tag{2.126}$$

($l = 1, \ldots, d$) are calculated. Keeping in mind that $\frac{\partial}{\partial h_l}$ is the partial derivative with respect to the l-th argument $h_l = x_l^{(i)} - x_l$ of $R(h)$, the chain rule has to be applied for the calculation of $\frac{\partial}{\partial x_l}$:

$$\begin{aligned}
\frac{\partial}{\partial x_l}\left(r^{\partial}(x)\right)_j &= \frac{\partial}{\partial x_l} R(x^{(j)} - x) \\
&= \frac{\partial}{\partial h_l} R(x^{(j)} - x) \cdot \frac{\partial}{\partial x_l}(x_l^{(j)} - x_l) \\
&= -\frac{\partial}{\partial h_l} R(x^{(j)} - x) \quad (j = 1, \ldots, N) \\
\text{and} \quad \frac{\partial}{\partial x_l}\left(r^{\partial}(x)\right)_{Nk+j} &= \frac{\partial}{\partial x_l}\left(\frac{\partial}{\partial h_k} R(x^{(j)} - x)\right) \\
&= \frac{\partial^2}{\partial h_l \partial h_k} R(x^{(j)} - x) \cdot \frac{\partial}{\partial x_l}(x_l^{(j)} - x_l) \\
&= -\frac{\partial^2}{\partial h_l \partial h_k} R(x^{(j)} - x) \quad (j = 1, \ldots, N;\ k = 1, \ldots, d).
\end{aligned}$$

Now inserting any $x^{(i)}$ for x,

$$\begin{aligned}
\frac{\partial}{\partial x_l}\left(r^{\partial}(x^{(i)})\right)_j &= -\frac{\partial}{\partial h_l} R(x^{(j)} - x^{(i)}) \\
&= \frac{\partial}{\partial h_l} R(x^{(i)} - x^{(j)}) \\
&= \left[R^{0\partial_l}\right]_{j,i} \\
&= \left[R^{\partial}\right]_{j,Nl+i}, \\
\frac{\partial}{\partial x_l}\left(r^{\partial}(x^{(i)})\right)_{Nk+j} &= -\frac{\partial^2}{\partial h_l \partial h_k} R(x^{(j)} - x^{(i)}) \\
&= -\frac{\partial^2}{\partial h_l \partial h_k} R(x^{(i)} - x^{(j)}) \cdot (-1)^2 \\
&= \left[R^{\partial_k \partial_l}\right]_{j,i} \\
&= \left[R^{\partial}\right]_{Nk+j,Nl+i}, \\
\frac{\partial}{\partial x_l}\left(r^{\partial}(x^{(i)})\right)_{Nl+j} &= -\frac{\partial^2}{\partial h_l^2} R(x^{(j)} - x^{(i)}) \\
&= -\frac{\partial^2}{\partial h_l^2} R(x^{(i)} - x^{(j)}) \\
&= \left[R^{\partial_l \partial_l}\right]_{j,i} \\
&= \left[R^{\partial}\right]_{Nl+j,Nl+i}
\end{aligned}$$

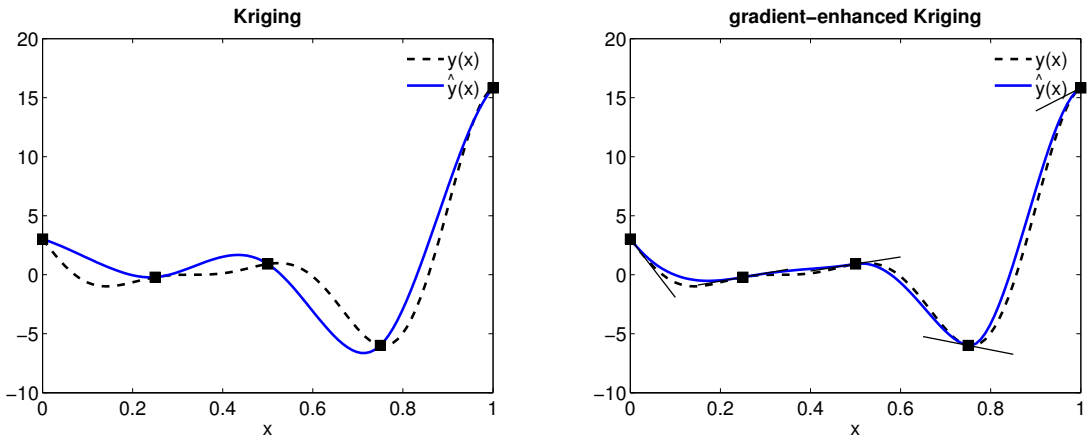

Figure 2.13: Kriging (left) and GEK (right) interpolations of $y(x) = (6x-2)^2 \sin(12x-4)$.

$(j = 1, \ldots, N;\, k = 1, \ldots, d,\, k \neq l)$. Also,

$$\frac{\partial}{\partial x_l} f_j(x^{(i)}) = \left[F^{\partial_l}\right]_{i,j} = \left[F^{\partial}\right]_{Nl+i,j} = \left[F^{\partial^\top}\right]_{j,Nl+i}$$

for $j = 1, \ldots, K$, consequently $\begin{pmatrix} \frac{\partial}{\partial x_l} r^{\partial}(x^{(i)}) \\ \frac{\partial}{\partial x_l} f(x^{(i)}) \end{pmatrix}$ is identical to the $Nl+i$-th column (and row) of $\begin{bmatrix} R^{\partial} & F^{\partial} \\ F^{\partial^\top} & 0 \end{bmatrix}$ in (2.126). The partial derivative of the predictor in $x^{(i)}$

$$\frac{\partial}{\partial x_l} \widehat{y}(x^{(i)}) = Y^{\partial^\top} e_{Nl+i} = (Y^{\partial})_{Nl+i} = \frac{\partial}{\partial x_l} y(x^{(i)}) \tag{2.127}$$

yields the correct partial derivative of the response in $x^{(i)}$.

Example

The benefit of incorporating gradient information in the Kriging surrogate is highlighted with a simple example. Again,

$$y(x) = (6x-2)^2 \sin(12x-4) \tag{2.128}$$

from [41] serves as a one-dimensional test case. The function is interpolated with Kriging and gradient-enhanced Kriging in $\Omega = [0,1]$ based on only $N = 5$ equidistant samples. As a correlation function, the biquadratic spline (2.72) is chosen. The hyperparameter estimation yields $\theta = 1.30$ in Kriging and $\theta = 1.45$ in GEK. The results are depicted in Fig. 2.13. While the Kriging interpolation is a poor surrogate with an average error of $\int_0^1 |y(x) - \widehat{y}(x)| dx = 1.3$, the GEK interpolation performs much better. The global behavior is represented more accurately by interpolating the true response as well as matching the gradients in the samples $x^{(i)}$ $(i = 1, \ldots, 5)$. The average error $\int_0^1 |y(x) - \widehat{y}(x)| dx = 0.4$ is less than a third of the Kriging error.

2.4.3 Correlation functions

In Sect. 2.3.3, the topic of correlation functions of stochastic processes for Kriging was discussed. With the characterization by Fourier transformations, a necessary and sufficient condition for the validness of $R(h)$ was presented. Also, an extension of the condition was given which not only guarantees the positive semidefiniteness, but actually the positive definiteness of any corresponding correlation matrix R, which is sufficient for the solvability of the Kriging equation (2.38). Within this section, requirements for correlation functions for GEK are discussed, which are not straightforward from the Kriging case.

First of all, a univariate correlation function must be at least twice differentiable in $\mathbb{R}$. Though only first order derivatives of the response are considered, in the covariance formulation (2.96) second order derivatives appear:

$$\operatorname{Cov}\left[\frac{\partial z(x)}{\partial x_l}, \frac{\partial z(\tilde{x})}{\partial x_l}\right] = -\sigma^2 R_l''(h_l) \prod_{\substack{k=1 \\ k \neq l}}^{d} R_k(h_k). \tag{2.129}$$

The GEK predictor $\widehat{y}(x)$ (2.124) is a weighted sum of $R(x^{(i)} - x)$, $\frac{\partial}{\partial x_l} R(x^{(i)} - x)$ $(i = 1, \ldots, N;$ $l = 1, \ldots, d)$ and $f_j(x)$ $(j = 1, \ldots, K)$, so here only first order partial derivatives appear. This implies that if $\widehat{y}(x) \in C^n(\mathbb{R}^d)$ is desired, the univariate correlation function is required to satisfy $R(h) \in C^{n+1}(\mathbb{R})$.

For example, the first and second order derivatives of the biquadratic spline correlation function (2.72) are

$$R'(h, \theta) = \begin{cases} \left(-30(\theta|h|) + 105(\theta|h|)^2 - \frac{195}{2}(\theta|h|)^3\right) \theta \operatorname{sgn}(h), & 0 \leq \theta|h| < 0.4 \\ \left(-\frac{20}{3} + 20(\theta|h|) - 20(\theta|h|)^2 + \frac{20}{3}(\theta|h|)^3\right) \theta \operatorname{sgn}(h), & 0.4 \leq \theta|h| < 1 \\ 0, & 1 \leq \theta|h| \end{cases} \tag{2.130}$$

$$R''(h, \theta) = \begin{cases} \left(-30 + 210(\theta|h|) - \frac{585}{2}(\theta|h|)^2\right) \theta^2, & 0 \leq \theta|h| < 0.4 \\ \left(20 - 40(\theta|h|) + 20(\theta|h|)^2\right) \theta^2, & 0.4 \leq \theta|h| < 1 \\ 0, & 1 \leq \theta|h| \end{cases} \tag{2.131}$$

They are depicted for $\theta = 1$ in Fig. 2.14. The correlation function satisfies $R(h) \in C^2(\mathbb{R})$ and $R(h) \in C^3(\mathbb{R} \setminus \{0\})$ for all $\theta \in \mathbb{R}_+$, so for the GEK predictor $\widehat{y}(x) \in C^1(\mathbb{R}^d)$ follows.

A second requirement for a valid correlation function is the positive semidefiniteness of the correlation matrix R^{∂} (2.107) and, in order to ensure the existence of a solution of the GEK system (2.112), even the positive definiteness. The topic of positive definiteness will not be covered here but, exemplarily, an extension of the arguments for positive semidefiniteness in the Kriging case to GEK is now presented. A Cokriging framework is used, i.e. the response and its partial derivatives are each understood as a Gaussian process which possesses an auto-correlation and cross-correlations. Let the $d + 1$ correlated Gaussian processes

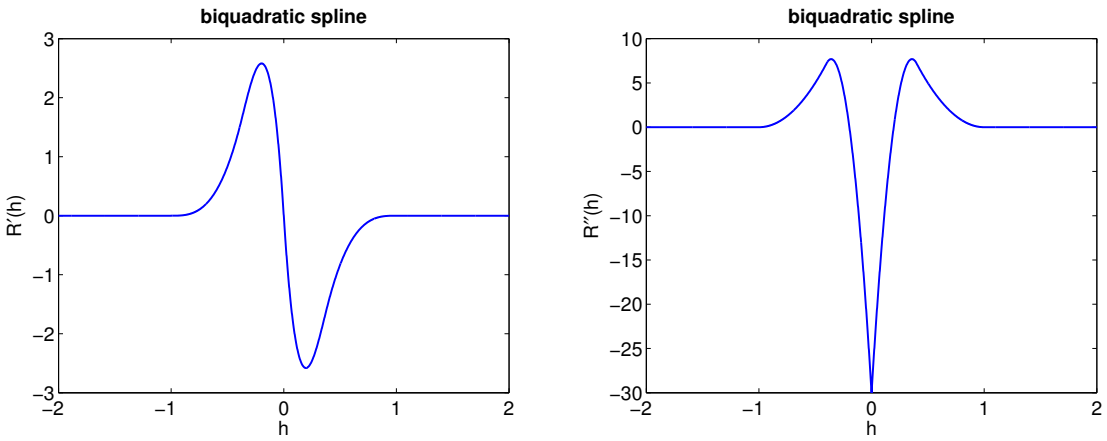

Figure 2.14: First and second derivative of the biquadratic spline correlation function, $\theta = 1$.

$\{z(x), \frac{\partial}{\partial x_1} z(x), \ldots, \frac{\partial}{\partial x_d} z(x)\}$ be identified by $\{z_0(x), z_1(x), \ldots, z_d(x)\}$ for simplicity. Then

$$\begin{aligned} \mathrm{E}\left[z_l(x)\right] &= 0, \\ \mathrm{Cov}\left[z_l(x), z_k(\tilde{x})\right] &= \sigma^2 R_{kl}(\tilde{x} - x) \quad \forall x, \tilde{x} \in \Omega,\ k, l = 0, \ldots, d, \end{aligned}$$

cf. (2.95), and the $R_{kl}(h)$ are defined by the direct and the cross-correlations (2.97)–(2.100). The positive semidefiniteness of R^{∂} is then equivalent to

$$\sum_{i,j=1}^{N} \sum_{k,l=0}^{d} w_{Nk+i} w_{Nl+j} R_{kl}(x^{(j)} - x^{(i)}) \geq 0 \quad \forall w \in \mathbb{R}^{(d+1)N}. \tag{2.132}$$

Following Wackernagel [146], the validness of $R_{kl}(h)$ $(k, l = 0, \ldots, d)$ as direct and cross-correlation functions can be characterized by an extension of Bochner's theorem by Cramér [20]. Thus, $R_{kl}(h)$ $(k, l = 0, \ldots, d)$ are valid correlation functions for $\{z_0(x), z_1(x), \ldots, z_d(x)\}$ (implying (2.132)), if and only if they have the spectral representation

$$R_{kl}(h) = \int_{\mathbb{R}^d} \exp\{i\omega^\top h\}\, dS_{kl}(\omega), \tag{2.133}$$

where the $S_{kl}(\omega)$ $(k, l = 0, \ldots, d)$ are spectral distribution functions, whose matrices of increments $[\Delta S_{kl}(\omega)]_{k,l=0}^{d,d} \in \mathbb{C}^{(d+1)\times(d+1)}$ are Hermitian and positive semidefinite for all increments $\Delta\omega \in \mathbb{R}^d$ (with $\Delta S_{kl}(\omega)$ and $\Delta\omega$ defined as in [149], Sect. 16). If the spectral density functions $s_{kl}(\omega)$ exist, the condition is equivalent to $[\int_I s_{kl}(\omega)\, d\omega]_{k,l=0}^{d,d} \in \mathbb{C}^{(d+1)\times(d+1)}$ being a Hermitian positive semidefinite matrix for all d-dimensional intervals $I := [a_1, b_1] \times \ldots \times [a_d, b_d] \subset \mathbb{R}^d$ $(a_k < b_k)$. This especially implies that all $s_{kk}(w)$ are real and nonnegative functions $(k = 0, \ldots, d)$, whereas the off-diagonal terms $s_{kl}(\omega)$ $(k \neq l)$ can also be complex-valued.

As in Sect. 2.3.3, this condition of positive semidefiniteness is now illustrated again with the biquadratic spline correlation function (2.72) and its first and second derivative for $d = 1$, i.e. its validness as a correlation function for $\{z_0(x), z_1(x)\} = \{z(x), z'(x)\}$ is reviewed. Again,

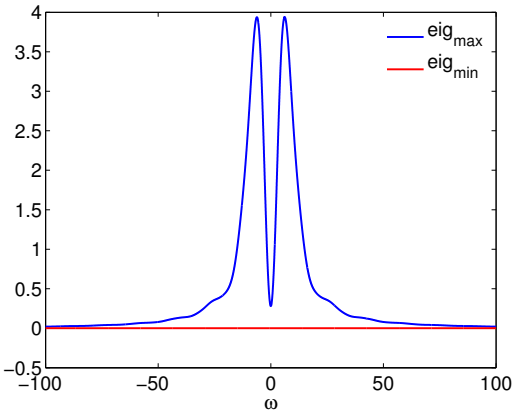

Figure 2.15: Eigenvalues of the matrix (2.134) depending on ω for the biquadratic spline correlation function, $\theta = 1$.

only the case $\theta = 1$ is considered here without loss of generality. $R(h)$, $R'(h)$ and $R''(h)$ are continuous, bounded and have a compact support. Generally, the case $d = 1$ yields $R_{00}(h) = R(h)$, $R_{01}(h) = R'(h)$, $R_{10}(h) = -R'(h)$ and $R_{11}(h) = -R''(h)$. Also, the former condition is equivalent to the positive semidefiniteness of the Hermitian matrix assembled by the density functions

$$\begin{bmatrix} s_{00}(\omega) & s_{01}(\omega) \\ s_{10}(\omega) & s_{11}(\omega) \end{bmatrix} \tag{2.134}$$

for all $\omega \in \mathbb{R}$ (cf. [149], Sect. 15). Calculating the spectral densities as the Fourier transforms of the (cross-) correlation functions (cf. Sect. 2.3.3)

$$s_{kl}(\omega) = \frac{1}{2\pi} \int_{\mathbb{R}} \exp\{-i\omega h\}\ R_{kl}(h)\ dh$$

and exploiting the symmetry properties $R(h) = R(-h)$, $R'(h) = -R'(-h)$ and $R''(h) = R''(-h)$ yields

$$\begin{aligned}
s_{00}(\omega) &= \frac{1}{2\pi} \int_{\mathbb{R}} \exp\{-i\omega h\}\ R(h)\ dh = \frac{1}{\pi} \int_0^\infty \cos(\omega h) R(h)\ dh, \\
s_{01}(\omega) &= \frac{1}{2\pi} \int_{\mathbb{R}} \exp\{-i\omega h\}\ R'(h)\ dh = -\frac{i}{\pi} \int_0^\infty \sin(\omega h) R'(h)\ dh, \\
s_{10}(\omega) &= \frac{1}{2\pi} \int_{\mathbb{R}} \exp\{-i\omega h\}\ (-R'(h))\ dh = \frac{i}{\pi} \int_0^\infty \sin(\omega h) R'(h)\ dh, \\
s_{11}(\omega) &= \frac{1}{2\pi} \int_{\mathbb{R}} \exp\{-i\omega h\}\ (-R''(h))\ dh = -\frac{1}{\pi} \int_0^\infty \sin(\omega h) R''(h)\ dh.
\end{aligned}$$

So the matrix (2.134) is Hermitian for all ω. Figure 2.15 depicts its eigenvalues for $\omega \in [-100, 100]$, where the matrix entries were computed with symbolic integration using Mathematica. The largest eigenvalue is nonnegative and the smallest eigenvalue is equal to zero for all ω up to the machine accuracy, thus the matrix (2.134) is positive semidefinite. Hence,

the $R_{kl}(h)$ $(k,l = 0,1)$ are valid direct and cross-correlation functions (2.97)–(2.100) for $\{z_0(x), z_1(x)\} = \{z(x), z'(x)\}$ and the correlation matrix R^∂ is positive semidefinite.

A discussion of analytical properties of $R(h)$, which guarantee the positive definiteness of the (cross-) correlation matrix R^∂, lies beyond the scope of this thesis. For further information, the reader is referred to discussions in [114, 146, 142, 57]. In this thesis' test cases and examples, and in all author's applications of GEK using the biquadratic spline correlation function, the correlation matrix $R^\partial \in \mathbb{R}^{(d+1)N}$ was in fact nonsingular for $x(i) \neq x^{(j)}$ $(i \neq j$, $i,j = 1,\ldots,N)$.

2.4.4 Hyperparameter estimation

As in Kriging, after choosing a correlation function $R(h,\theta)$, the hyperparameters $\theta \in \mathbb{R}_+^d$ have to be estimated. The observed data is assumed to be multivariate normally distributed $Y^\partial \sim \mathcal{N}_{(d+1)N}(F^\partial\beta, \sigma^2 R^\partial(\theta))$ with the correlation matrix depending on θ. Thus, maximum likelihood estimation can be applied for the unknowns $\beta \in \mathbb{R}^K$, $\sigma^2 \in \mathbb{R}_+$ and $\theta \in \mathbb{R}_+^d$ by maximizing

$$L(\beta,\sigma^2,\theta|Y^\partial) = (2\pi\sigma^2)^{-\frac{(d+1)N}{2}} (\det R^\partial(\theta))^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(Y^\partial - F^\partial\beta)^\top R^\partial(\theta)^{-1}(Y^\partial - F^\partial\beta)\right)$$

analogously to the Kriging case. In GEK, the entries of the correlation matrix $R^\partial(\theta)$ consist of partial derivatives up to second order of $R(h,\theta)$ with respect to h. Thus, $R^\partial(\theta)$ includes first and second order derivatives of the univariate $R_k(h_k,\theta_k)$ with respect to h_k, cf. (2.96). Therefore, if a gradient-based optimization routine is used for determining θ, not only $R_k(h_k,\theta_k)$, but also $R_k'(h_k,\theta_k)$ and $R_k''(h_k,\theta_k)$ have to be differentiable with respect to $\theta_k > 0$. For example, this is not the case for the cubic spline correlation function (2.63). The biquadratic spline correlation function (2.72), on the other hand, satisfies $R_k(h_k,\theta_k) \in C^2(\mathbb{R}\times\mathbb{R}_+)$ and $R_k(h_k,\theta_k) \in C^3((\mathbb{R}\setminus\{0\})\times\mathbb{R}_+)$, which is easily proven to be sufficient for the continuous differentiability of all entries of R^∂ with respect to $\theta \in \mathbb{R}_+^d$.

Figure 2.16 highlights the problem of insufficient smoothness for a one-dimensional example (data as in Fig. 2.8). The objective function $-\log L$, which is to be minimized, is depicted for the GEK case using the cubic spline correlation function, both for the reduced problem and for the original problem where β and σ^2 are kept fixed in the optimum (cf. Sect. 2.3.4). The objective is not differentiable in e.g. $\theta \approx 2.4$ and also has a local minimum in this point. For a larger number N of samples, even more local minima occur in which the objective function is not differentiable and an optimization algorithm will likely end up in one of them. On the contrary, the biquadratic correlation function is smooth enough and always produces a smooth objective function (see Fig. 2.17)

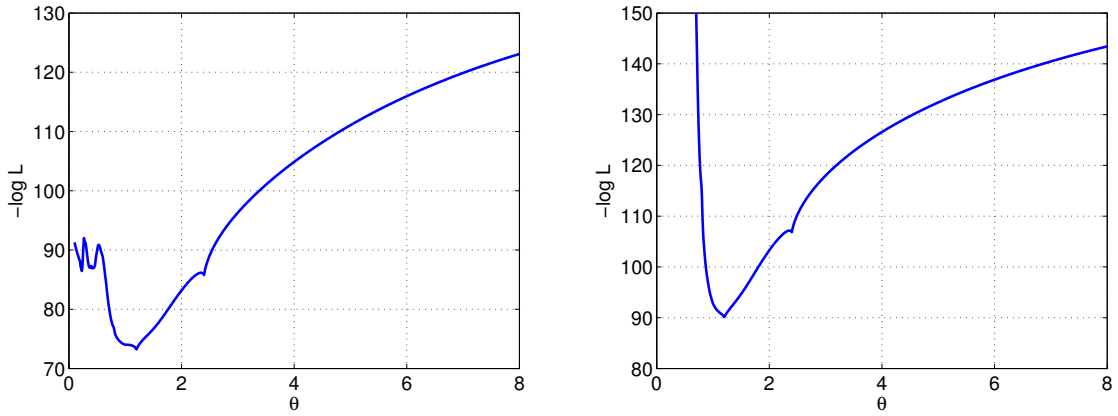

Figure 2.16: $-\log L$ for the data as in Fig. 2.8. Cubic spline correlation function. Left: reduced problem. Right: original problem (fixed β, σ^2).

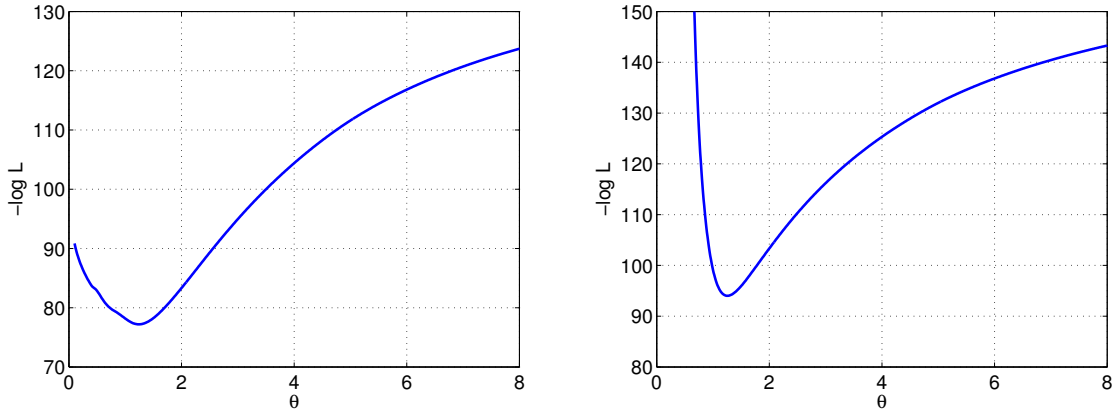

Figure 2.17: $-\log L$ for the data as in Fig. 2.8. Biquadratic spline correlation function. Left: reduced problem. Right: original problem (fixed β, σ^2).

3 Design of experiments

As design of experiments (DOE) one denotes the method of selecting the samples $x^{(i)}$ in which the response $y(x)$ is evaluated ($i = 1, \ldots, N$). The set $X = \{x^{(1)}, \ldots, x^{(N)}\} \subset \Omega$ is referred to as design or sampling. The evaluations of $y(x)$ are assumed to be expensive. Thus, dense computations, which describe the global behavior of the response accurately, are out of reach due to limited computational resources. So for the performance of the surrogate model $\widehat{y}(x)$, which is based on the sampling X of size N and the corresponding evaluations $Y = (y_1, \ldots, y_N)^\top$, there is a natural tradeoff between accuracy and efficiency. As there are different goals in surrogate modeling, like global approximation, target region approximation, optimization of an expensive target function or uncertainty quantification, the design of experiments has to be chosen accordingly. This thesis aims at globally valid surrogate models for expensive computer experiments and sampling strategies have to be applied which use as few samples as possible without sacrificing accuracy.

Two classes of designs can be distinguished: one-stage approaches and adaptive designs. In one-stage approaches, also called a-priori designs or input-based designs, all samples of X are determined before the responses Y are evaluated and the surrogate model $\widehat{y}(x)$ is generated. Since no information of the response is incorporated in the choice of X, these approaches mainly aim at space-filling designs. Adaptive sampling strategies, also referred to as a-posteriori designs, sequential designs or output-based designs, generate X successively. Starting with an initial design, in each stage of the adaptive process a surrogate model is generated based on the current design and assessed. Depending on some infill criterion, one or more new samples are chosen, in which the response is evaluated and the samples and evaluations are added to the current design. In this way, the sampling strategy learns about the behavior of the response. Knowledge about $y(x)$ can be integrated in the choice of the samples X, such that a problem-specific design is generated.

In this chapter, only designs with pairwise disjunct samples are considered, i.e. $x^{(i)} \neq x^{(j)}$ for all $i \neq j$. Designs with multiple samples $x^{(i)} = x^{(j)}$ can e.g. be used in regression models for stochastic outputs. However, in this thesis only outputs of deterministic computer experiments are investigated. Multiple samples do not add supplementary information to the model, since $y(x^{(i)}) = y(x^{(j)})$ for $x^{(i)} = x^{(j)}$. Note also that for Kriging and gradient-enhanced Kriging, multiple samples produce equal rows and columns in the correlation matrix R, which then becomes singular. Furthermore, the investigation is limited to sampling strategies for hypercube input parameter domains $\Omega = [a_1, b_1] \times \ldots \times [a_d, b_d] \subset \mathbb{R}^d$, which is usually transformed to $\Omega = [0, 1]^d$ without loss of generality.

The chapter is organized as follows. Section 3.1 is devoted to one-stage approaches and gives a broad overview on space-filling designs, including factorial designs, distance-based designs, Latin hypercube designs, low-discrepancy designs and also model-based designs, which depend on the structure of the surrogate model for which they are generated. In Sect. 3.2, adaptive sampling strategies are discussed, particularly regarding the concepts of

exploration and exploitation. Three selected adaptive designs are introduced. They have already been investigated in [117] and are also discussed in this thesis for the completeness of the presentation. Additionally, a more profound review is given, including the illustration of the selection criteria with academic test cases.

3.1 One-stage approaches

In this section, an overview on popular designs for deterministic computer experiments, which use a one-stage approach, is presented. They cover design strategies which determine $X = \{x^{(1)}, \dots, x^{(N)}\} \subset \Omega$ before any observations $y(x^{(i)})$ $(i = 1, \dots, N)$ are computed. They also go under the names of a-priori designs or, since $y(x)$ has no influence on the selection of X at all, input-based designs. The goal is to produce surrogate models which are globally accurate. Without information on $y(x)$, a design X therefore is expected to be space-filling, i.e. the samples $x^{(i)}$ are evenly spread in the design space Ω. Several quantities, which describe a measure of the space-filling quality, are presented in the remainder of this section.

Generally, among one-stage approaches, it can be differentiated between deterministic and random designs as well as between model-based and model-independent approaches. For model-independent designs, only knowledge about Ω and the number of samples N is required. Model-based approaches additionally require information on the structure of the surrogate model $\widehat{y}(x)$, which would be the regression matrix $F \in \mathbb{R}^{N \times K}$ and the correlation matrix $R \in \mathbb{R}^{N \times N}$ in the Kriging case (cf. Sect. 2.3.1). The correlation matrix depends not only on the design X, but also on the correlation function, namely the hyperparameters $\theta \in \mathbb{R}^d$ (2.57). Usually, they are determined by maximum likelihood estimation, a procedure which requires both X and the evaluations $Y = (y_1, \dots, y_N)^\top$ (cf. Sect. 2.3.4 and 2.4.4). Unless the hyperparameters are known and kept fixed, model-based approaches are not applicable in Kriging and gradient-enhanced Kriging.

Contrary to adaptive sampling strategies, which are able to learn about the behavior of the yet unknown response $y(x)$, one-stage approaches are independent of $y(x)$ and can therefore only produce generic designs. Evidently, for most responses $y(x)$, problem-adapted designs of some sample size N generally produce more accurate surrogate models than one-stage approaches of the same sample size. This raises the question, why a-priori designs are investigated at all and why not for each surrogate modeling problem a problem-specific adaptive sampling strategy is used. First of all, they can be used for screening. Real life applications can comprehend a high-dimensional input parameter space $\Omega \in \mathbb{R}^d$ (e.g. $d \geq 10$), for which surrogate modeling becomes extremely challenging and computationally expensive. Initial screening studies can be applied to identify redundant variables x_k, which have no or comparatively little influence on $y(x)$ and can be removed from the model in order to reduce the dimension of the problem. However, in the test cases of this work, all variables are considered to be equally important and screening is not applied. Second, the possibility that after determining X, all evaluations $y(x^{(1)}), \dots, y(x^{(N)})$ can be computed in parallel, seems to create a huge benefit regarding wall clock time and efficiency. Adaptive sampling strategies add only one or few new samples in every stage of the adaptive process, such that only one or few evaluations can be computed simultaneously. But nowadays, codes in practical applications excessively use parallel computing. Unless the performance of the

computer experiment itself scales enormously bad with parallel computer architectures, there is no need to make do with one-stage approaches. For a parallel computer architecture and a given number of N samples, computing $y(x^{(1)}), \ldots, y(x^{(N)})$ successively (with adaptively chosen $x^{(i)}$), while for each using full parallelization, means a more accurate surrogate model compared to one-stage approaches, while the sacrifice in efficiency is marginal. Third, any adaptive sampling strategy requires a (small) initial design, which is generated by a one-stage approach. Very small initial designs may be suitable for a low number of input paramaters d, but for higher-dimensional problems (e.g. $d \geq 5$) the initial design should be chosen carefully in order to produce reasonable surrogate models already in the first stages of the adaptive process.

In this thesis, one-stage approaches are investigated for the following reasons. As stated above, small a-priori designs are used as initial designs for adaptive sampling strategies. Also, model-dependent one-stage approaches can be utilized to develop adaptive designs by selecting a new sample based on a model-dependent infill criterion in every stage of the adaptive design process. At last, one-stage approaches serve as a benchmark for the performance of sequential sampling strategies: an adaptive sampling strategy should perform at least as good as any one-stage design of the same sample size N.

A brief literature survey on one-stage designs is now presented. Designs for regression models have been investigated before surrogate modeling for computer experiments was introduced, see e.g. [18, 36, 7, 8]. Already in the first publications on the design and analysis of computer experiments [122, 121, 26], different model-based one-stage approaches like the entropy criterion, the integrated mean squared error (IMSE) criterion and the maximum mean squared error (MMSE) criterion were discussed. In [65], maximin and minimax designs were introduced. For Latin hypercube designs, which were introduced in [95], the authors of [108, 136, 98] used several approaches like orthogonal arrays or the maximin criterion to improve the design performance. The first publication on gradient-enhanced Kriging [99] also contained a discussion of Latin hypercube designs, maximin designs and ϕ_p designs. A detailed review of one-stage approaches for computer experiments was given in [73]. The authors of [72] presented a computational investigation of low-discrepancy designs like the Halton sequence or the Sobol sequence. Surveys on one-stage designs for different surrogate modeling techniques can be found in [131, 132, 111, 64] and in the textbooks [123, 33]. In [81], the performance of different designs was reviewed in an aerodynamical test case. The author of [115] presented greedy algorithms for one-stage designs by subset selection from large datasets. Recent reviews were given in [66, 60, 3]. Also recently, a survey on Latin hypercube designs and a new algorithm for optimizing Latin hypercubes was presented in [144], see also [23] for a comparative study and a Voronoi-tessellation-based strategy.

3.1.1 Factorial designs

Full factorial designs are probably the easiest approach of evenly distributing samples in a d-dimensional hypercube Ω. The design consists of $N = \prod_{k=1}^{d} N_k$ points on a grid, where the coordinates of the samples are $x_k = \frac{i_k - 1}{N_k - 1}$ ($i_k = 1, \ldots, N_k$, $k = 1, \ldots, d$), see Fig. 3.1. If there is no information on the influence of the input parameters x_k on the black-box function $y(x)$, there is little need to choose different values for each N_k ($k = 1, \ldots, d$) and the $N = n^d$ points are distributed on a regular grid. The performance of full factorial designs is limited

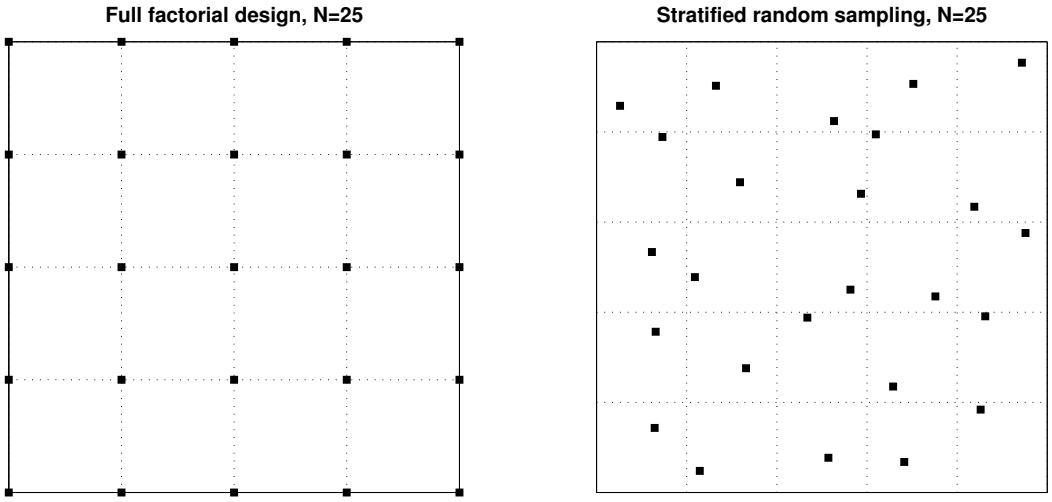

Figure 3.1: Full factorial design and stratified random sampling for $\Omega = [0, 1]^2$.

due to the following reasons. Though the design consists of n^d samples in total, each variable x_k takes only n disjunct values. This also means that if a variable x_k is removed from the model, the design will collapse, i.e. only n^{d-1} disjunct samples will remain in the design after the projection from $\mathbb{R}^d$ onto $\mathbb{R}^{d-1}$. Another problem is the curse of dimensionality. The total number of samples $N = n^d$ grows exponentially with the dimension d and for larger numbers of input parameters, only very coarse grids are computationally affordable. Furthermore, due to their method of construction, full factorial designs are not available for arbitrary sample size N and they only exist for sample sizes of $N = \prod_{k=1}^{d} N_k$, i.e. if N is not a prime number.

In *fractional factorial designs*, particular samples from a full factorial design are deleted in order to overcome the curse of dimensionality and to generate designs which perform acceptably even in high-dimensional problems. They are mostly used for the screening of variables and for the reduction of the number of samples in 2^d-factorial designs. For instance, if a computer experiment has 10 variables, already a full 2^{10}-factorial design consists of 1024 samples. In order to merely apply screening for eliminating redundant variables, 1024 expensive evaluations of the response $y(x)$ are required, which is computationally out of reach in most test cases. More information on fractional factorial designs can be found in [8, 101, 132].

Stratified random samplings can overcome some of the problems associated with full factorial designs. Instead of choosing the n^d grid points as a design, the domain Ω is subdivided into $N = n^d$ equal bins or strata $[\frac{i_1-1}{n}, \frac{i_1}{n}] \times \ldots \times [\frac{i_d-1}{n}, \frac{i_d}{n}]$ $(i_k = 1, \ldots, n,\ k = 1, \ldots, d)$. From each bin a sample is randomly chosen. See Fig. 3.1 for an example. In this way, each region of the input parameter space is represented in the design and each variable x_k generally takes more than the n values of the full factorial design. But, despite the stratification, still comparatively large holes in the design are possible, while clustering of samples

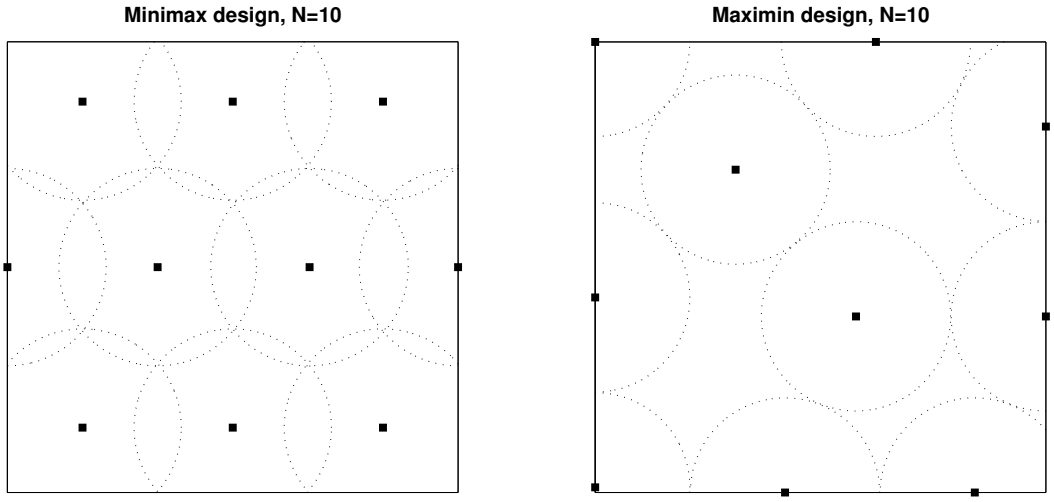

Figure 3.2: Minimax and maximin designs for $\Omega = [0,1]^2$, Euclidean distance $\|\cdot\|_2$.

can occur in other regions due to the random sampling. Particularly, two samples can be picked arbitrarily close. This is a problem in Kriging and GEK models, since closely adjacent points produce two similar rows and columns in the correlation matrix R which then becomes ill-conditioned. For further reading, see e.g. [95, 138, 119].

3.1.2 Distance-based designs

In [65], minimax and maximin distance designs were introduced as space-filling designs for computer experiments. Both use distance-based quantities which assess the space-filling quality of the design. Let

$$\begin{aligned} \text{dist} : \mathbb{R}^d \times \mathbb{R}^d &\to \mathbb{R}, \\ (x, \tilde{x}) &\mapsto \text{dist}(x, \tilde{x}) \end{aligned} \tag{3.1}$$

be a metric on $\mathbb{R}^d$ and $\text{dist}(x, X) := \min_{\tilde{x} \in X} \text{dist}(x, \tilde{x})$ for any set $X \subset \mathbb{R}^d$. A design X^* is a *minimax design* for $\Omega \subset \mathbb{R}^d$ of size N, if the maximum distance of all points $x \in \Omega$ to the design is minimal, i.e.

$$\max_{x \in \Omega} \text{dist}(x, X^*) = \min_{\substack{X \subset \Omega \\ \#X = N}} \max_{x \in \Omega} \text{dist}(x, X). \tag{3.2}$$

Thus the minimax design covers Ω with N closed balls $\overline{B}_\epsilon(x^{(i)}) := \{x : \text{dist}(x, x^{(i)}) \leq \epsilon\}$ with the smallest $\epsilon > 0$. This means that each $x \in \Omega$ lies within a distance of ϵ or less to the closest $x^{(i)} \in X^*$.

A design X^* is called *maximin design* for $\Omega \subset \mathbb{R}^d$ of size N, if the minimum distance of

each pair $x^{(i)}, x^{(j)} \in X^*$ $(i \neq j)$ is maximal,

$$\min_{\substack{x^{(i)},x^{(j)} \in X^* \\ i \neq j}} \operatorname{dist}(x^{(i)}, x^{(j)}) = \max_{\substack{X \subset \Omega \\ \#X = N}} \min_{\substack{x^{(i)},x^{(j)} \in X \\ i \neq j}} \operatorname{dist}(x^{(i)}, x^{(j)}). \tag{3.3}$$

So for the maximin design, the samples are chosen such that the N open balls $B_\epsilon(x^{(i)}) := \{x : \operatorname{dist}(x, x^{(i)}) < \epsilon\}$ are non overlapping $(B_\epsilon(x^{(i)}) \cap B_\epsilon(x^{(j)}) = \emptyset,\ i \neq j)$ and ϵ is maximal. All samples have a distance of at least ϵ.

ϕ_p-designs are a generalization of maximin distance designs [99]. While in maximin designs, only the minimum distance pair $(x^{(i)}, x^{(j)})$ is considered and all other pairs are neglected, ϕ_p-designs use an averaged distance function $(p > 0)$

$$\phi_p(X) := \left(\frac{1}{\binom{N}{2}} \sum_{i=1}^{N-1} \sum_{j>i} \operatorname{dist}(x^{(i)}, x^{(j)})^{-p} \right)^{\frac{1}{p}}, \tag{3.4}$$

which considers all combinations $x^{(i)}, x^{(j)} \in X$ $(j > i)$. A design $X \subset \Omega$ is then chosen such that $\phi_p(X)$ is minimal. For any $0 < p < \infty$, some mean value of the distances between all samples of $X \subset \Omega$ is maximized, for instance the harmonic mean $(p = 1)$. For $p \to \infty$, the design is a maximin design, see Sect. 5.3 in [123].

Finding an optimal distance-based design for a given N is a highly challenging task, even in low-dimensional spaces. Minimax designs are related to covering problems and maximin designs to packing problems in combinatorics. Often, for a given N there exist only best-known results with the best yet known minimax score $(\max_{x \in \Omega} \operatorname{dist}(x, X))$ or maximin score $(\min_{x^{(i)},x^{(j)} \in X} \operatorname{dist}(x^{(i)}, x^{(j)}))$. See for instance [97, 104, 105] for more information on coverings and packings. Nevertheless, the three quantities minimax score (3.2), maximin score (3.3) and ϕ_p (3.4) can be used to measure and compare the space-filling quality of any design X, cf. Sect. 3.1.3.

3.1.3 Latin hypercube designs

The main idea behind Latin hypercube designs is to generate a stratified design X of size N, in which each variable x_k takes exactly N different, generally evenly distributed values $(k = 1, \ldots, d)$. They were first introduced in [95]. Let the interval $[0, 1]$ be partitioned into the N subintervals $I_j := [\frac{j-1}{N}, \frac{j}{N}]$ $(j = 1, \ldots, N)$. Furthermore, let

$$\begin{aligned} \pi_k : \{1, \ldots, N\} &\to \{1, \ldots, N\}, \\ i &\mapsto \pi_k(i) \end{aligned} \tag{3.5}$$

be d random permutations of the numbers $\{1, \ldots, N\}$ $(k = 1, \ldots, d)$. A Latin hypercube design then consists of N samples

$$x^{(i)} \in I_{\pi_1(i)} \times \ldots \times I_{\pi_d(i)}. \tag{3.6}$$

The samples can be e.g. be chosen as the centers of the hypercubes $I_{\pi_1(i)} \times \ldots \times I_{\pi_d(i)}$ or by random sampling of the coordinates $x_k^{(i)}$ from the uniform distributions $\mathcal{U}(I_{\pi_k(i)})$ $(i = 1, \ldots, N,$

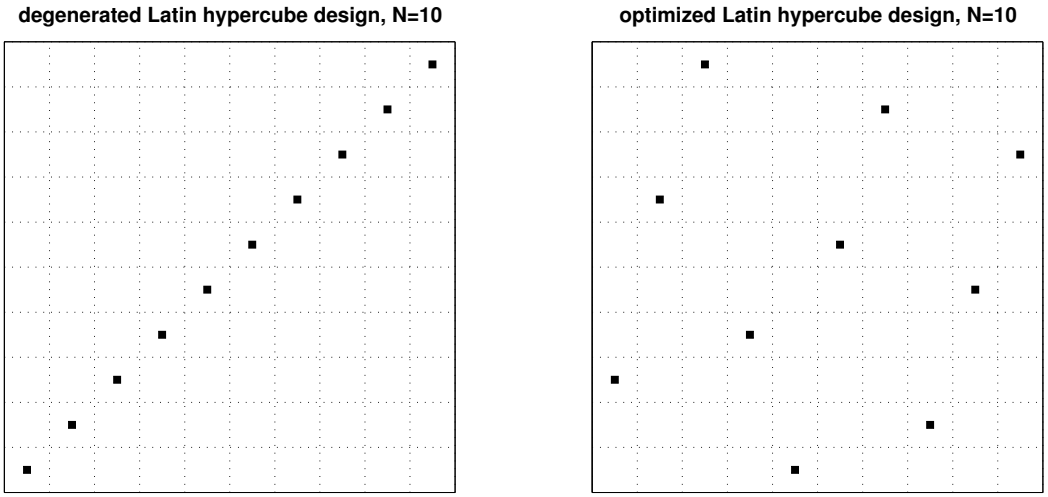

Figure 3.3: Latin hypercube designs for $\Omega = [0,1]^2$. Left: degenerated ($\pi_k(i) = i$), right: optimized maximin score.

$k = 1, \ldots, d$). Note again that, if the samples inside the hypercubes are chosen randomly, two samples $x^{(i)}, x^{(j)}$ from adjacent hypercubes could become arbitrary close and cause an ill-conditioned correlation matrix R. A Latin hypercube design ensures that every subinterval $I_j \subset [0,1]$ ($j = 1, \ldots, N$) on the x_k-axis ($k = 1, \ldots, d$) is represented in the design using N samples in total only. Full factorial designs and stratified random samplings on the contrary need N^d samples for a complete representation of the intervals (cf. Sect. 3.1.1). By their method of construction, Latin hypercube designs are also non collapsing. This especially means that if any variable x_k is removed from the model, an existing design $X \subset \mathbb{R}^d$ becomes a new Latin hypercube design $\widetilde{X} \subset \mathbb{R}^{d-1}$ of the same size N after projection onto $\mathbb{R}^{d-1}$. For instance, after removing the variable x_d (without loss of generality), the samples of the new design satisfy $\tilde{x}^{(i)} \in I_{\pi_1(i)} \times \ldots \times I_{\pi_{d-1}(i)}$ ($i = 1, \ldots, N$).

Due to the random permutations π_k, it is still possible to produce bad designs which are not space-filling at all and leave large regions of the input parameter space $[0,1]^d$ uncovered. For instance, consider a worst-case scenario where all permutations are the identity function ($\pi_k(i) = i$). Then, a degenerated design is produced which presumably will result in an inaccurate surrogate model $\widehat{y}(x)$, see Fig. 3.3. In the last 20 years, efforts have been made to improve the designs and to find optimized solutions within the class of Latin hypercube designs. In [108, 136], orthogonal arrays were used to generate Latin hypercube designs, in which all regions of the input parameter space $[0,1]^d$ are evenly represented. Also, the maximin score (3.3) and ϕ_p (3.4), respectively, were used in [98] in order to improve the space-filling quality of the designs. A recent survey on improvement techniques and new algorithms can be found in [144]. Figure 3.3 depicts a Latin hypercube design with an optimized maximin score, generated by the Matlab function `lhsdesign`.

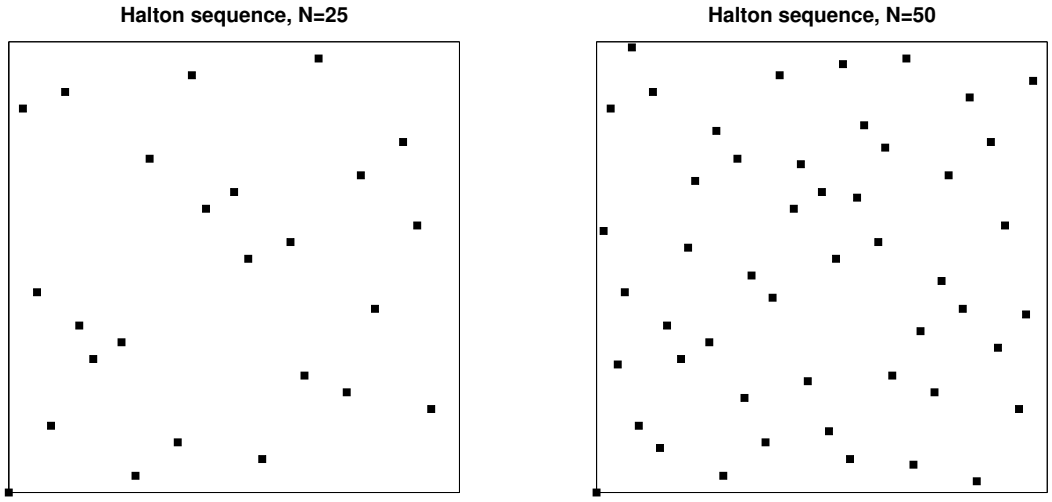

Figure 3.4: Low-discrepancy designs based on the Halton sequence for $\Omega = [0,1]^2$.

3.1.4 Low-discrepancy designs

In [102], a new indicator for the space-filling quality of a design X was introduced. Let $\mathcal{I} := \{I \subset \mathbb{R}^d : I = [a_1, b_1] \times \ldots \times [a_d, b_d], a_k < b_k\}$ be the set of d-dimensional intervals and $\mathrm{Vol}(I)$ denote the d-dimensional Lebesgue measure. Furthermore, let $\#X$ identify the number of elements in a set X. Then, the discrepancy

$$D_N(X) := \sup_{\substack{I \in \mathcal{I} \\ I \subset [0,1]^d}} \left| \frac{\#(X \cap I)}{N} - \mathrm{Vol}(I) \right| \tag{3.7}$$

of a design X for $\Omega = [0,1]^d$ of size N quantifies how much the distribution of X deviates from the d-dimensional uniform distribution $\mathcal{U}([0,1]^d)$ on the unit hypercube. The expression under the norm in (3.7) compares the ratio of $\mathrm{Vol}(I)$ and $\mathrm{Vol}([0,1]^d) = 1$ to the ratio of the number of samples in I and the total number of samples N. This implies that the lower the discrepancy $D_N(X) \in (0,1]$, the more evenly X is distributed in $[0,1]^d$.

Monte Carlo samplings of size N have a discrepancy of $D_N(X) = \mathcal{O}(\frac{1}{\sqrt{N}})$, see e.g. [102], Sect. 1.2. In order to improve this convergence rate, low-discrepancy sequences were developed. Examples for such designs are the Hammersley set, the van der Corput sequence, the Sobol sequence and the Halton sequence. While the samples of the Hammersley set depend on N, elements of the Sobol sequence and the Halton sequence are independent of the total number of samples N. Therefore they can also be used for sequential sampling, i.e. for the refinement of an existing low-discrepancy design X, see Fig. 3.4 for Halton designs generated by the Matlab function `haltonset`. Again, see [102] for the definition and a theoretical discussion of the mentioned designs. A computational investigation can be found in [72]. Because of their good space-filling property, low-discrepancy designs have become more

popular in surrogate modeling during the last decade, cf. [131, 123, 81]. Moreover, efforts have been made to produce Latin hypercube designs which minimize a discrepancy score, see e.g. [64, 33, 60].

3.1.5 Model-based designs

As the name suggests, the difference between model-based designs and the one-stage approaches in the previous sections is that they are not independent of the surrogate model which they are used for. For Kriging and gradient-enhanced Kriging, this particularly means that the correlation function $R(h, \theta) = \prod_{k=1}^{d} R_k(|h_k|, \theta_k)$ (2.57) has to be known a-priori, before a design X can be obtained. Typically, this is not the case because the data $Y = (y(x^{(1)}), \ldots, y(x^{(N)}))^\top$ is not available in the design phase. After a sampling X is determined, the responses Y are evaluated and the hyperparameters $\theta = (\theta_1, \ldots, \theta_d)^\top$ are computed via maximum likelihood estimation using the data X and Y, cf. Sect. 2.3.4. Unless the user does have a guess about the hyperparameters θ, e.g. by best-practice experience from similar computer experiments, model-based criteria for the design of experiments are not applicable in Kriging and gradient-enhanced Kriging.

Maximum entropy designs are based on a Bayesian approach. For the time being, let Ω be a finite set of size M. Let the design $X \subset \Omega$ be a subset of size $N < M$ and $X^C := \Omega \setminus X$ be the complement of X in Ω. $Y_X \in \mathbb{R}^N$ denotes the random vector containing all evaluations in $x \in X$, analogously for $Y_{X^C} \in \mathbb{R}^{M-N}$. Following the authors of [26] and [130], the knowledge of $y(x)$ in X^C is embedded in the distribution of $Y_{X^C}|Y_X$, which is defined as a random vector conditional on Y_X. Then the design X should be chosen such that the uncertainty in $Y_{X^C}|Y_X$ is minimized. The amount of uncertainty in a random vector Z can be quantified by the entropy

$$H(Z) := \mathrm{E}_Z[-\log p_Z(Z)], \tag{3.8}$$

where p_Z denotes the probability density function. By a decomposition of $H(Y_\Omega) = H(Y_X) + \mathrm{E}_{Y_X}[H(Y_{X^C}|Y_X)]$, the authors of [130] deduced that minimizing the expected posterior entropy $\mathrm{E}_{Y_X}[H(Y_{X^C}|Y_X)]$ is equivalent to maximizing the prior entropy $H(Y_X)$. This, on the other hand, is equivalent to maximizing $\det(R) \cdot \det(F^\top R^{-1} F)$ and, furthermore, to maximizing only the determinant of the correlation matrix (2.34)

$$\det(R), \tag{3.9}$$

if $\beta \in \mathbb{R}^K$ and $\sigma^2 \in \mathbb{R}$ are assumed constant in the Kriging model (2.31). Since the correlation matrix $R \in \mathbb{R}^{N \times N}$ depends on the design X as well as the hyperparameters θ, the maximum entropy criterion is a model-based criterion.

Two other criteria make use of the predicted mean squared error (2.49)

$$\mathrm{MSE}\left[\widehat{y}(x)\right] = \sigma^2 \left\{ 1 - \begin{pmatrix} r(x) \\ f(x) \end{pmatrix}^\top \begin{bmatrix} R & F \\ F^\top & 0 \end{bmatrix}^{-1} \begin{pmatrix} r(x) \\ f(x) \end{pmatrix} \right\} \tag{3.10}$$

of the surrogate model, cf. [122]. A design X of size N is *integrated mean squared error*

optimal (IMSE-optimal), if it minimizes

$$\frac{1}{\sigma^2} \int_\Omega \mathrm{MSE}[\widehat{y}(x)]\, dx \tag{3.11}$$

among all designs of size N, i.e. the average predicted error is minimal. The expression is equal to

$$\int_\Omega dx - \mathrm{trace}\left(\begin{bmatrix} R & F \\ F^\top & 0 \end{bmatrix}^{-1} \int_\Omega \begin{bmatrix} r(x)r(x)^\top & r(x)f(x)^\top \\ f(x)r(x)^\top & f(x)f(x)^\top \end{bmatrix} dx \right),$$

where the integral is taken component-wise. For rectangular domains Ω, a product correlation model $R(\tilde{x} - x)$ as in (2.57) and most simple regression models $f(x)$, the integrals can be computed as products of one-dimensional integrals, see e.g. Sect. 6.2.2 in [123]. A design X of size N is *maximum mean squared error optimal* (MMSE-optimal), if it minimizes

$$\max_{x \in \Omega} \left\{ \frac{1}{\sigma^2} \mathrm{MSE}[\widehat{y}(x)] \right\} \tag{3.12}$$

among all designs of size N, i.e. the maximum predicted error is minimal. Division by σ^2 makes both expressions (3.11) and (3.12) independent of the process variance, still depending on the design X and the hyperparameters θ, such that both criteria are also model-based criteria.

Entropy designs, IMSE designs and MMSE designs are related to D-, A- and G-optimal designs (in that order) for regression models (for the definition of the optimality criteria see Sect. 8.2 in [101]). The authors of [123] showed in Sect. 6.2.2, that IMSE designs are A-optimal and MMSE designs are G-optimal, if the stochastic process $z(x)$ in the model (2.27) is not a Gaussian process, but a white noise process. Also, for $R = I$, it is easily deduced that maximizing $\det(R) \cdot \det(F^\top R^{-1} F)$ is the same as maximizing $\det(F^\top F)$, which is the classical maximum entropy approach for regression models and is called D-optimality. Furthermore, in [65] asymptotics for entropy designs and MMSE designs were given. The authors demonstrated that maximin designs (3.3) are asymptotically entropy-optimal and minimax designs (3.2) are asymptotically MMSE-optimal, by taking the limit of a sequence of correlation functions, for which $R(\tilde{x} - x) = \mathrm{Corr}[y(x), y(\tilde{x})]$ converges to zero ($x \neq \tilde{x}$).

Finding optimal model-based designs is no trivial task and it involves solving a constrained global optimization problem with $N \cdot d$ variables. Heuristics and approaches to the solution can be found in [121, 122, 26, 73, 66]. But there is no workaround for the problem that in model-based designs the hyperparameters θ have to be known up-front, which makes them impractical for Kriging and gradient-enhanced-Kriging in most applications. However, if adaptive design strategies are used instead of a one-stage approach, in each stage (N_i) of the adaptive process, a current design $X^{(N_i)}$ of size N_i and the corresponding data $Y^{(N_i)} \in \mathbb{R}^{N_i}$ are available and also a preliminary surrogate model $\widehat{y}^{(N_i)}(x)$. One of the model-based criteria (3.9)–(3.12) can then be used to find the $(N_i + 1)$-th sample $x^{(N_i+1)}$, for which the design $X^{(N_i+1)} = X^{(N_i)} \cup \{x^{(N_i+1)}\}$ is optimal, cf. Sect. 3.2.1.

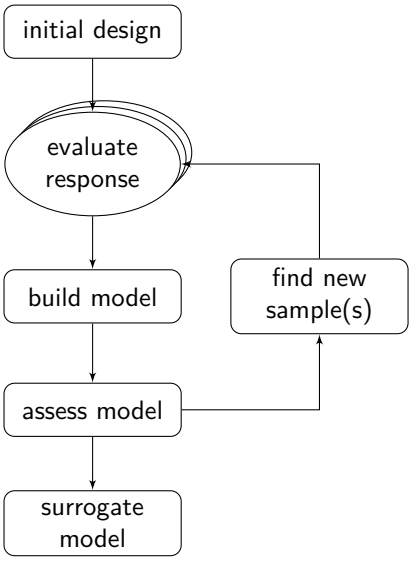

Figure 3.5: Adaptive sampling strategy framework.

3.2 Adaptive designs

Adaptive sampling strategies are a viable alternative to one-stage approaches, which received a growing attention during the last decade. Contrary to the a-priori designs in the last section, in a-posteriori sampling, also called sequential, adaptive or output-based designs, the design X as well as the surrogate model $\widehat{y}(x)$ is built step-by-step. Starting off with a small initial design $X^{(N_0)}$ of size N_0 (e.g. determined by a one-stage approach), the response $y(x)$ is evaluated in all samples $x \in X^{(N_0)}$ and an initial surrogate model $\widehat{y}^{(N_0)}(x)$ is built based on the data $X^{(N_0)}$ and $Y^{(N_0)} \in \mathbb{R}^{N_0}$. For Kriging and gradient-enhanced Kriging, this involves the hyperparameter estimation and the solution of the linear Kriging system (2.51)/(2.123). Then, in each stage N_i of the adaptive process, one or more new sample(s) $x \in \Omega$ are chosen by some selection criterion. The criterion makes use of the information at hand, which is included in the current surrogate model $\widehat{y}^{(N_i)}(x)$. The sample(s) are added to the current design, the response is evaluated in the new sample(s) and a new surrogate model $\widehat{y}^{(N_{i+1})}(x)$ is built based on the augmented data $X^{(N_{i+1})}$ and $Y^{(N_{i+1})} \in \mathbb{R}^{N_{i+1}}$ including a new hyperparamter estimation, cf. Fig. 3.5. The procedure can be iterated until a predefined maximum number of samples $N_{\max}$ is reached or until some error estimator falls below a certain threshold. This also makes adaptive sampling strategies more flexible than one-stage approaches, since the total number of samples does not have to be known in advance.

Thus, information about the true response $y(x)$ is directly included in the process of finding an optimal design X. The design is therefore both adapted to the specific test case (by including the data Y during the adaptive process) as well as to the research goal (by an appropriate criterion of selecting new samples). By creating a problem-adapted design, a more intelligent distribution of the samples in the input parameter domain $\Omega \subset \mathbb{R}^d$ is pursued, for instance by a higher concentration of the samples in critical regions of Ω, which are identified automatically. In this way, a reduction of the total number of samples $N_{\max}$

is targeted compared to the space-filling designs of the one-stage approaches, which bear the risk of wasting many samples on uncritical regions. For example, in surrogate modeling for optimization, adding most samples in regions where the value of an objective function is not improved is barely desired. Also, in global approximation, refining the design where the approximation quality is already high does not produce a better surrogate model.

In adaptive sampling strategies for the global optimization of expensive black-box functions, two main concepts can be distinguished: *exploitation* and *exploration*, cf. [41] for instance. An update strategy based on pure exploitation would add a new sample $x^{(N+1)} = \arg\min_{x\in\Omega}\{\widehat{y}^{(N)}(x)\}$ where the current surrogate objective $\widehat{y}^{(N)}(x)$ is optimal. If this strategy is to converge to the global optimum of the true response $x^* = \arg\min_{x\in\Omega}\{y(x)\}$, the surrogates $\widehat{y}^{(N)}(x)$ $(N = 1, 2, \dots)$ have to be suitable global approximations. Otherwise the iterates could converge to a local optimum. Further regions of Ω where $\widehat{y}^{(N)}(x) > \min_{x\in\Omega}\{\widehat{y}^{(N)}(x)\}$ would be disregarded, though little is known there about the behavior of the true response $y(x)$. On the other hand, a pure exploration method would add a new sample $x^{(N+1)}$ where the approximation error is assumed highest or where a space-filling criterion from the previous sections can be improved. The approximation quality of $\widehat{y}^{(N)}(x)$ is improved globally in every stage N, but a large number of samples N could be required in order that the global optimum of the surrogate $\arg\min_{x\in\Omega}\{\widehat{y}^{(N)}(x)\}$ and the true response $\arg\min_{x\in\Omega}\{y(x)\}$ coincide. A balanced strategy between exploitation and exploration is generally used in order to converge faster to the global optimum.

The concept of exploitation and exploration can roughly be adopted in global approximation methods. The target is to reduce the interpolation error, so an exploitation method would place a new sample $x^{(N+1)}$ where a predicted error is highest. Using Kriging, the error in $\widehat{y}^{(N+1)}(x^{(N+1)}) = y(x^{(N+1)})$ is set to zero. But using exploitation only, there might be regions in the input parameter domain Ω where a low error is assumed, but in fact again little is known about the true response $y(x)$. For instance, if not enough samples are distributed there, a peak in the true response might stay undetected. A pure exploration method on the other hand would use a distance-based criterion in order to ensure that all regions of the input parameter space are attended. But this again would merely produce a space-filling design without any priorities for critical regions. These regions typically require a denser distribution of samples, e.g. for resolving large gradients, curvature or even erratic behavior of the response $y(x)$. Therefore, also in surrogate modeling for global approximation of black-box functions, a mixed strategy between exploitation and exploration is assumed to be beneficial. It ensures not only the accuracy of the surrogate model, but also an efficient usage of expensive response evaluations. Such strategies are both local and global. They are local in the sense that samples are distributed more densely in critical regions, global in the sense that other regions are not neglected either and they still produce a space-filling design. Not using a single selection criterion, but multiple or mixed criteria for the refinement of a design X has been the subject of recent studies, see e.g. [46, 23]. In this thesis, cross-validation methods are proposed as a local component of a refinement criterion. They measure local errors by deleting the i-th sample from the surrogate model and comparing the new prediction to the true response $y_i = y(x^{(i)})$ $(i = 1, \dots, N)$, see Sect. 3.2.2. As a global component, the mean squared error prediction $\mathrm{MSE}[\widehat{y}(x)]$ (2.49) is proposed. Its computation does not involve the response evaluations Y directly, but only indirectly via the global hyperparameters θ in the correlation matrix and vector. Thus, the

MSE is rather a distance-based quantification of the uncertainty in the prediction than a local error estimator, see Sect. 3.2.1.

An algorithm for the efficient global optimization of expensive black-box functions (EGO) based on an expected improvement criterion was presented in [67]. It received a lot of attention and reviews, enhancements and further investigations can be found in [125, 124, 123, 58, 40, 41]. Surveys on adaptive sampling strategies for optimization were given in [111, 39, 129]. Recent applications in aerodynamic optimization can be found in [79, 80, 150]. New studies in reliability-based optimization were presented in [151, 30]. The use of adaptive sampling strategies in global approximation methods for expensive computer experiments dates back to the pioneering article [122], where the IMSE criterion was used for an adaptive refinement. The authors of [107] also used a model-based criterion for updating Kriging surrogates adaptively. In [63], a new cross-validation approach was introduced and was compared to model-based and distance-based adaptive design, while in [71] jackknifing was used to estimate the variance of cross-validation errors. New approaches in adaptive maximum entropy designs were presented in [85, 34] by adjusting the entries of the correlation matrix according to observed irregularities in the data Y. Adaptive designs based on decompositions of the input parameter domain Ω were introduced in [11] and in [24]. In the former, a method called adaptive gridding was developed for a regular partitioning, while in the latter a Voronoi tessellation was used, see also [23, 25]. A sampling strategy for surrogate models based on nonstationary Gaussian processes was developed in [49], which exploits the assumption that the correlation structure is not global, but can vary in different parts of the input parameter domain. Designs based on active learning were discussed in [16]. The authors of [90] investigated optimal design extraction strategies based on a dense set of validation, compared the distributions of the designs to local complexity and error estimators and discussed possible new sampling strategies. For applications of adaptive sampling strategies in aerodynamic test cases see for instance [81, 41, 49, 117].

3.2.1 MSE-based strategies

The $\mathrm{MSE}\,[\widehat{y}(x)]$ (2.49) is a measurement of the uncertainty of the prediction, so it is quite intuitive to use it as a selection criterion for adaptive sampling. It is zero in every $x^{(i)} \in X$ and grows with distance to the existing samples. Expanding the expression

$$\begin{aligned}
\mathrm{MSE}\,[\widehat{y}(x)] &= \sigma^2 \left\{ 1 - \begin{pmatrix} r(x) \\ f(x) \end{pmatrix}^\top \begin{bmatrix} R & F \\ F^\top & 0 \end{bmatrix}^{-1} \begin{pmatrix} r(x) \\ f(x) \end{pmatrix} \right\} \\
&= \sigma^2 \left\{ 1 - \begin{pmatrix} r(x) \\ f(x) \end{pmatrix}^\top \begin{pmatrix} \lambda(x) \\ \mu(x) \end{pmatrix} \right\} \\
&= \sigma^2 \left\{ 1 - r(x)^\top \lambda(x) - f(x)^\top \left(F^\top R^{-1} F \right)^{-1} \left(F^\top R^{-1} r(x) - f(x) \right) \right\}, \qquad (3.13)
\end{aligned}$$

it can be observed that it tends to $\sigma^2\{1 + f(x)^\top (F^\top R^{-1} F)^{-1} f(x)\}$ for $\mathrm{dist}(x, X) \to \infty$, since $r(x) = (R(x^{(i)} - x))_{i=1}^N \to 0$. Its computation does not contain Y directly. Thus, for any new sample $x^{(N+1)} \in \Omega \setminus X^{(N)}$, the mean squared error prediction for $\widehat{y}^{(N+1)}$ can be evaluated based on the data $X^{(N)} \cup \{x^{(N+1)}\}$ only and without any knowledge of $y(x^{(N+1)})$,

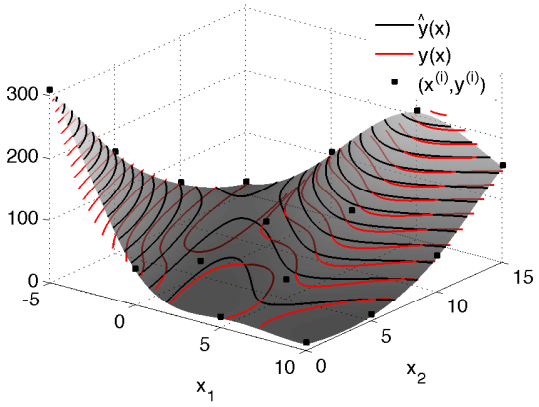

Figure 3.6: Kriging interpolation of the Branin test function.

if σ^2 and the hyperparameters θ of $\widehat{y}^{(N)}$ are used.

Therefore, model-based criteria from Sect. 3.1.5 like the MMSE criterion and the IMSE criterion can be applied in adaptive sampling. In an MMSE-based adaptive sampling strategy, the new sample is obtained by

$$x^{(N+1)} = \arg \min_{x^{(N+1)} \in \Omega} \left\{ \max_{x \in \Omega} \mathrm{MSE}[\widehat{y}^{(N+1)}(x)] \right\}. \tag{3.14}$$

It is the sample for which the new surrogate $\widehat{y}^{(N+1)}$ has the lowest maximum MSE, provided σ^2 and θ are kept fixed. Analogously, an IMSE-based adaptive sampling strategy chooses

$$x^{(N+1)} = \arg \min_{x^{(N+1)} \in \Omega} \left\{ \int_{\Omega} \mathrm{MSE}[\widehat{y}^{(N+1)}(x)] \, dx \right\}, \tag{3.15}$$

the sample for which the new surrogate $\widehat{y}^{(N+1)}$ has the lowest integrated MSE. These approaches have been proposed and investigated e.g. in [122, 63, 49].

A much simpler approach of selecting a new sample is determining

$$x^{(N+1)} = \arg \max_{x \in \Omega} \left\{ \mathrm{MSE}[\widehat{y}^{(N)}(x)] \right\}, \tag{3.16}$$

the maximum of the predicted mean squared error of the current surrogate $\widehat{y}^{(N)}$. Given the fact that $\mathrm{MSE}[\widehat{y}^{(N+1)}(x^{(N+1)})] = 0$, choosing $x^{(N+1)}$ where $\mathrm{MSE}[\widehat{y}^{(N)}(x)]$ is largest will also lead to a probably adequate decrease of the MMSE and the IMSE score of $\widehat{y}^{(N+1)}$. Criterion (3.16) is connected to maximum entropy designs (cf. Sect. 3.1.5). The authors of [63] proved that (3.16) is equivalent to maximizing

$$\det\left\{ \begin{bmatrix} R & r(x) \\ r(x)^\top & 1 \end{bmatrix} \right\} \cdot \det\left\{ \begin{bmatrix} F \\ f(x)^\top \end{bmatrix}^\top \begin{bmatrix} R & r(x) \\ r(x)^\top & 1 \end{bmatrix}^{-1} \begin{bmatrix} F \\ f(x)^\top \end{bmatrix} \right\}.$$

This, on the other hand, is the same as the maximization of the entropy of the dataset

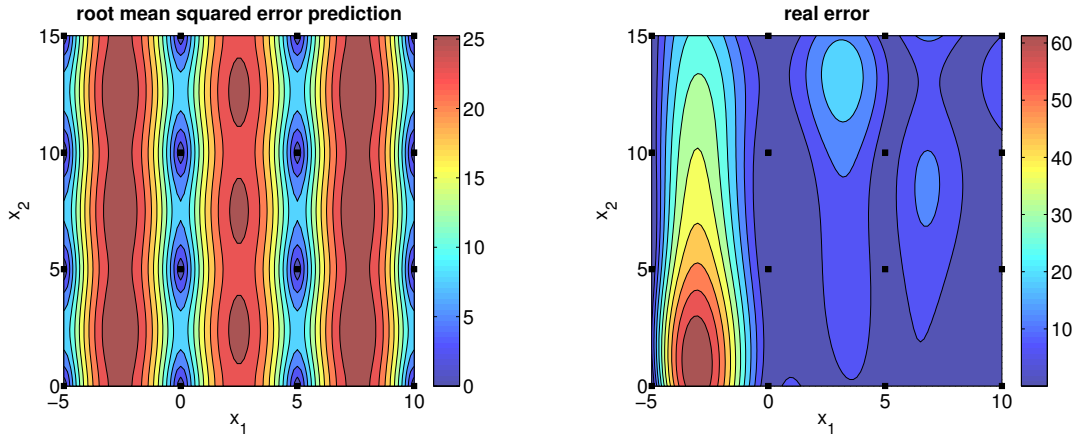

Figure 3.7: RMSE$[\widehat{y}(x)]$ (left) and $|\widehat{y}(x) - y(x)|$ (right) for the interpolation in Fig. 3.6.

$X^{(N+1)} = X^{(N)} \cup \{x^{(N+1)}\}$, if $X^{(N)}$ is kept fixed and $x^{(N+1)}$ is the only degree of freedom. This adaptive sampling strategy was used e.g. in [93], a modified adaptive approach is found in [34] and comparisons to other techniques were presented in [63, 81, 117].

The use of MSE-based sampling strategies for efficient global approximation is limited. Since the evaluation of MSE$[\widehat{y}(x)]$ does not include the response Y directly, it has only a poor performance as a local error indicator. Y rather has a globalized and indirect contribution via the hyperparameters $\theta \in \mathbb{R}^d$ in the correlation function. The following example illustrates the issues with MSE-based approaches. Figure 3.6 shows a Kriging interpolation $\widehat{y}(x)$ of the Branin function (2.88) based on an 4×4 factorial design. Both the root mean squared error $\text{RMSE}[\widehat{y}(x)] = \sqrt{\text{MSE}[\widehat{y}(x)]}$ and the real error $|\widehat{y}(x) - y(x)|$ are depicted in Fig. 3.7. The RMSE fails to uniquely identify the region where the real error is highest, which would be a natural choice for a new sample for global approximation. Besides the zeros in the samples X, both functions do not share any similarities in this test case. The RMSE is rather a distance-based measure, in which the distance to the nearest samples $x^{(i)} \in X$ is weighted by the correlation function $R(x^{(i)} - x, \theta)$. Using the biquadratic spline correlation function (2.72), the maximum likelihood hyperparameter estimation yields $\theta_1 \approx 2\theta_2$ and the RMSE grows approximately twice as fast in $(x_1^{(i)} - x_1)$-direction as in $(x_2^{(i)} - x_2)$-direction. Note that the correlation length along the x_k-axis is $\frac{1}{\theta_k}$, i.e. $R_k(x_k^{(i)} - x_k) = 0$ for $|x_k^{(i)} - x_k| \geq \frac{1}{\theta_k}$ (cf. Figs. 2.4 and 2.6). Thus, selecting a new sample where the predicted RMSE is highest primarily aims at a space-filling design. But compared to one-stage approaches, in every stage of the sequential process the distance measure is based on the current surrogate model, i.e. the current hyperparameters θ. Speaking in terms of this section's introduction, it is a pure exploration method.

3.2.2 Cross-validation-based strategies

In this section, cross-validation is introduced as an exploitation criterion for adaptive sampling criteria in order to design a mixed strategy between exploration and exploitation. Kriging and gradient-enhanced Kriging are interpolators of the data $Y \in \mathbb{R}^N$ and an assess-

ment of the approximation error in $\Omega \setminus X$ based on RMSE performs poorly, see Sect. 3.2.1. Cross-validation is a method of estimating error information of the surrogate model by comparing predictions based on a reduced dataset to the existing data $Y \in \mathbb{R}^N$. Let $\widehat{y}_{-i}(x)$ denote the Kriging predictor based on $X \setminus \{x^{(i)}\}$. Keeping the hyperparameters θ of $\widehat{y}(x)$ fixed, $\widehat{y}_{-i}(x)$ can be computed by deleting the i-th column and row of the system matrix and the i-th entries of the vectors in the Kriging equation (2.50). In gradient-enhanced Kriging, all $(kd+i)$-th columns and rows of the system matrix in (2.122) are deleted $(k = 0, \ldots, d)$ and all $(kd+i)$-th vector entries, respectively. The *leave-one-out* cross-validation error for the sample $x^{(i)}$ is defined by

$$e_i := \left| \widehat{y}_{-i}(x^{(i)}) - y_i \right| \tag{3.17}$$

and an error estimator can be evaluated by $\frac{1}{N}\sum_i^N e_i$ or $\frac{1}{N}\sum_i^N e_i^2$. Cross-validation has been widely used as a global error measure and model assessment, see [67, 96, 93, 143, 47].

But for an adaptive sampling strategy a local error estimator is required, which can be evaluated in arbitrary candidates $x \in \Omega$. The cross-validation error e_i measures the influence of the data pair $(x^{(i)}, y_i)$ on the prediction $\widehat{y}(x)$. If e_i is low, $(x^{(i)}, y_i)$ is redundant and no additional samples are needed in the proximity of $x^{(i)}$. If, on the contrary, e_i is high, $(x^{(i)}, y_i)$ has an essential contribution to the performance of $\widehat{y}(x)$. This can indicate a critical region with e.g. large gradients or high curvature in the response $y(x)$. The concept of cross-validation is extended to Ω in the following way:

$$e(x) := \frac{1}{N} \sum_{i=1}^{N} \left| \widehat{y}_{-i}(x) - \widehat{y}(x) \right| . \tag{3.18}$$

Cleary, $e(x^{(i)}) = \frac{1}{N} e_i$ holds since $\widehat{y}_{-i}(x)$ interpolates all y_j in $x^{(j)} \neq x^{(i)}$ as well as $\widehat{y}(x)$. Now choosing $x^{(N+1)}$ where the predicted error $e(x)$ is highest can be impractical, because $|\widehat{y}_{-i}(x) - \widehat{y}(x)|$ tends to be highest in some $x^{(i)}$ and $x^{(N+1)} = \arg\max_{x\in\Omega} e(x)$ is likely to be close to or even identical with an existing $x^{(i)} \in X$. In [63], $e(x)$ was multiplied with a minimum distance function $\operatorname{dist}(x, X)$, which guarantees the product to be zero in every $x^{(i)} \in X$, cf. (3.1). Then,

$$x^{(N+1)} = \arg\max_{x\in\Omega} \left\{ e(x) \operatorname{dist}(x, X) \right\} \tag{3.19}$$

was used as a selection criterion for an adaptive sampling strategy. In [81], the authors multiplied $e(x)$ with the root mean square error $\operatorname{RMSE}[\widehat{y}(x)] = \sqrt{\operatorname{MSE}[\widehat{y}(x)]}$ and called the criterion a mixed *sample sensitivity error* strategy. It is also zero in every $x^{(i)} \in X$ and

$$x^{(N+1)} = \arg\max_{x\in\Omega} \left\{ e(x) \operatorname{RMSE}[\widehat{y}(x)] \right\} \tag{3.20}$$

was regarded the best method in that study on adaptive and nonadaptive sampling strategies. Using the RMSE instead of the MSE increases the contribution of the exploitation component $e(x)$ compared to the exploration component MSE. See also [117] for a review of the method.

Criterion (3.20) is now illustrated with the example from the previous section, see Fig. 3.6 for the Kriging interpolation of the Branin function. Figure 3.8 depicts the RMSE, which is a poor indicator for the real error $|\widehat{y}(x) - y(x)|$. The cross-validation error $e(x)$ identifies

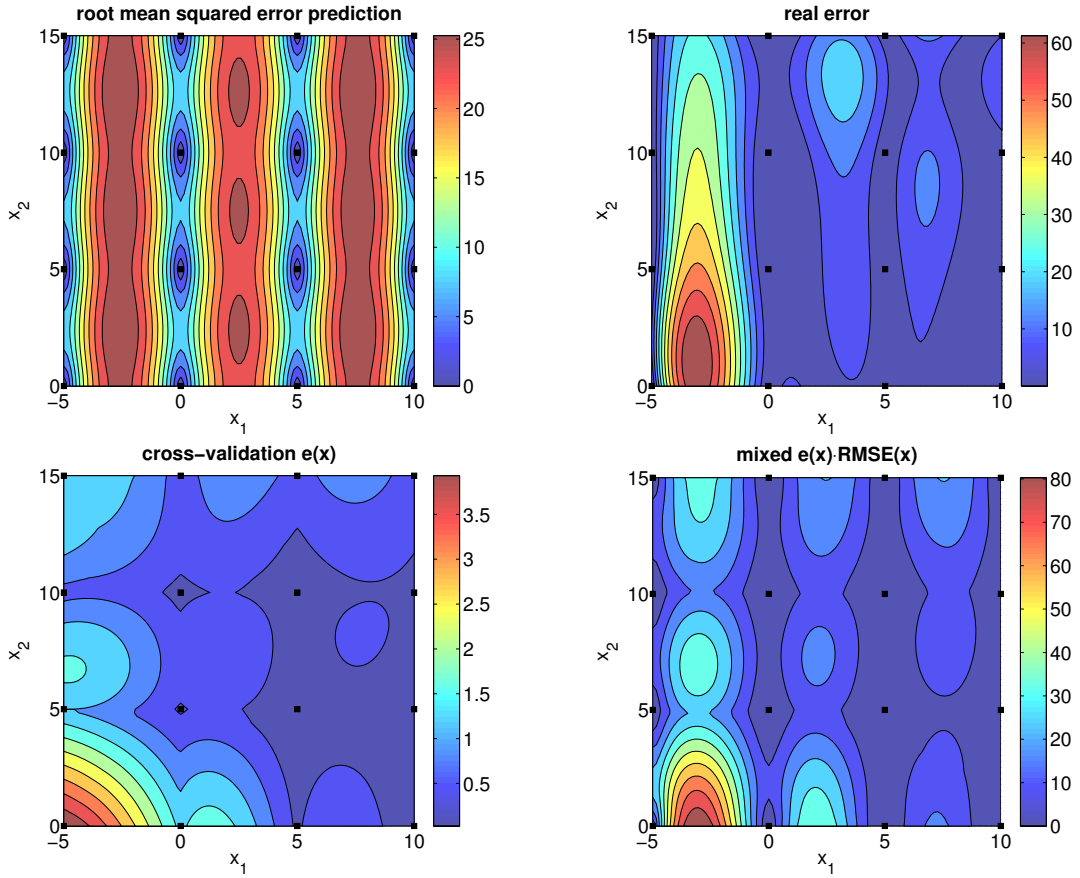

Figure 3.8: RMSE$[\widehat{y}(x)]$ (top left), real error $|\widehat{y}(x) - y(x)|$ (top right), cross-validation error $e(x)$ (bottom right) and mixed error indicator (bottom left) for the interpolation in Fig. 3.6.

the lower left corner as the source of the highest error, but is maximal in an already existing sample $x^{(i)} \in X$. Multiplying $e(x)$ with the RMSE sets the error indicator to zero in all existing samples, while the region with the highest real error is identified correctly.

3.2.3 Adaptive gridding

The methods from the previous sections have some limitations regarding efficiency. Even if the evaluation of the computer experiment $y(x)$ clearly dominates the cost of surrogate model generation (e.g. solving the Kriging equation), finding a new sample $x^{(N+1)} \in \Omega \subset \mathbb{R}^d$ can be a bottleneck in the adaptive process if the number of input parameters d is large. Both MSE-based and cross-validation-based methods require the global optimization of a target function over Ω. The evaluation of this functional is comparably cheap and if d is small, in practice it can just be evaluated on a dense set of candidates for $x^{(N+1)}$ and the candidate with the best score is chosen. After inverting the system matrices once, for arbitrary $x \in \mathbb{R}^d$ the numerical effort of computing MSE$[\widehat{y}(x)]$ is $\mathcal{O}(N^2)$ and for $e(x)$ it is also $N \cdot \mathcal{O}(N-1) = \mathcal{O}(N^2)$, for gradient-enhanced Kriging $\mathcal{O}(N^2d^2)$ and $\mathcal{O}(N^2d)$, respectively. If d is large, this procedure can become computationally out of reach. For example, consider

an application with $d = 6$ variables and the candidates for a new $x^{(N+1)}$ are located on a 50^6-tensorgrid (or factorial design, see Sect. 3.1.1), then $1.5625 \cdot 10^{10}$ evaluations of some target function have to be computed.

Also, practitioners might be interested in adding more than one new sample per stage of the adaptive process. If adding m new samples $x^{(N_k+1)}, \ldots, x^{(N_k+m)}$ to the current design $X^{(N_k)}$, the evaluations $y(x^{(N_k+1)}), \ldots, y(x^{(N_k+m)})$ can be computed independently of each other in parallel. If for instance the parallelization of a computer code for the evaluation $y(x)$ scales extremely poorly with the number of available processors m_{CPU}, scheduling m evaluations simultaneously, using $\frac{m_{\mathrm{CPU}}}{m}$ processors each, can produce benefits regarding the total wall-clock time.

In [11], a new method for adaptive sampling was introduced, which can overcome both of these limitations. It combines a domain decomposition strategy with a local optimal design. Both terms will be described in the following. Ω is assumed to be a hypercuboid, again without loss of generality $\Omega = [0,1]^d$. It is decomposed into equal-sized cells ζ. This is accomplished by equidistant partitionings of the intervals $I_j = [0,1]$ on the x_j-axis into the n_j closed intervals $I_j^1, \ldots, I_j^{n_j}$ $(j = 1, \ldots, d)$. Then

$$\zeta_{i_1,\ldots,i_d} := I_1^{i_1} \times \ldots \times I_d^{i_d} \tag{3.21}$$

are defined and

$$\Omega = \bigcup_{i_1=1}^{n_1} \ldots \bigcup_{i_d=1}^{n_d} \zeta_{i_1,\ldots,i_d} \tag{3.22}$$

holds. Clearly, the lengths of the intervals (and therefore the lengths of the cells' edges) are $|I_j^i| = \frac{1}{n_j}$ for every $j = 1, \ldots, d$. What makes this gridding method adaptive (meaning related to the current dataset) is now choosing the lengths of the cells' edges $|I_j^i|$ proportional to the correlation length $\frac{1}{\theta_j}$ along the x_j-axis, cf. Sect. 2.3.3. So $\frac{|I_j|}{n_j} \sim \frac{1}{\theta_j}$ which is equivalent to $n_j \sim \theta_j |I_j|$ and n_j can be chosen as

$$n_j = \lfloor c\, \theta_j |I_j| \rfloor \tag{3.23}$$

with a constant $c > 0$, $\lfloor \cdot \rfloor$ denotes the floor-function. One ends up with Ω decomposed into equal-sized cells ζ with pairwise disjunct interiors. The hyperparameters are updated in every stage of the adaptive sampling routine, as are the lengths of the cells' edges. The correlation lengths $\frac{1}{\theta_j}$ describe the influence of the value $y_i = y(x^{(i)})$ in the region around $x^{(i)} \in X$. $R(x^{(i)} - x)$ is monotonically decreasing with growing distance $|x_j^{(i)} - x_j|$ in x_j-direction. For instance, using the correlation functions with local support (2.60), (2.63) and (2.72), $R(x^{(i)} - x) = 0$ holds for all $|x_j^{(i)} - x_j| > \frac{1}{\theta_j}$. This corresponds to having many cells along axes with low correlation and few cells along axes with higher correlation, depending on the data Y and therefore on the global behavior of the function $y(x)$.

Now every cell ζ is assigned a cell prediction error η_ζ by cross-validation (3.18). Let $X_\zeta := X \cap \zeta$, then

$$\eta_\zeta := \max_{x^{(i)} \in X_\zeta} \left| \widehat{y}_{-i}(x^{(i)}) - y_i \right| \tag{3.24}$$

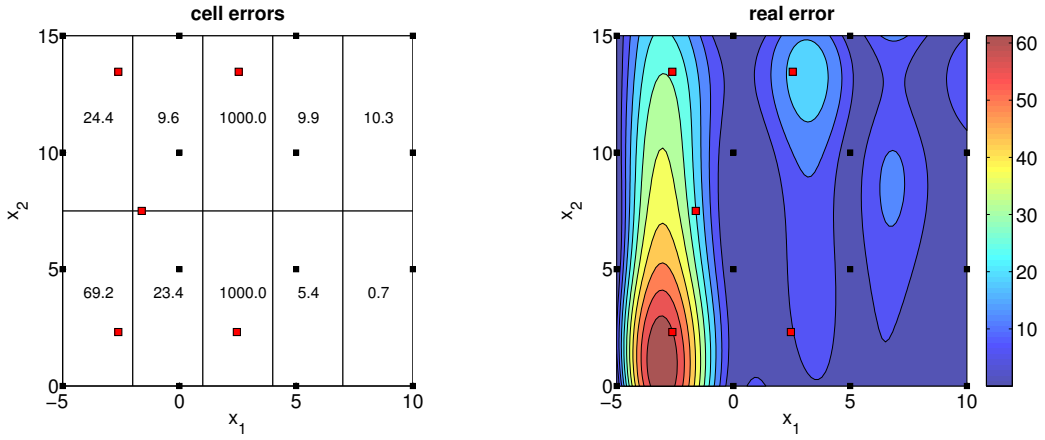

Figure 3.9: Cell errors η_ζ (left) and $|\widehat{y}(x) - y(x)|$ (right) for the interpolation in Fig. 3.6. Red squares depict the 5 new samples.

or, if $X_\zeta = \emptyset$, $\eta_\zeta := \eta_{\max}$. In [11], a target accuracy $\eta^* < \eta_{\max}$ was defined and one new sample is added in every cell ζ with $\eta_\zeta \geq \eta^*$. Alternatively, the $n = \prod_{j=1}^{d} n_j$ cells $\zeta_1, \ldots, \zeta_n$ can be sorted decreasingly by their cell prediction error $\eta_1 \geq \cdots \geq \eta_n$. Then, for a fixed number $m < n$, in the m worst cells $\zeta_1, \ldots, \zeta_m$ a new sample is added. The new samples are obtained by a local optimal design. A maximum entropy criterion is chosen which is equivalent to

$$x^{(N_k+i)} = \arg\max_{x \in \zeta_i} \{\mathrm{MSE}[\widehat{y}(x)]\} \quad (i = 1, \ldots, m), \tag{3.25}$$

cf. Sect. 3.2.1. For the selection of any $x^{(N_k+i)} \in \zeta_i$, not only the current dataset $X^{(N_k)}$ can be used for the computation of the MSE in (3.25), but also the previously selected ones $X^{(N_k)} \cup \{x^{(N_k+1)}, \ldots, x^{(N_k+i-1)}\}$. Otherwise, if 2 adjacent cells ζ_i, ζ_j are identified with a bad prediction error, criterion (3.25) could produce close or even identical $x^{(N_k+i)}, x^{(N_k+j)}$. Finally, one ends up with $X^{(N_{k+1})} = X^{(N_k)} \cup \{x^{(N_k+1)}, \ldots, x^{(N_k+m)}\}$. Summing up, the adaptive gridding method can be classified as a mixed strategy. The domain decomposition adds an exploration component, while using cross-validation-based error estimators adds an exploitation component to the strategy.

Again, the strategy is illustrated with the interpolation of the Branin function (Fig. 3.6). The estimated hyperparameters for the Kriging interpolation are $\theta = (0.593, 0.273)^\top$ (after transformation of Ω to $[0,1]^2$), indicating a weak correlation of $y(x)$ along the x_1-axis and a stronger correlation along the x_2-axis. Using $c = 10$ in (3.23) yields $n_1 = 5$ and $n_2 = 2$ numbers of cells per dimension. The cells ζ and the cross-validation-based predicted cell errors η_ζ (3.24) are depicted in Fig. 3.9, along with the $m = 5$ new samples (red). The two cells in the middle do not contain any samples from the current design $X^{(16)}$, so a maximum error $\eta_{\max} := 1000.0$ is assigned. In the 5 cells with the highest error η_ζ, a new sample is added each by finding the maximum MSE (3.25). Comparing the locations of the new samples $x^{(17)}, \ldots, x^{(21)}$ with the real error $|\widehat{y}(x) - y(x)|$, one can observe that the main sources of surrogate's inaccuracy have been identified.

4 Numerical investigation of sampling strategies

In this chapter, sampling strategies for Kriging and gradient-enhanced Kriging surrogate models are validated with numerical test cases. In Sect. 4.1, the aerodynamic test cases are presented and numerical details of the surrogate models are discussed. Section 4.2 contains a comparison of three space-filling one-stage approaches: factorial designs, Latin hypercube designs and low-discrepancy designs based on the Halton sequence. Additionally, a comparison of Kriging and GEK regarding efficiency is given based on one-stage designs. Section 4.3 is devoted to the investigation of the performance of three adaptive sampling strategies, both in terms of accuracy and distribution of the samples in Ω. For providing a more profound review of these strategies, they are compared to two theoretical sampling strategies which make use of a dense set of validation data in Sect. 4.4. Based on these results, an adaptive sampling strategy for vector-valued responses is discussed in Sect. 4.5. Section 4.6 contains a summary of the most important results of this chapter.

4.1 Numerical test cases

In the aircraft industry, computational fluid dynamics (CFD) have become indispensable for research and development. The airflow around a wing or a complete aircraft configuration can be described by a system of nonlinear partial differential equations. The conservation laws of mass, momentum and energy are known as the Navier-Stokes equations for the state variables density, velocity and energy. If viscous forces are assumed negligible, a simplification is given by the Euler equations for inviscid flow. Discretized versions of these equations can be solved numerically by finite differences, finite volumes or finite elements approaches. For a derivation of the equations and numerical methods see for instance [1, 37]. The German Aerospace center (Deutsches Zentrum für Luft- und Raumfahrt, DLR) developed the flow solvers FLOWer [75] and TAU [127]. FLOWer is a finite volume solver for structured meshes, while TAU is a finite volume solver for unstructured and hybrid meshes.

Following test cases will be investigated in this chapter. For the RAE2822 airfoil, the stationary 2d Euler equations are solved numerically with the FLOWer code using a structured grid of 6369 points and a residual of 10^{-10}, cf. Fig. 4.1. In a second test case, the stationary 2d Reynolds-averaged Navier-Stokes (RANS) equations are solved numerically using the TAU code for an unstructured grid of 27874 points with a Reynolds number of $6.5 \cdot 10^6$, a Spalart-Allmaras turbulence model and a residual of 10^{-6}. The flow solvers are used as a black-box which require the two input parameters Mach number $\mathrm{Ma} \in [0.1, 2.0]$ and angle of attack α, $\alpha \in [-5°, +10°]$ in the Euler case, $\alpha \in [-5°, +8°]$ in the RANS case. Depending on the input parameters $x = (\mathrm{Ma}, \alpha)$, scalar responses are observed. The lift force F_L, the drag force F_D and the moment F_M (cf. Fig. 4.2) can be computed as integrals over the

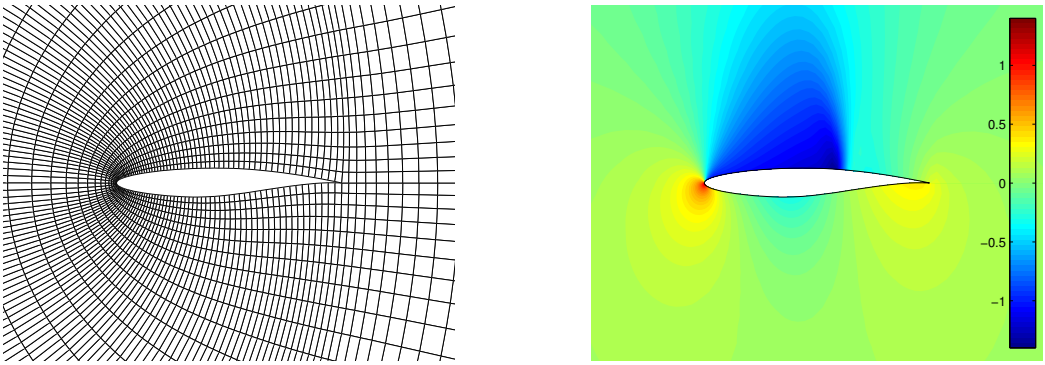

Figure 4.1: Zoom of a structured mesh around the RAE2822 airfoil and solution of the discretized Euler equations with DLR FLOWer, pressure distribution for Mach 0.73, angle of attack $\alpha = 2°$.

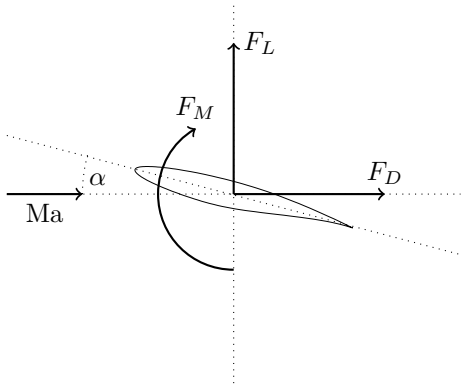

Figure 4.2: Aerodynamic forces and moment on an airfoil.

airfoil's surface based on the flow solution. The aerodynamic coefficients lift c_l, drag c_d and pitching moment c_m are their dimensionless quantities. See e.g. [2] for more information on the aerodynamic coefficients or on aerodynamics in general.

A global representation for the scalar functions $c_l(\mathrm{Ma}, \alpha)$, $c_d(\mathrm{Ma}, \alpha)$ and $c_m(\mathrm{Ma}, \alpha)$ in Ω is sought. On an AMD Opteron 6100 series computer with 2.3GHz processors, a single evaluation of the response in the Euler case takes up to 5 minutes using one processor, depending on the input parameter configuration $x = (\mathrm{Ma}, \alpha)$. In the RANS case however, the evaluation can take up to several hours of CPU time. The responses are assumed to be deterministic. Moreover, the flow solutions converge for all $x \in \Omega$ and a certain regularity of the responses is implied, i.e. $y(x) \in C^1(\Omega)$. For the purpose of comparing the performance of surrogate models regarding their design of experiments, a dense set of validation data is produced. Figures 4.3 and 4.4 depict the responses based on a 50×50

factorial design Ω^{val}. In more complex real-life applications such as a three-dimensional aircraft configuration with the CFD solver depending on several input parameters, said dense evaluations are beyond the limits of tolerable computation time. But already in these two-dimensional test cases depending on two input parameters Ma and α, the difficulties of global approximations are revealed. The input parameter domain Ω covers the sub-, trans- and supersonic flow regimes, where the flow solutions behave strongly different. This leads to different characteristics of the response in different regions of the input parameter domain. While the responses $y(x)$ behave almost linear in the subsonic flow regime and resemble some second order polynomials in the supersonic flow regime, this is not the case in the transsonic flow regime. They behave strongly nonlinear and large gradients as well as large curvature appear. Compared to the Euler case, the nonlinearity is even increased in the RANS case and e.g. multiple local extrema appear for the c_l and the c_m response. Thus, global surrogate modeling is challenging already for only two input parameters ($d = 2$). The responses investigated in this chapter represent relevant test cases, which provide valuable insights into the performance of sampling strategies for global surrogate modeling.

As surrogate models both Kriging and gradient-enhanced Kriging are considered. For the purpose of this study, the partial derivatives required for GEK are computed by finite difference approximations. Note that in more complex industrial applications this is hardly practicable. The computation of the gradient in $x^{(i)}$ requires d additional costly evaluations in $x^{(i)} + \delta e_k$ (e_k denoting the k-th unit vector, $k = 1, \ldots, d$), while the whole gradient can often be accessed much more efficient by one adjoint computation [91, 45]. In the following, let N denote the number of samples. For GEK, suppose in every $x^{(i)}$ the response $y(x^{(i)})$ as well as the partial derivatives with respect to the Mach number $\frac{\partial}{\partial x_1} y(x^{(i)})$ and to the angle of attack $\frac{\partial}{\partial x_2} y(x^{(i)})$ are available. Hence, for GEK the total amount of information is $3N$. As a regression model in the Kriging model (2.27) just a constant term is chosen, i.e. $K = 1$, $f(x) := f_1(x) \equiv 1$. This is also known under the name of ordinary Kriging and usually suffices for the interpolation of highly nonlinear functions, cf. Sect. 2.3.2. As a correlation function, a biquadratic spline (2.72) is chosen. The popular Gaussian correlation function (2.62) suffers from ill-conditioned correlation matrices. In this chapter's test cases, already for $N > 50$ singular matrices with respect to the machine accuracy could be observed when using the Gaussian correlation. In Kriging, also cubic spline functions (2.63) and exponential functions (2.64) were tested with results comparable to the biquadratic splines. In GEK however, exponential functions are not applicable for $p < 2$. In $R''(h, \theta)$ terms with h^{p-2} appear and the second derivative has a singularity in $h = 0$, making it impossible to compute the correlation matrix entries. The biquadratic spline function does not suffer from such limitations. It produces positive definite correlation functions in Kriging and GEK (cf. Sect. 2.3.3 and 2.4.3), which are well-conditioned due to its local support. It is also smooth enough for the use in GEK and its corresponding likelihood function is continuously differentiable ($L(\beta, \sigma^2, \theta) \in C^1(\mathbb{R}^K \times \mathbb{R}_+ \times \mathbb{R}_+^d)$, see Sect. 2.4.4). Regardless of the utilized correlation function, the maximum likelihood problem (2.74) can be challenging to tackle. Especially when the correlation lengths differ (e.g. $\theta_1 \gg \theta_2$), the target function can be ill-conditioned in the optimum. The computational effort of evaluating the target function and its derivatives is dominated by the inversion of R in each iteration, which is $\mathcal{O}(N^3)$ with standard methods. As discussed in Sect. 2.3.4, the full problem $\min_{\beta, \sigma^2, \theta} \{-\log L(\beta, \sigma^2, \theta | Y)\}$ is

considered instead of a reduced formulation. It is solved numerically with a gradient-based Quasi-Newton algorithm using a BFGS update formula for the approximation of the Hessian. A trust region framework is utilized for step size control and for enforcing the constraints $\sigma^2 > 0$, $\theta_k > 0$ $(k = 1, \ldots, d)$ [103]. The input parameter domain Ω is transformed to $[0, 1]^2$ for robustness of the algorithm. The solution is generally found after 30–100 iterations. For problems with up to $N = 100$ samples, the correlation matrix has a maximum size of 100 in Kriging and $100 \cdot (d + 1) = 300$ in GEK and the computation of the solution takes less than one minute in the investigated test cases.

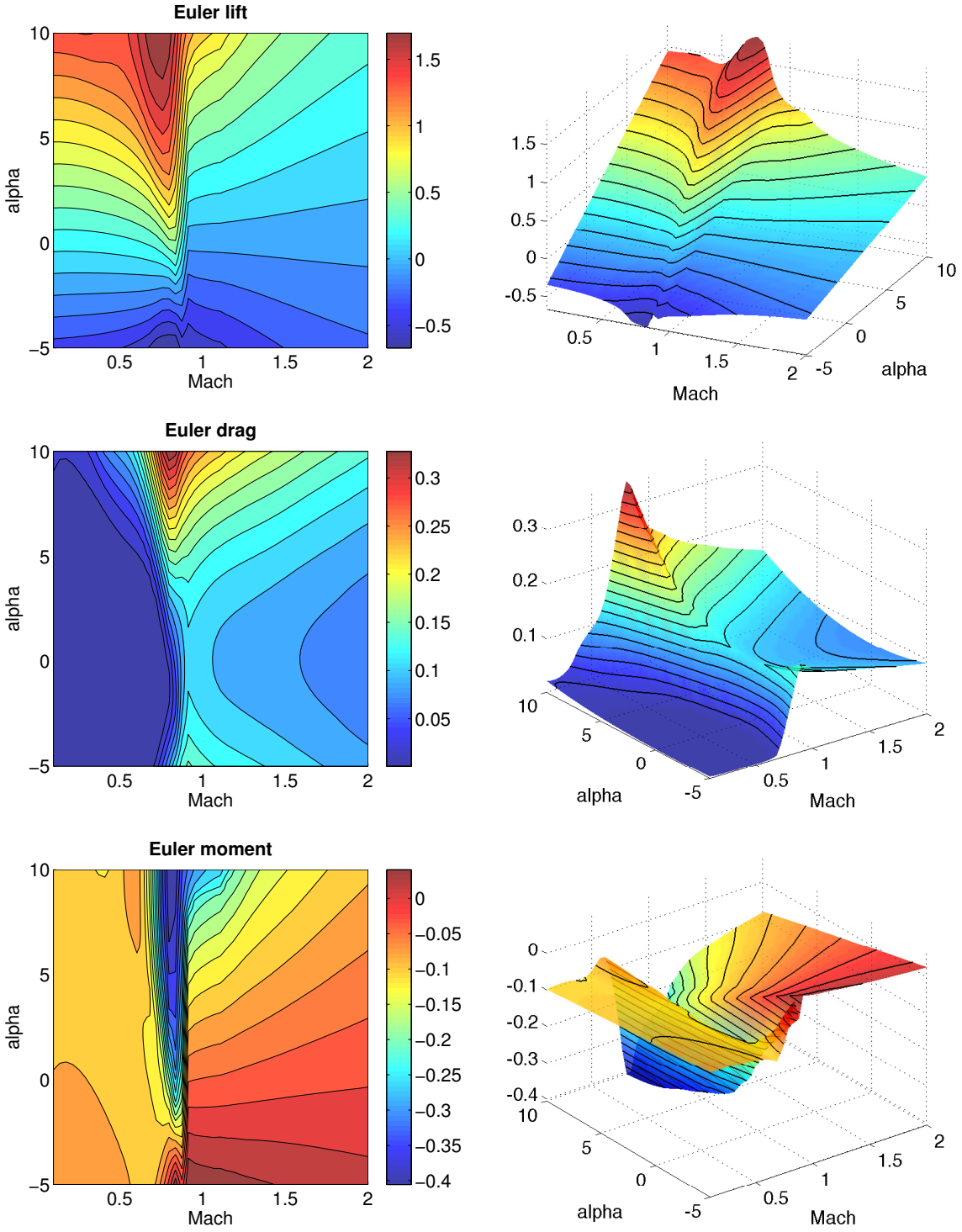

Figure 4.3: Validation data in the Euler case.

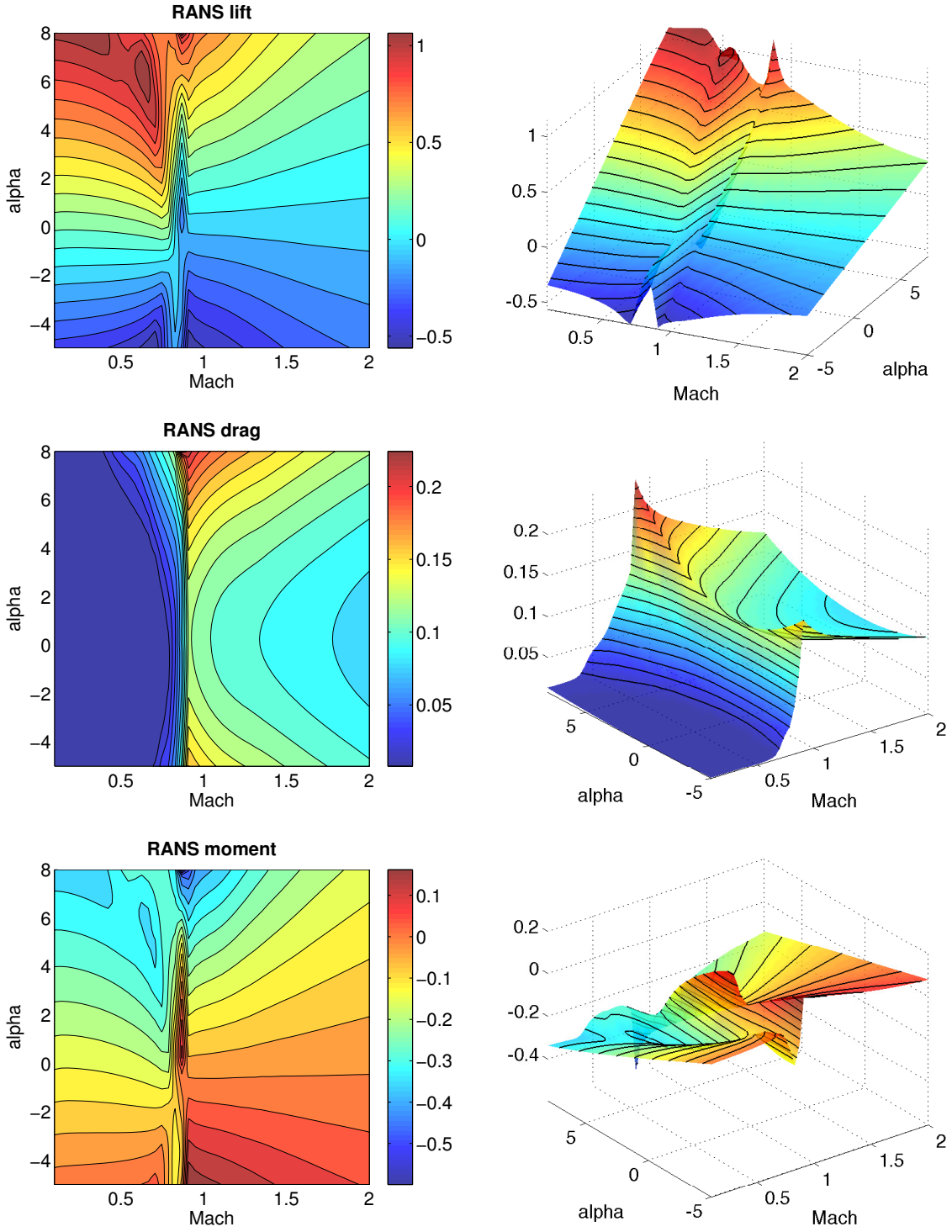

Figure 4.4: Validation data in the RANS case.

4.2 One-stage approaches

4.2.1 Comparison of approximation quality

In a first study, the performance of one-stage designs for Kriging and gradient-enhanced Kriging models for the global interpolation of the aerodynamic coefficients is investigated. Exemplarily, factorial designs, optimized Latin hypercube designs and low-discrepancy designs based on the Halton sequence are compared regarding the approximation quality of the corresponding surrogate models. For the factorial designs, samplings of size $N = 9, 16, 25, \ldots, 100$ are generated using a regular $\sqrt{N} \times \sqrt{N}$-grid in $[0,1]^2$. Pure distance-based design are not considered because of the difficulties of finding an optimal design for arbitrary sample size N. Instead, Latin hypercube (LHC) designs with an optimized maximin-score are used. For $N = 10, 20, 30, \ldots, 100$, 10 different designs each are generated by the Matlab function `lhsdesign`. It offers the feature of maximizing the Latin hypercube design's minimum distance between the samples. For the same sample sizes N as the Latin hypercube designs, Halton designs are generated with the Matlab function `haltonset`. Model-based designs are not considered for one-stage approaches because the Kriging model's hyperparameters θ are not available in the design phase.

For assessing the global approximation quality of the surrogate models, the Kriging and GEK interpolations are compared to the validation data on Ω^{val}. A relative averaged error η_1, a relative root mean squared error η_2 and a relative maximum error η_∞ are defined by

$$\eta_1 := \frac{1}{\sigma} \frac{1}{\#\Omega^{\mathrm{val}}} \sum_{x \in \Omega^{\mathrm{val}}} \left| \widehat{y}(x) - y(x) \right|, \tag{4.1}$$

$$\eta_2 := \frac{1}{\sigma} \left(\frac{1}{\#\Omega^{\mathrm{val}}} \sum_{x \in \Omega^{\mathrm{val}}} \left(\widehat{y}(x) - y(x) \right)^2 \right)^{\frac{1}{2}}, \tag{4.2}$$

$$\eta_\infty := \frac{1}{\sigma} \max_{x \in \Omega^{\mathrm{val}}} \left| \widehat{y}(x) - y(x) \right|. \tag{4.3}$$

Here, σ denotes the standard deviation of the validation data set $\{y(x)\}_{x \in \Omega^{\mathrm{val}}}$

$$\sigma := \left(\frac{1}{\#\Omega^{\mathrm{val}} - 1} \sum_{x \in \Omega^{\mathrm{val}}} \left(y(x) - \overline{y} \right)^2 \right)^{\frac{1}{2}}, \tag{4.4}$$

where its mean is

$$\overline{y} := \frac{1}{\#\Omega^{\mathrm{val}}} \sum_{x \in \Omega^{\mathrm{val}}} y(x). \tag{4.5}$$

The relative root mean squared errors η_2 (4.2) for the Kriging and GEK interpolations of the lift $c_l(\mathrm{Ma}, \alpha)$, the drag $c_d(\mathrm{Ma}, \alpha)$ and the pitching moment $c_m(\mathrm{Ma}, \alpha)$ are depicted in Fig. 4.6 for the Euler case and in Fig. 4.7 for the RANS case. The average LHC performance is computed as the mean of the 10 LHC designs which are considered for each sample size N. Except for the Halton designs, the samplings are not nested, i.e. $X^{(N_1)} \not\subset X^{(N_2)}$ for $N_1 < N_2$. But typically, a higher approximation quality is expected with a larger sample

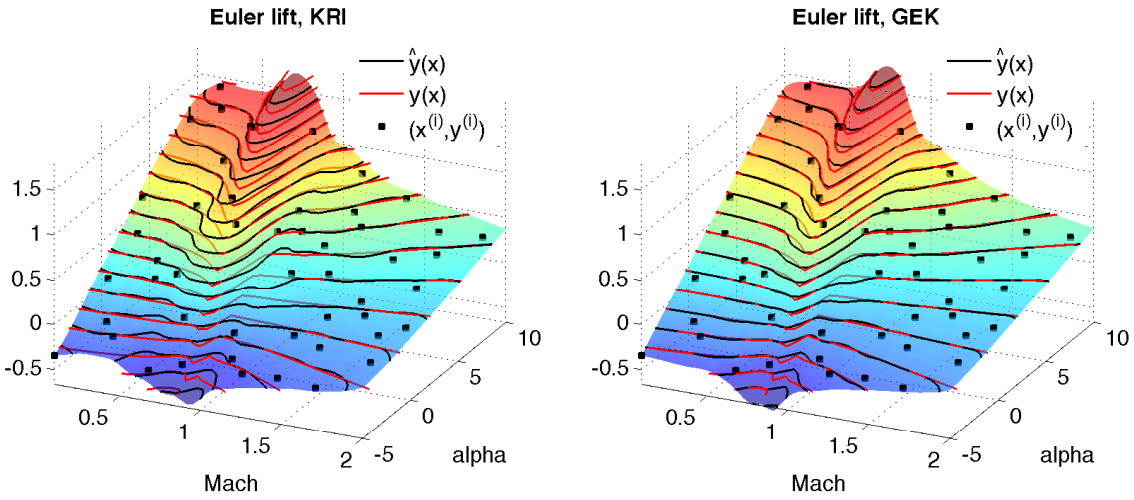

Figure 4.5: Interpolations of $c_l(\mathrm{Ma}, \alpha)$ based on a Halton design, $N = 50$.

size N. This is not the case for the *factorial designs* in general. Compared to the other two strategies, these designs produce the poorest interpolations in the investigated test cases. Also, the error η_2 is not necessarily reduced for larger sample sizes. For instance, in most cases the error for $N = 100$ is is only slightly lower or even higher than for $N = 50$, although double the amount of information is used for $\widehat{y}(x)$. For the *Latin hypercube designs*, generally an improvement of the average approximation quality can be observed with growing sample size N. The corresponding surrogates also mostly perform second-best. For $N < 50$, in most cases, the errors η_2 are comparable to the Halton designs' errors. For $N > 50$, the surrogate models based on the *Halton designs* are most accurate and outperform both the LHC and the factorial designs. These results mainly reproduce numerical results found in the literature. In other applications low-discrepancy designs also proved to be superior to Latin hypercube designs, cf. [68, 131, 81, 60].

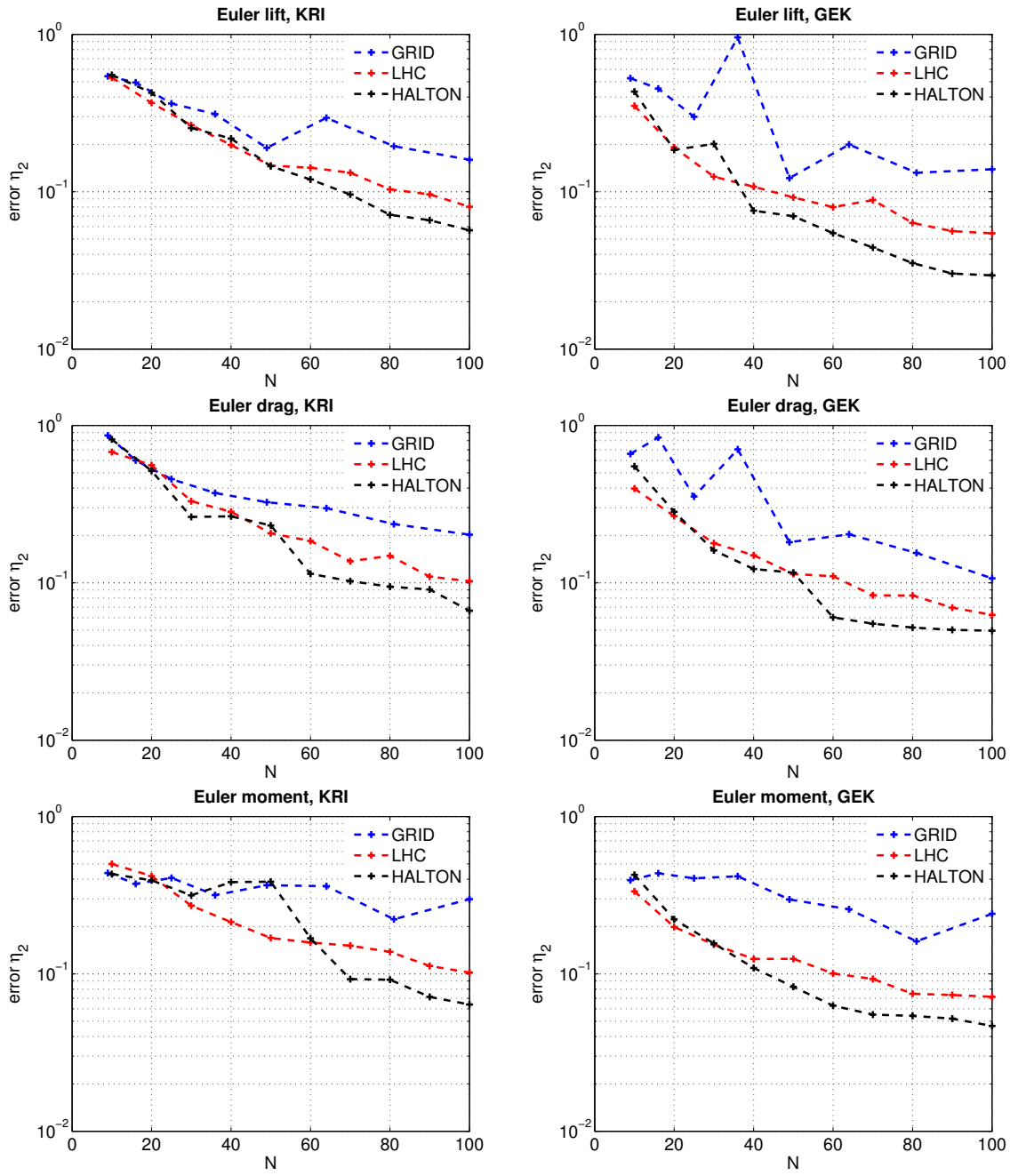

Figure 4.6: Root mean squared error η_2 for interpolation of FLOWer Euler responses.

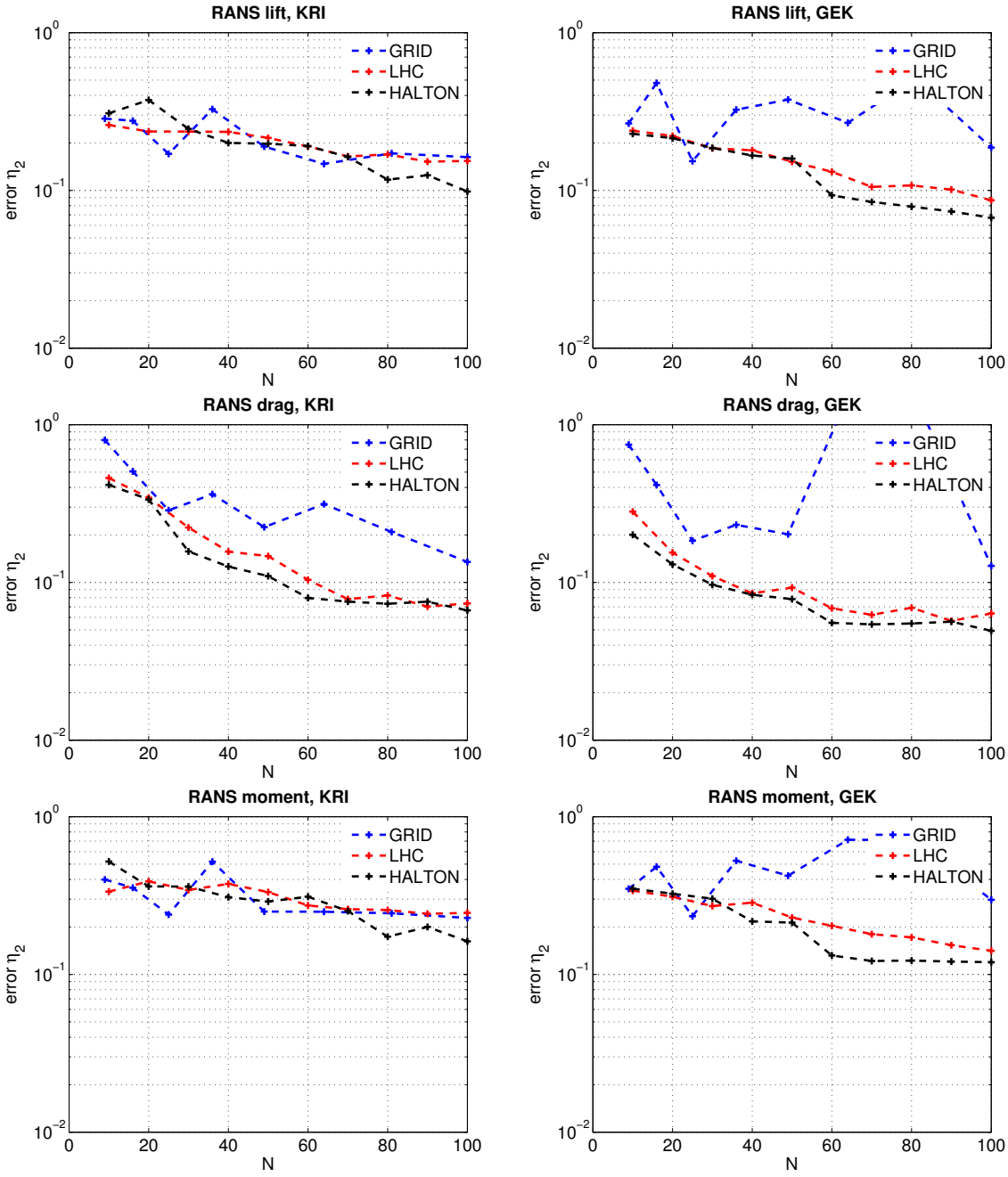

Figure 4.7: Root mean squared error η_2 for interpolation of TAU RANS responses.

4.2.2 Benefits of GEK

A brief comparison of Kriging and gradient-enhanced Kriging regarding efficiency is now presented on the basis of the investigated test cases. It can generally be observed that, for a given design of size N, the GEK interpolation is more accurate than the Kriging interpolation (Figs. 4.5–4.7). But obviously, in GEK more computational effort has to be brought up for generating not only the evaluations $y(x^{(i)})$, but also the gradient information $\frac{\partial}{\partial x_k} y(x^{(i)})$ $(i = 1, \ldots, N,\ k = 1, \ldots, d)$. So for investigating the benefit of GEK compared to Kriging, a review of the surrogate models' accuracy vs. the numerical effort has to be considered, instead of reviewing the accuracy vs. the number of samples N. In this section's interpolations the gradients $\nabla y(x^{(i)}) = (\frac{\partial}{\partial x_1} y(x^{(i)}), \ldots, \frac{\partial}{\partial x_d} y(x^{(i)}))^\top$ are generated by finite difference approximations, which means one additional expensive evaluation of the flow solver $y(x)$ per dimension $k = 1, \ldots, d$. But if adjoints are used, the complete gradient $\nabla y(x)$ can be obtained by one additional adjoint computation regardless of the number of input parameters d. A theoretical best case study is considered, which means that the use of adjoints is assumed. For the investigated test cases suppose that, with twice as many computations (flow solutions plus adjoint computations), three times the amount of information (responses plus partial derivatives with respect to Ma and α) can be obtained. Once more, the numerical effort of generating the surrogate model is neglected, because it is assumed to be insignificant compared to the evaluation of the response. In Fig. 4.8, the interpolation error η_2 (4.2) is plotted vs. the number of computations for Kriging and GEK based on Halton designs. The GEK data is the same as in Figs. 4.6 and 4.7, now with N samples corresponding to $2N$ computations. The Kriging data is the same with N samples corresponding to N computations and additional data for $N = 110, 120, \ldots, 200$. The pictures clearly show that the approximation quality is basically comparable for Kriging and GEK for the same amount of numerical effort (assuming adjoint computations). However, taking the actual numerical effort of these test cases as a basis, the numerical effort for GEK is $3N$ (finite differences) instead of $2N$ (adjoints) and GEK is less efficient than Kriging. This two-dimensional survey reveals that GEK cannot be more efficient than Kriging unless the gradient information is obtained by adjoints or automatic differentiation. Also, the benefit of GEK using adjoints compared to Kriging is assumed to grow with the number of input parameters d, since $(d+1)$ times the amount of information can be included in the GEK surrogate model using only twice as many computations (response evaluations plus adjoints).

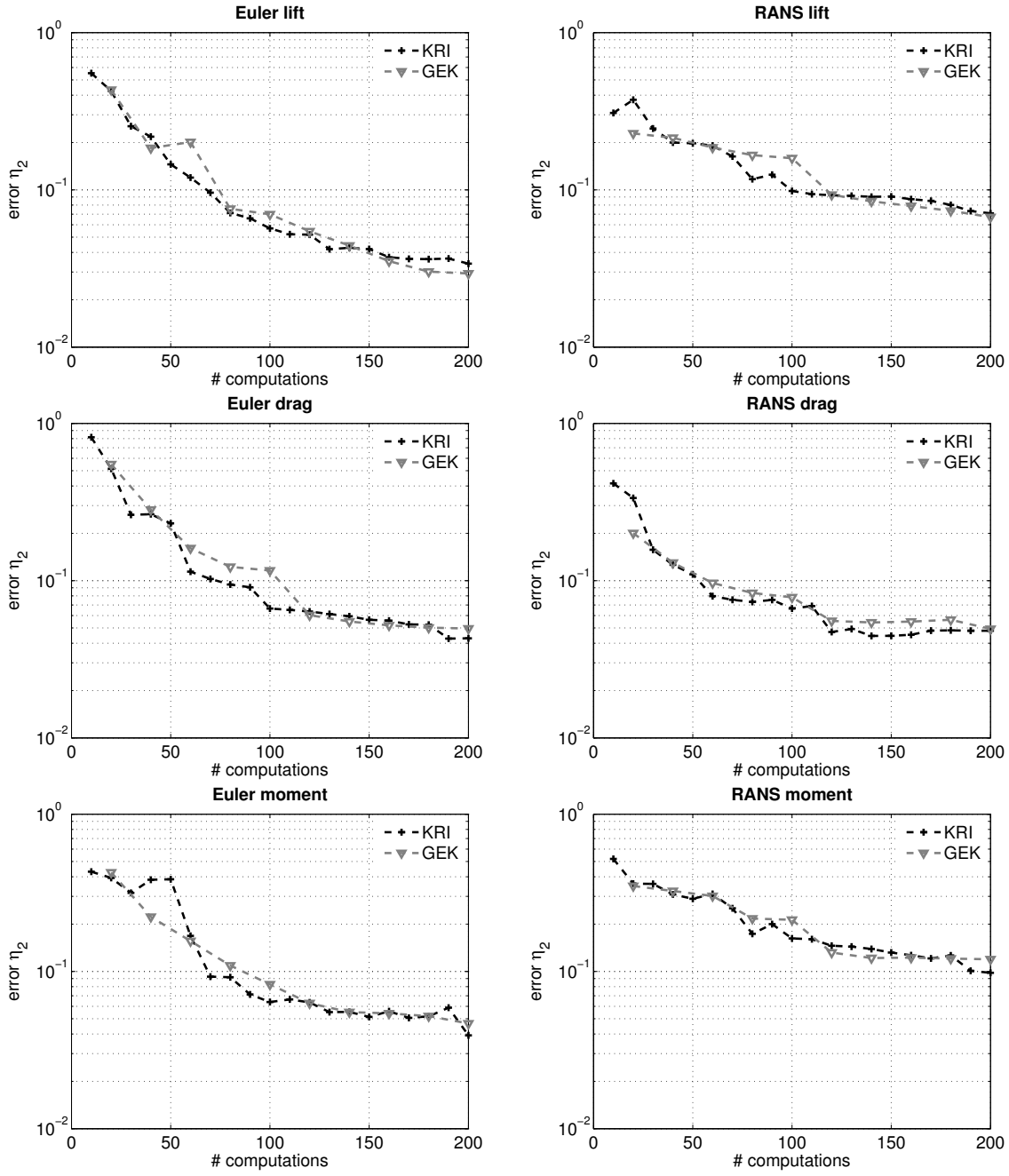

Figure 4.8: Root mean squared error η_2 vs. numerical effort for interpolations of FLOWer Euler (left) and TAU RANS (right) responses based on Halton designs.

4.3 Adaptive designs

This section is devoted to the study of the performance of adaptive sampling strategies for Kriging and gradient-enhanced Kriging in the aerodynamic test cases. The accuracy of the surrogates based on Halton designs from the previous section serves as a benchmark for the adaptive strategies. The performance of three sequential designs is assessed and the distribution of the samples is discussed. The maximum MSE strategy (3.16) is considered exemplarily for the purely exploration-based MSE strategies. As a mixed strategy between exploration and exploitation, the cross-validation-based sample sensitivity error (SSE) strategy (3.20) is investigated. Adaptive gridding (ADG) is also a mixed strategy (3.25), except that it is capable of adding multiple new samples in every stage of the adaptive process. As an initial design $X^{(N_0)}$, a 3×3 factorial design is chosen. Each sampling strategy then adaptively adds samples until a maximum number of $N = 100$ samples is reached. For applying the selection criteria, in the MSE and the SSE method a global optimization problem has to be solved and in the ADG method multiple localized optimization problems with constraints have to be solved. These problems classically have numerous local optima and global optimization is challenging. Instead of solving these demanding optimization problems numerically, a dense set of candidates Ω^{val} is considered, from which the new sample $x^{(N_k+1)}$ with the best selection criterion score is chosen. Note that in the SSE and the ADG method using GEK, for performing cross-validation the reduced surrogate models $\widehat{y}_{-i}(x)$ are generated by not only deleting $y(x^{(i)})$ from the data, but also the partial derivatives $\frac{\partial}{\partial x_k} y(x^{(i)})$ $(i = 1, \ldots, N,\ k = 1, \ldots, d)$. In ADG, the number of new samples per stage is set to $m = 10$. The factor for defining the number of cells n_j per dimension in (3.23) is chosen as $c = 5$. It is augmented if necessary to ensure that at least two cells per axis are defined $(n_j \geq 2)$ and that the total number of cells does not fall below the number of samples to add $(n = n_1 \cdot n_2 \geq m)$. The number of cells n_j is proportional to the hyperparameter θ_j, so for other correlation functions (linear, exponential, Gaussian, cf. Sect. 2.3.3) other best practice constants c have to be determined. In a similar study [117], the author of this thesis compared adaptive sampling strategies to Latin hypercube designs and the remainder of this section clarifies if they can also outperform the superior Halton designs and provides a more comprehensive examination of the sampling distributions.

4.3.1 Comparison of approximation quality

The performance of the adaptive sampling strategies is now reviewed and compared to nonadaptive Halton designs. In Figs. 4.9 and 4.10, the validation-data-based relative root mean squared error η_2 (4.2) is plotted for the Kriging and the GEK interpolations of the aerodynamic functions in the Euler and in the RANS case. Comparing all relative error levels, the lift responses can be approximated most accurately, while the interpolations of the drag responses perform second-best and the moment responses seem to be most difficult to interpolate. Also, the global approximation quality is generally higher in the Euler case than in the RANS case, which is not a surprise, comparing the validation data for both test cases (cf. Figs. 4.3 and 4.4). As expected, the error commonly decreases with increasing sample size N. In some cases, the error η_2 can increase when adding a new sample to the interpolation. This may be due to a local variation of the new surrogate model's behavior

around the new sample $x^{(N_k+1)}$ or by a change of the global behavior of $\widehat{y}^{(N_k+1)}$ compared to $\widehat{y}^{(N_k)}$ because the hyperparameters θ have changed. In most cases, the error is reduced again after adding few more new samples. The error levels clearly indicate a hierarchy between the three investigated adaptive sampling strategies. For the better part of the test cases, the MSE-based surrogates perform poorest, while adaptive gridding performs second best and the cross-validation-based SSE strategy produces the most accurate interpolations. After some starting phase ($N < \widetilde{N}$) which can differ from case to case, this hierarchy holds for most $N > \widetilde{N}$. The accuracy levels of the individual adaptive designs are reviewed in detail in the following.

The interpolations defined by the pure exploration-based *MSE strategy* are more accurate than the ones based on the nonadaptive Halton designs or perform at least comparable in most cases for small sample sizes (e.g. for $N < 50$). For larger designs however, they mostly cannot compete with the Halton sequence. When the maximum number of $N = 100$ samples is reached in the Euler case, the MSE-based surrogates have a larger error than the nonadaptive benchmark (with one exception). In the RANS case, they perform slightly better (also with one exception). Overall, in these two-dimensional test cases, it is only recommended to use MSE-based designs for moderate sample sizes. For larger sample sizes, the Halton sequence seems to provide a better space-filling quality than the MSE method (cf. Sect. 4.3.2). Regarding efficiency, the nonadaptive Halton sequence seems more profitable, because for any given sample size N, all evaluations Y can be computed in parallel whereas for the MSE method all evaluations must be computed sequentially without a notable improvement of the accuracy.

For the assessment of the (mixed exploration- and exploitation-based) *adaptive gridding strategy*, one has to differentiate between its performance in the Euler and in the RANS case. In the Euler case, it can outperform the Halton designs already for small sample sizes. The approximation error generally decreases in each stage of the adaptive process after adding $m = 10$ new samples. In the RANS case, the adaptive strategy's advantage over the nonadaptive Halton sequence is not that obvious. However, it mostly generates more accurate surrogate models with one exception. In all other cases, after reaching its maximum number of $N = 99$ samples, the approximation quality is higher or at least comparable to the Halton design.

The (also mixed) *sample sensitivity error strategy* overall produces the best results. In some cases, a small staring phase is required, after which its surrogate models outperform the nonadaptive Halton designs as well as MSE designs and mostly all ADG designs. Except for the c_d response in the RANS case, the interpolations based on the SSE strategy are most accurate for all $N > \widetilde{N}$ for some case-dependent $\widetilde{N}$. Using not only an exploration component in the selection criterion ($\mathrm{RMSE}[\widehat{y}(x)]$), but combining it with a cross-validation-based exploitation criterion, which automatically identifies regions with a poor approximation quality, the method proves best in the aerodynamic test cases. Its accuracy outmatches not only the nonadaptive benchmark, but also the pure exploration method (MSE) and the second mixed strategy (ADG).

Some comments about the bad performance of adaptive sampling strategies for the c_d response in the RANS case are now given. Investigating the true response's validation data (Fig. 4.4), one can observe a very steep ridge at approximately Mach 0.9, which is steeper than in the Euler case and may even not be differentiable. When error-based adaptive

sampling strategies try to dissolve this strong curvature, many samples are added in the area around Mach 0.9 (cf. the subsequent discussion of sample distributions). Also, the hyperparameter estimation chooses a very large θ_1, which is equivalent to a small correlation length $\frac{1}{\theta_1}$, i.e. the influence of a response evaluation y_i decreases rapidly along the Ma-direction with growing distance to the corresponding sample $x^{(i)}$. If the correlation length becomes too small, this can result in a poor interpolation quality in other areas with few samples like the supersonic flow regime, cf. also Fig. 2.8 in the section on hyperparameters 2.3.4. This issue shows that the capacity of global surrogate models is limited. If the response's behavior differs strongly in different regions of the input parameter domain (e.g. different levels of smoothness), a single surrogate model using global hyperparameters might not be the right choice. It generally raises the question, if the response y satisfies the model assumptions, particularly if z in $y(x) = f(x)^\top \beta + z(x)$ is a stationary Gaussian process (cf. Sect. 2.3.1). Stationarity means that the correlation function $R(x, \tilde{x})$ only depends on the distance vector $\tilde{x} - x$ and not on the the positions x and $\tilde{x}$ in space. For approaches using non-stationary covariance models, the reader is referred to [148, 48, 49]. While in the former one a non-linear mapping of the input parameter space was used, in the latter two a treed partitioning was applied to Ω and distinct stationary models were used in each partition.

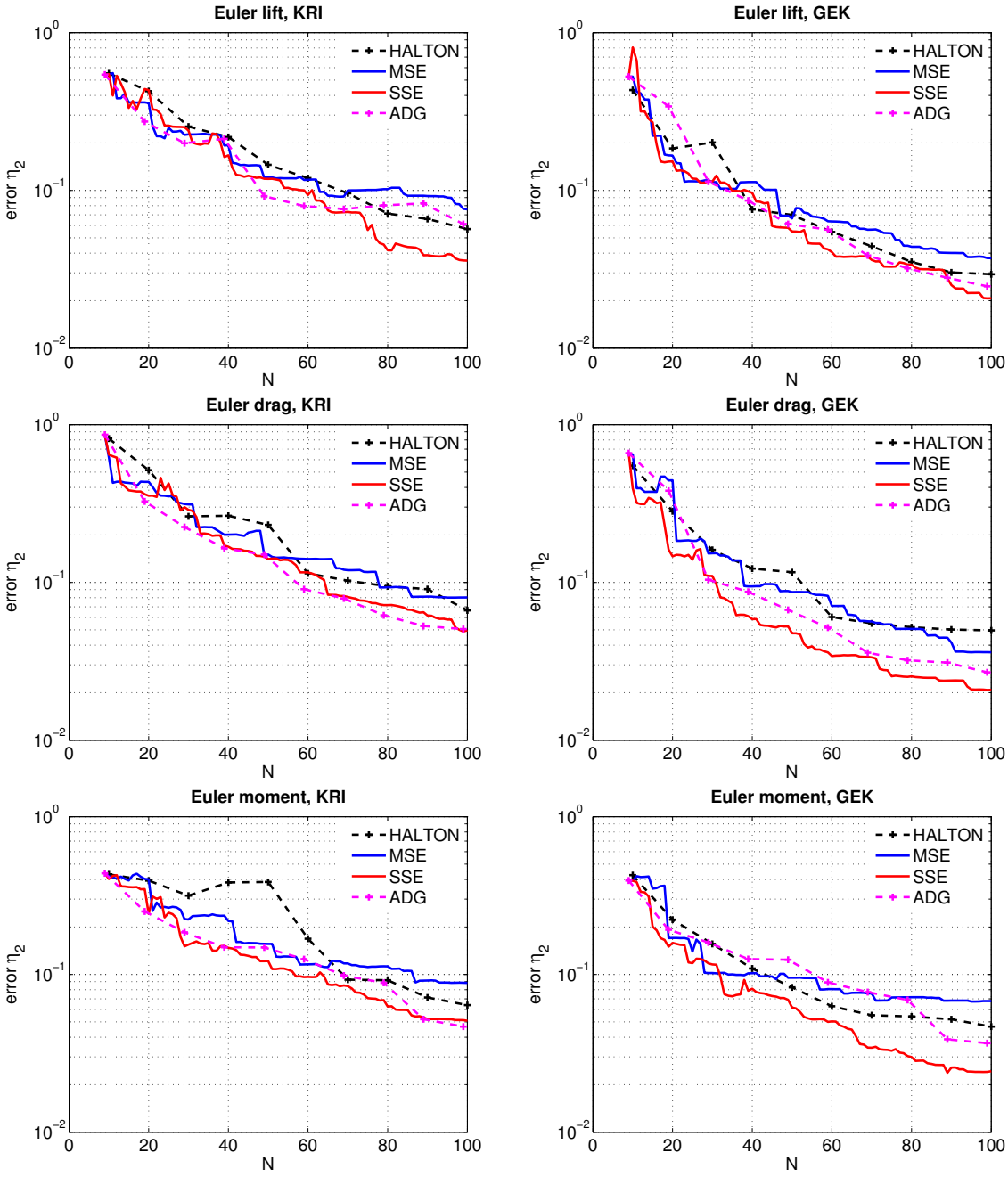

Figure 4.9: Root mean squared error η_2 for interpolation of FLOWer Euler responses.

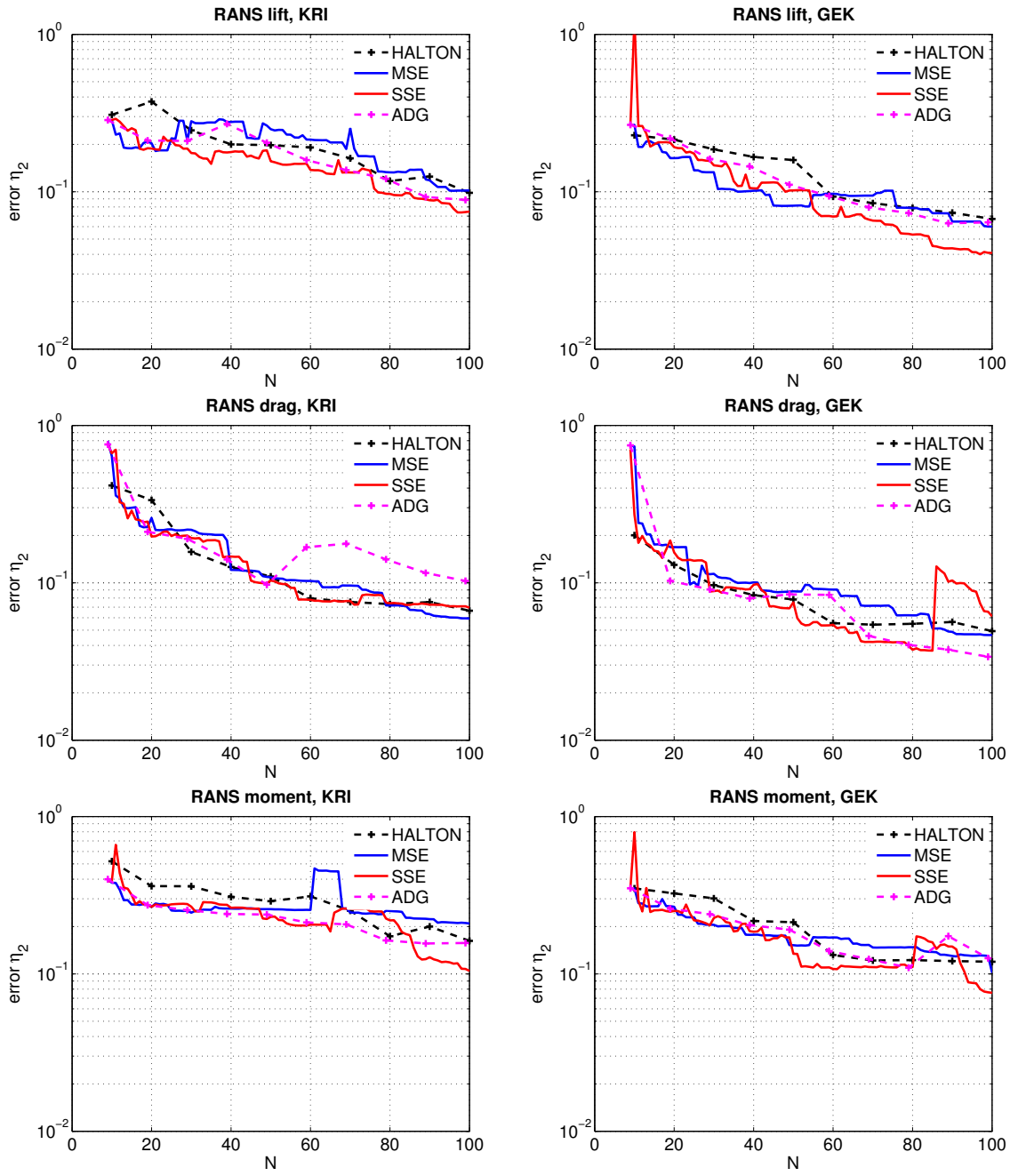

Figure 4.10: Root mean squared error η_2 for interpolation of TAU RANS responses.

4.3.2 Comparison of sample distributions

For a more thorough investigation of the adaptive designs, the distribution of the samples X in the input parameter domain Ω is discussed. Figures 4.11–4.16 depict the samplings for all test cases combined with a contour line plot of the final surrogate model $\widehat{y}(x)$ based on the particular design for $N = 100$ ($N = 99$ in ADG, respectively). The initial 3×3 factorial design is depicted in black squares, while the sequentially added samples are depicted in colored lozenges which correspond to the color of previous section's error plots.

The exploration-based *MSE strategy* (Figs. 4.11 and 4.12) aims at a space-filling design. The samples are equally distributed in the sub-, trans- and supersonic flow regime without any preference for previously identified critical transsonic region of the input parameter domain. This is plausible, since the information on the unknown $y(x)$ contained in the evaluations Y is not involved locally in the selection criterion, but only globally via the hyperparameters $\theta = (\theta_1, \theta_2)$, cf. Sect. 3.2.1. Of particular interest is the concentration of samples on the upper and lower α-boundary of the input parameter domain, while only few additional samples are located on the Ma-boundaries. The reason for this can be found in the strong correlation of the evaluations along the α-axis, while they are only weakly correlated along the Ma-axis ($\frac{1}{\theta_1} < \frac{1}{\theta_2}$). In the Euler case for instance, more samples are located on the α-boundaries in the design for c_d than for c_l, and even more for c_m. This corresponds to the hyperparameters' ratio $\frac{\theta_1}{\theta_2}$ in Table 4.1: the higher the ratio, the more samples are located on the α-boundaries. This behavior of a MSE-based strategy was also reported in [107]. At first sight, this property seems reasonable. A strong correlation along an axis indicates a linear behavior and, recalling that Kriging with local basis functions is rather an interpolation than an extrapolation technique, covering the boundaries of the identified axis with samples seems just appropriate. In a previous study [117], this was diagnosed as a reason for outperforming the nonadaptive Latin hypercube designs. Comparing the MSE designs to the nonadaptive Halton designs regarding accuracy, they can only compete for smaller designs. For larger designs however, it seems that the distribution of samples in the interior of Ω is too sparse or rather little intelligent. The Halton sequence, which is more evenly distributed in whole Ω (cf. Sect. 3.1.4), then mostly produces more accurate surrogate models.

	θ_1	θ_2	ratio $\frac{\theta_1}{\theta_2}$	samples on $\alpha_{\min,\max}$
c_l	1.481	0.438	3.381	30
c_d	2.062	0.399	5.168	35
c_m	3.099	0.440	7.043	54

Table 4.1: Hyperparameters for Kriging interpolation based on MSE design, $N = 100$ (Euler case).

The samples of the *sample sensitivity error strategy* (Figs. 4.13 and 4.14) are concentrated higher in and around the transsonic region, where the responses behave much less linear than for smaller or higher Mach numbers. Nonetheless, the sub- and supersonic regions are not neglected completely, but only provided with less samples. The concentration is generally stronger in GEK than in Kriging, since the GEK surrogate model contains more information on the behavior of $y(x)$. Also, like in the MSE-based designs, many samples are added on the

α-boundaries. Evidently, this is due to the influence of the RMSE component (exploration component) in the selection criterion (3.20). But unlike the MSE strategy, the samples in the interior of Ω are distributed more intelligent. Critical regions are identified automatically due to the cross-validation component (exploitation component). This combination of a local or exploitation component and a global or exploration component produces overall the most accurate surrogate models (cf. Sect. 4.3.1).

The distribution of the *adaptive gridding* samples (Figs. 4.15 and 4.16) is similar to the SSE strategy. Many samples are located on the boundaries of the α-axis, while in the interior of Ω, generally a higher concentration of samples in the critical transsonic region can be observed. In the RANS case, the concentration is less distinct than for the SSE strategy, but beyond that no structural difference is detected. Using also both an exploration and an exploitation component for selecting new samples, this characteristic is plausible. Nevertheless, the SSE designs generally perform better. The difference is that the SSE strategy adds only one sample per stage and not $m = 10$ at once. Adding multiple samples per stage can produce benefits regarding efficiency (by parallel evaluation). But the sampling strategy cannot react that precisely on changes of the N-th surrogate model $\widehat{y}^{(N)}(x)$ as if adding just one instead of m samples and possibly evaluations are wasted on uncritical regions of the input parameter space.

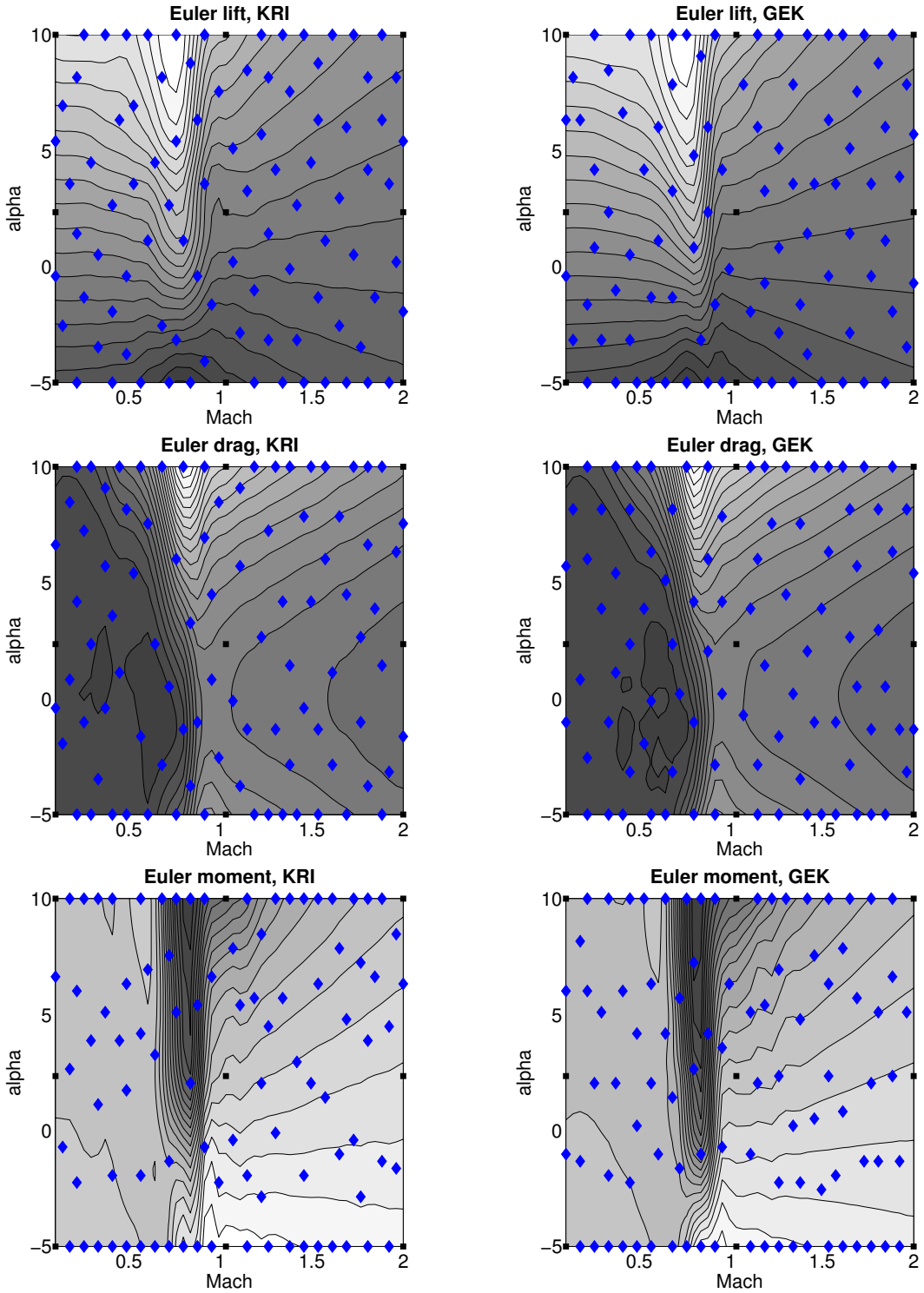

Figure 4.11: Sample distributions for MSE strategy, FLOWer Euler responses.

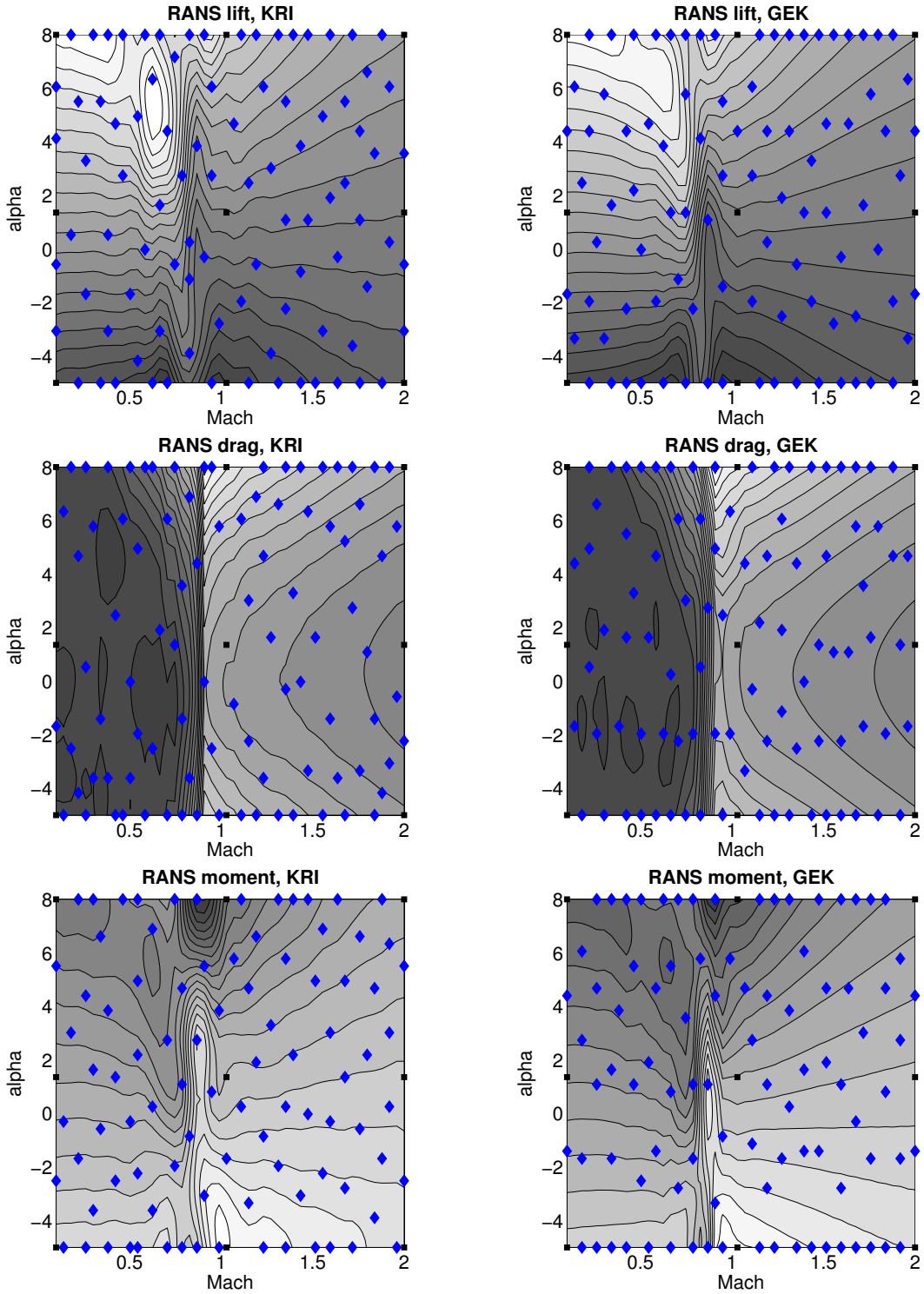

Figure 4.12: Sample distributions for MSE strategy, TAU RANS responses.

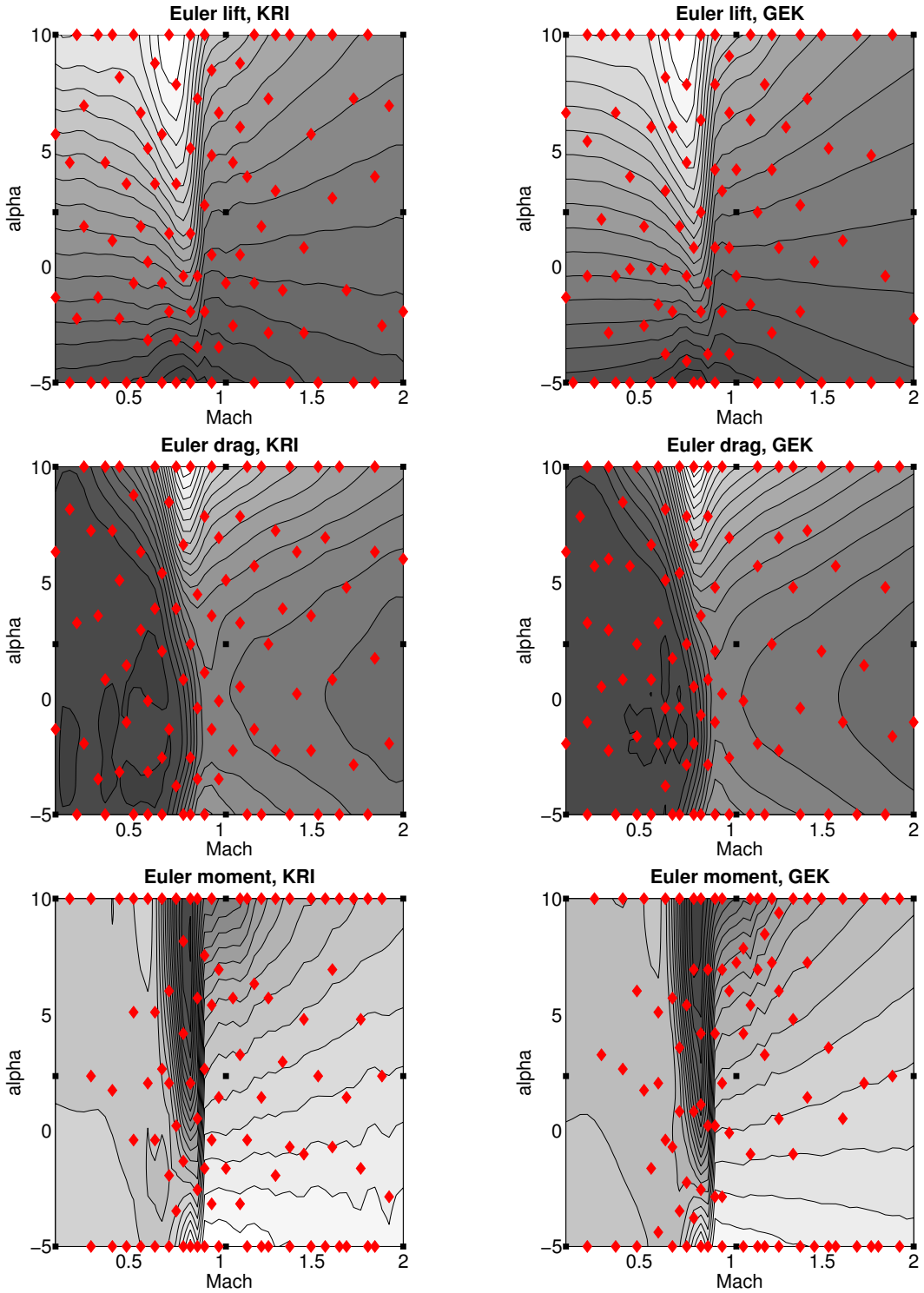

Figure 4.13: Sample distributions for SSE strategy, FLOWer Euler responses.

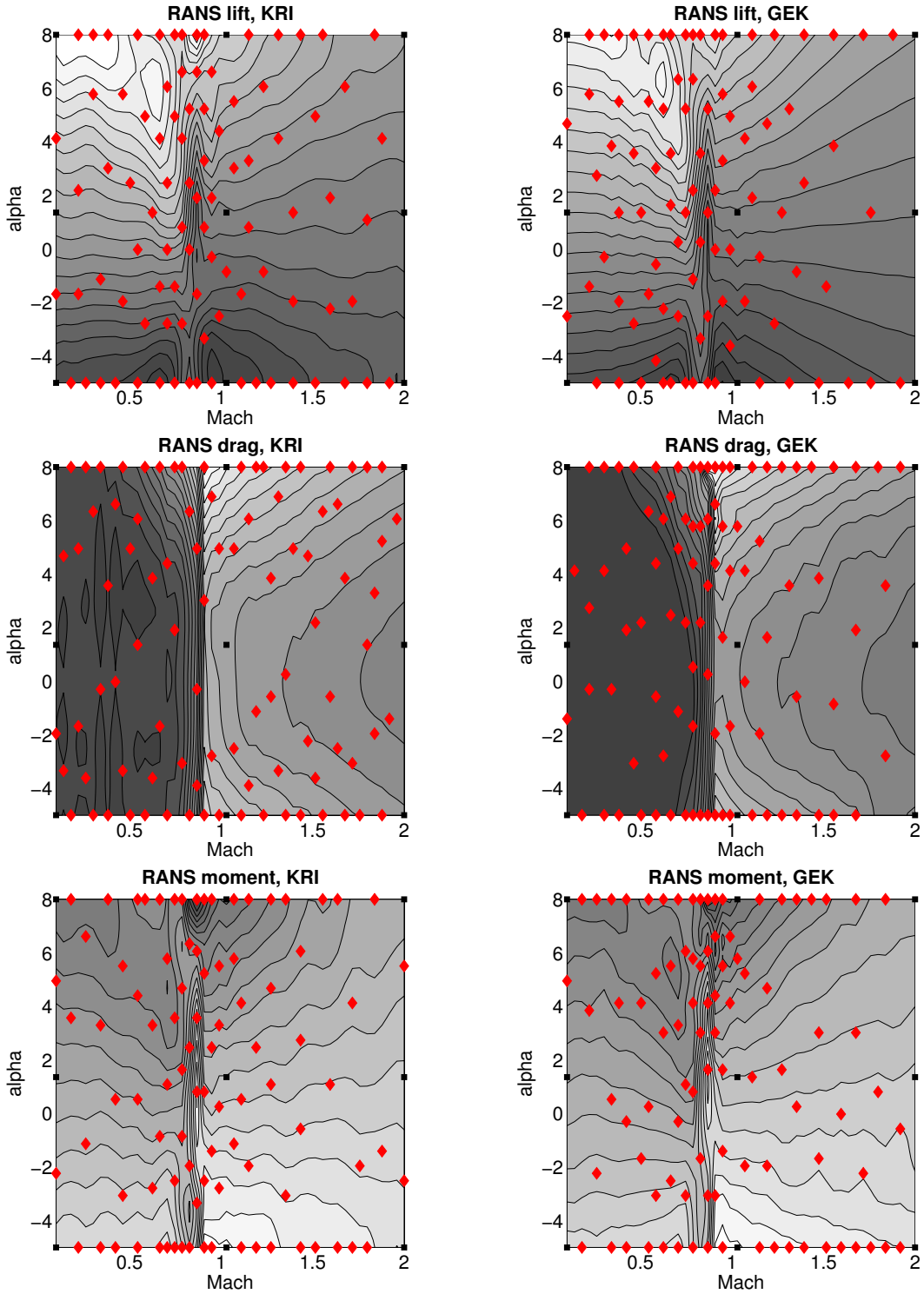

Figure 4.14: Sample distributions for SSE strategy, TAU RANS responses.

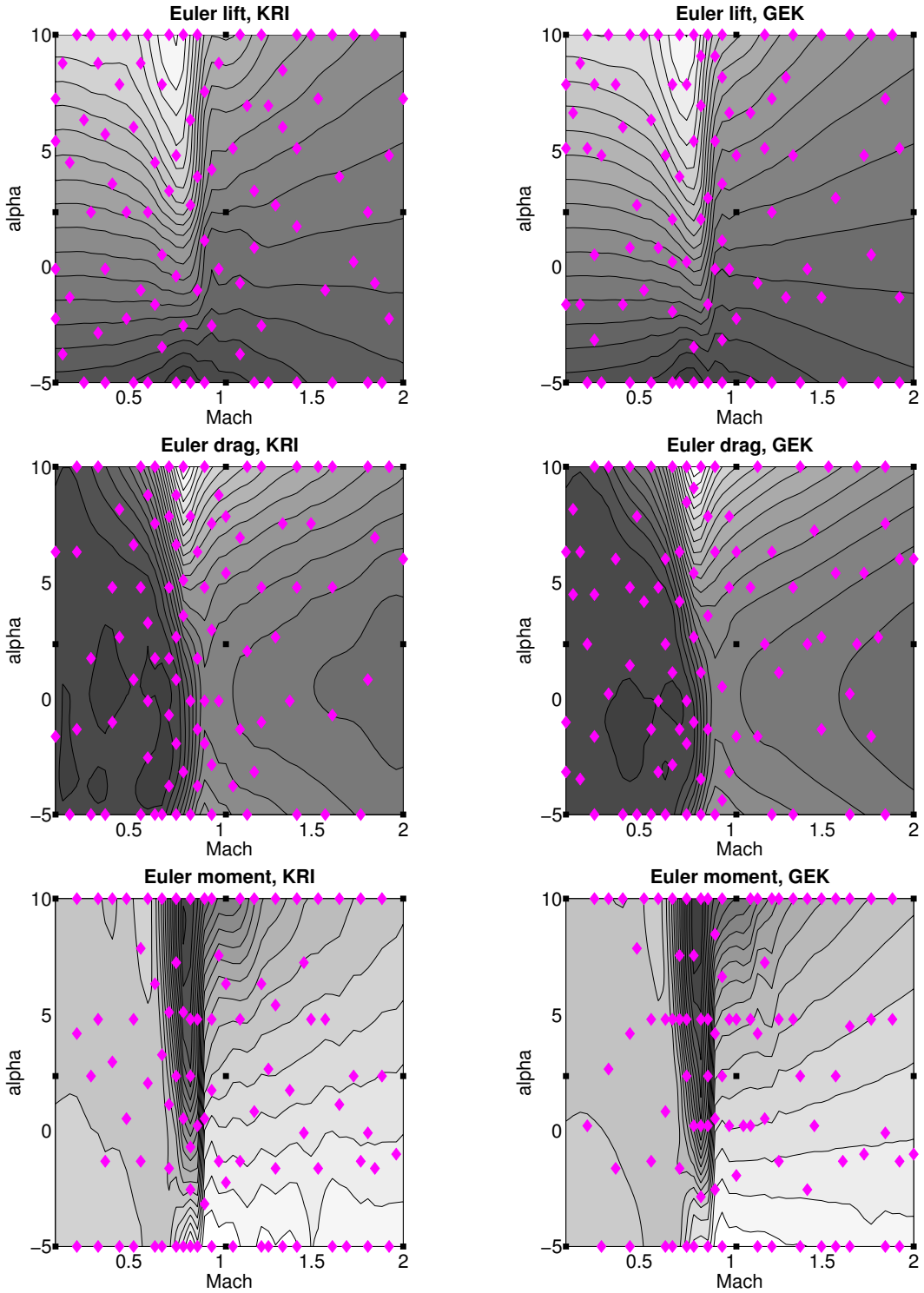

Figure 4.15: Sample distributions for ADG strategy, FLOWer Euler responses.

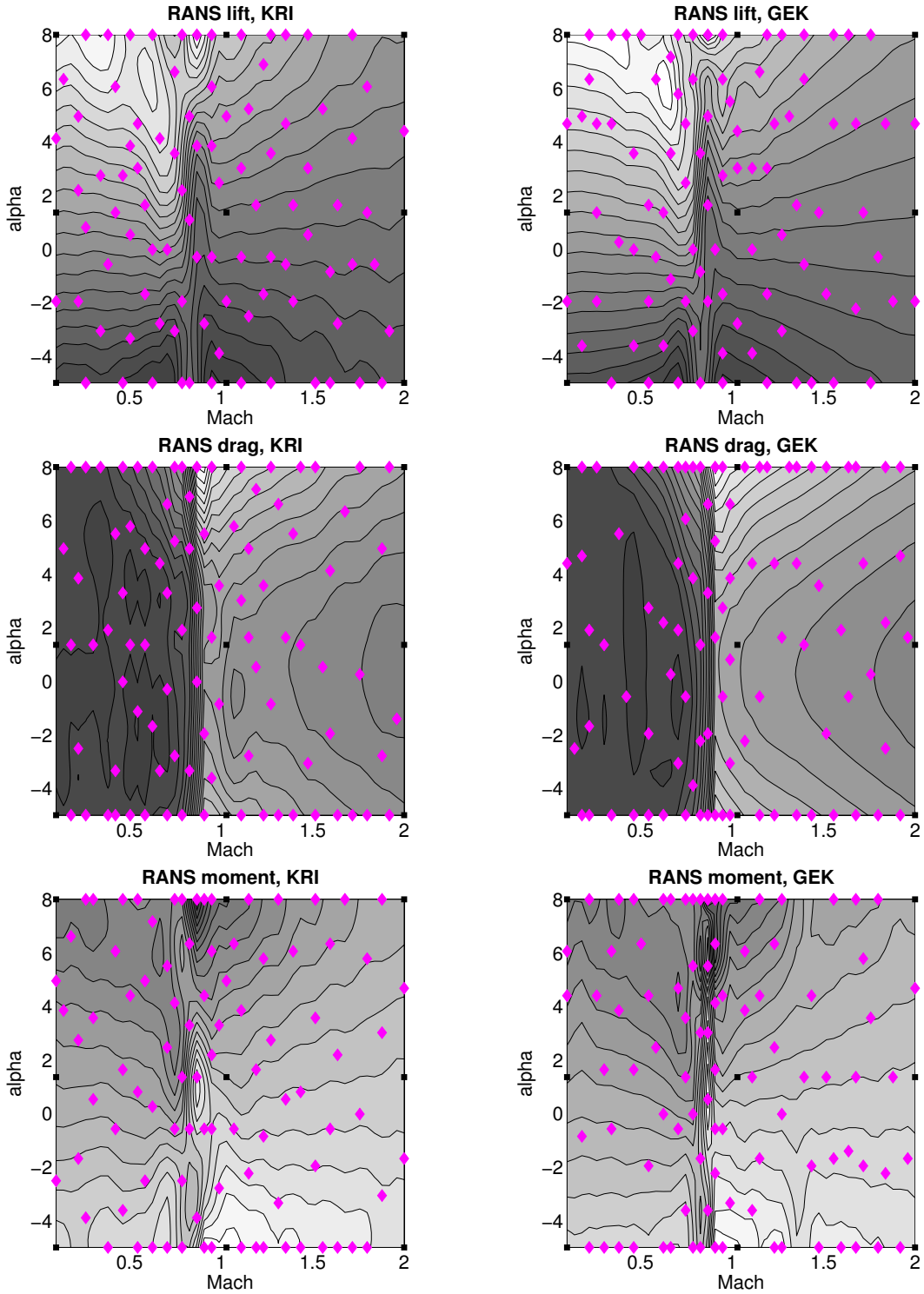

Figure 4.16: Sample distributions for ADG strategy, TAU RANS responses.

4.4 Theoretical sampling strategies

In order to provide an assessment standard for the performance of the adaptive sampling strategies, two new strategies are now introduced which avail themselves of the validation data on Ω^{val}, a 50×50 factorial design on Ω. Since in real life applications only the current data $Y \in \mathbb{R}^N$ on $X^{(N)}$ ($Y^{\partial} \in \mathbb{R}^{(d+1)N}$ for GEK, respectively) is available during the process of model generation, they are only of theoretical interest. In the following, these strategies are referred to as theoretical sampling strategies, while the previous approaches (MSE, SSE and ADG) are referred to as practical sampling strategies. Probably the first approach which comes in mind when sequentially adding samples from a set of validation data is comparing the current surrogate model with the true response on Ω^{val} in each step. A new sample is added where the true error $|\widehat{y}(x) - y(x)|$ is highest,

$$x^{(N+1)} = \arg \max_{x \in \Omega^{\mathrm{val}}} |\widehat{y}(x) - y(x)| \,. \tag{4.6}$$

After adding the evaluation $y(x^{(N+1)})$ (and the gradient $\nabla y(x^{(N+1)})$ in GEK) and generating a new surrogate, the error is set to zero in $x^{(N+1)}$ and also a global improvement, e.g. in terms of the relative root mean squared error η_2 (4.2), is expected. It will be denoted as *true error strategy* (ERR) in the following.

A second approach chooses the new sample $x^{(N+1)}$ from Ω^{val} such that η_2 is actually diminished. Let $\widehat{y}^{(N+1)}(x \,|\, x^*)$ denote a Kriging surrogate model, which is based on $X^{(N)}$ and one additional sample $x^* \in \Omega \setminus X^{(N)}$. It is generated by adding $y(x^*)$ (and $\nabla y(x^*)$ in GEK) to the model and solving the Kriging system anew. Having access to the validation data on Ω^{val} and being able to evaluate an error functional like η_2, the best possible choice for adding one new sample is minimizing this error functional for $\widehat{y}^{(N+1)}(x \,|\, x^*)$,

$$x^{(N+1)} = \arg \min_{x^* \in \Omega^{\mathrm{val}} \setminus X^{(N)}} \left\{ \sum_{x \in \Omega^{\mathrm{val}}} \left(\widehat{y}^{(N+1)}(x \,|\, x^*) - y(x) \right)^2 \right\}. \tag{4.7}$$

With this selection criterion, the new surrogate model $\widehat{y}^{(N+1)}(x)$ based on the updated design $X^{(N+1)} = X^{(N)} \cup \{x^{(N+1)}\}$ achieves the best possible improvement of global accuracy in terms of the root mean squared error η_2, when only considering Ω^{val} as a finite set of candidates. This strategy is referred to as *maximum improvement strategy* (IMPR). It is a theoretical best case scenario for sequential designs when adding one sample per stage. It does however not yield the best possible design $X^{(N_{\max})}$ of size $N_{\max}$ out of the finite set of candidates Ω^{val}. Though theoretically possible by generating surrogate models for all combinations of $N_{\max}$ samples out of Ω^{val} and assessing them by their error values, this is computationally out of reach even for small examples. There exist $\binom{\#\Omega^{\mathrm{val}}}{N_{\max}} = \frac{\#\Omega^{\mathrm{val}}!}{N_{\max}! \cdot (\#\Omega^{\mathrm{val}} - N_{\max})!}$ combinations and only for determining the best design of size $N_{\max} = 10$ out of a 10×10 grid validation design, already over $1.7 \cdot 10^{13}$ surrogates would have to be generated and assessed.

The performances of the theoretical sampling strategies (Figs. 4.17 and 4.18) as well as the sample distributions (Figs. 4.19–4.22) are now discussed in order to draw more profound conclusions on the previous results (Sects. 4.3.1 and 4.3.2). The *true error strategy* is able to

outperform the practical sampling strategies in the Euler case after a small starting phase. Still, the SSE strategy, which was identified as most accurate method in the previous sections, can compete to a great extend. In the RANS case on the contrary, the true error method fails for the most part. With few exceptions, the surrogates generated by the ERR strategy are the least accurate ones.

In order to interpret these heterogeneous performances, the sample distributions are now investigated. In the Euler case (Fig. 4.19), most samples are located in the critical transsonic region as expected. There, the true responses feature large gradients and high curvature and the largest errors appear. Also, many samples are located on the α-boundaries. These features can also be observed for the SSE and the ADG strategies, which indicates that these two mixed exploration and exploitation strategies correctly identify the most critical regions of the input parameter space. In the RANS case (Fig. 4.20), in which the ERR strategy produces inaccurate surrogates, the concentration of samples in the most critical regions is much higher. For instance, a clustering of the samples in the transsonic region for high angle of attack α can be observed for all responses. Other areas of Ω, where the true error is indeed not maximal, but the surrogate is nevertheless inaccurate, stay disregarded by the strategy. Also, the true responses $y(x)$ behave not as smooth as in the Euler case, especially in the transsonic region. Therefore, a lower correlation is detected by the hyperparameter estimation (i.e. larger hyperparameters θ) when many samples are located in the highly nonlinear areas. The consequence of large hyperparameters or, equivalently, low correlation is a globally erratic behavior of the surrogate $\widehat{y}(x)$, also in regions where the true response $y(x)$ is in fact very smooth or almost linear. It can be concluded that for globally accurate surrogate models, both highly concentrated clusters of samples in critical regions as well as leaving vast regions undersampled, even if the response is linear in that area, is not desired. Thus, rather mixed strategies, which have an exploitation and an exploration component, are the method of choice.

The *maximum improvement strategy* generally imposes itself as a lower error boundary for all practical sampling strategies as well as for the theoretical ERR strategy. Apart from one exception (Kriging for c_d in the Euler case), the surrogates based on the IMPR strategy mainly are most accurate. Despite the selection of the new sample as a minimum of the new surrogate's error, η_2 does not decrease in a few cases (e.g. GEK for c_m in the RANS case, $N = 60$). The reason is that for the construction of $\widehat{y}^{(N+1)}\left(x \,|\, x^*\right)$ in (4.7), the hyperparameters θ of $\widehat{y}^{(N)}(x)$ are used, while after adding $x^{(N+1)}$ and constructing $\widehat{y}^{(N+1)}(x)$, they are estimated anew. Usually, the estimated hyperparameters change only slightly after adding a new sample, but in some rare cases, this can be a source of error. After adding few more new samples, the error diminishes again. In some cases, the SSE and the ADG strategy can compete with the lower error boundary, proving that they are indeed powerful adaptive sampling strategies.

Investigating the sample distributions of the IMPR method (Figs. 4.21 and 4.22), it is noticeable that the samples are again concentrated both in the transsonic region and, to some extend, also on the α-boundaries. It can be observed that the samples tend to build clusters, especially in the transsonic region, which are less dense than the ones observed for the ERR strategy. Also, their location seems to be rather random. These clusters can establish a trend in their surroundings (in indirect GEK for instance, adjacent samples are used to substitute gradient information, cf. Sect. 2.4). Instead of placing samples where the

true error of the surrogate $\widehat{y}^{(N)}(x)$ is highest, the information on $\widehat{y}^{(N+1)}(x)$ is used for a global reduction of the error η_2. The samples are thus not necessarily placed where the true response $y(x)$ is most erratic, but in clusters in the neighborhood of these regions, such that the response is correctly reproduced by the trend beyond the clusters' locations. However, any practical sampling strategy can only use the information contained in $\widehat{y}^{(N)}(x)$ and is not able to look one step ahead like the IMPR strategy does.

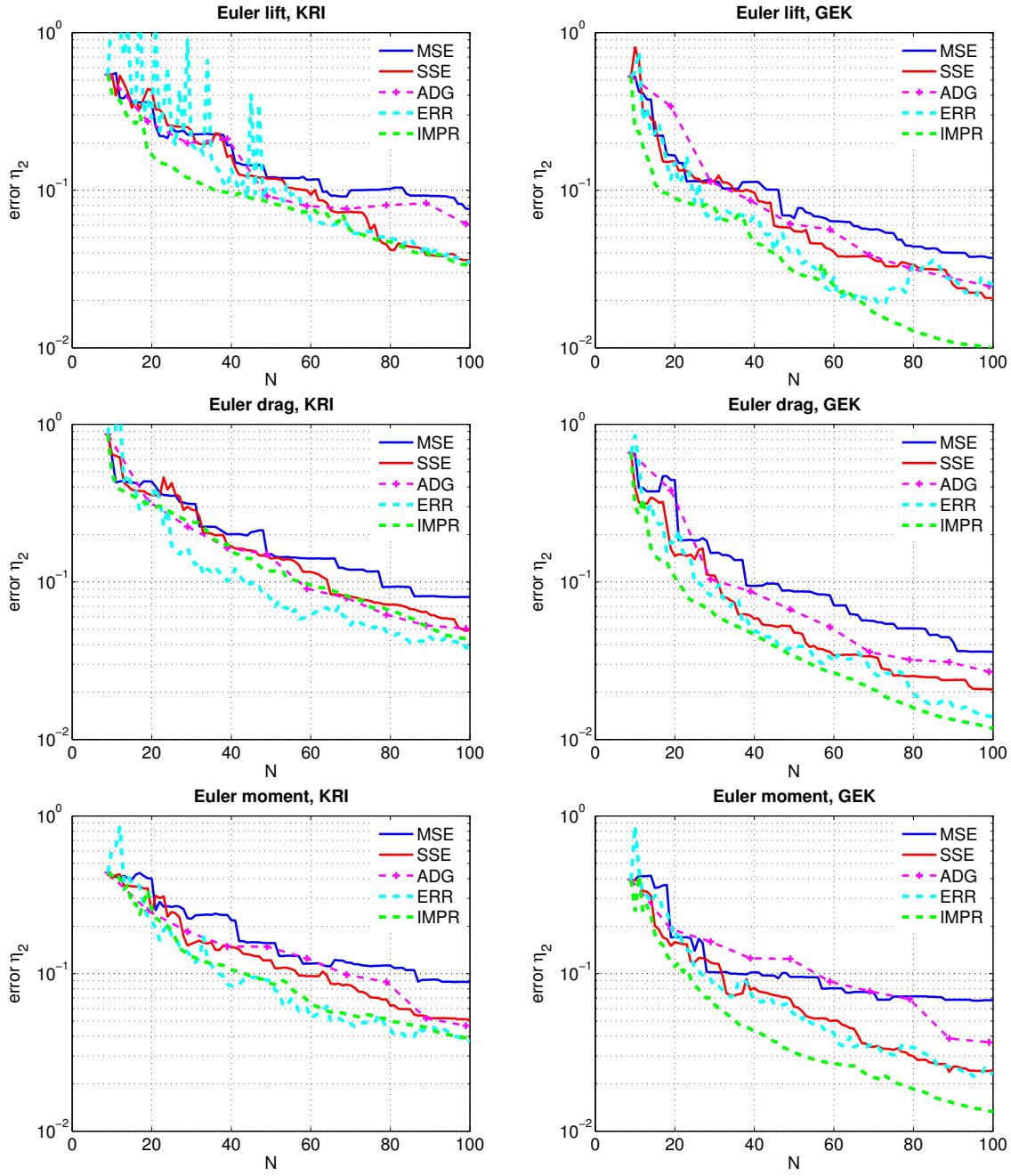

Figure 4.17: Root mean squared error η_2 for interpolation of FLOWer Euler responses.

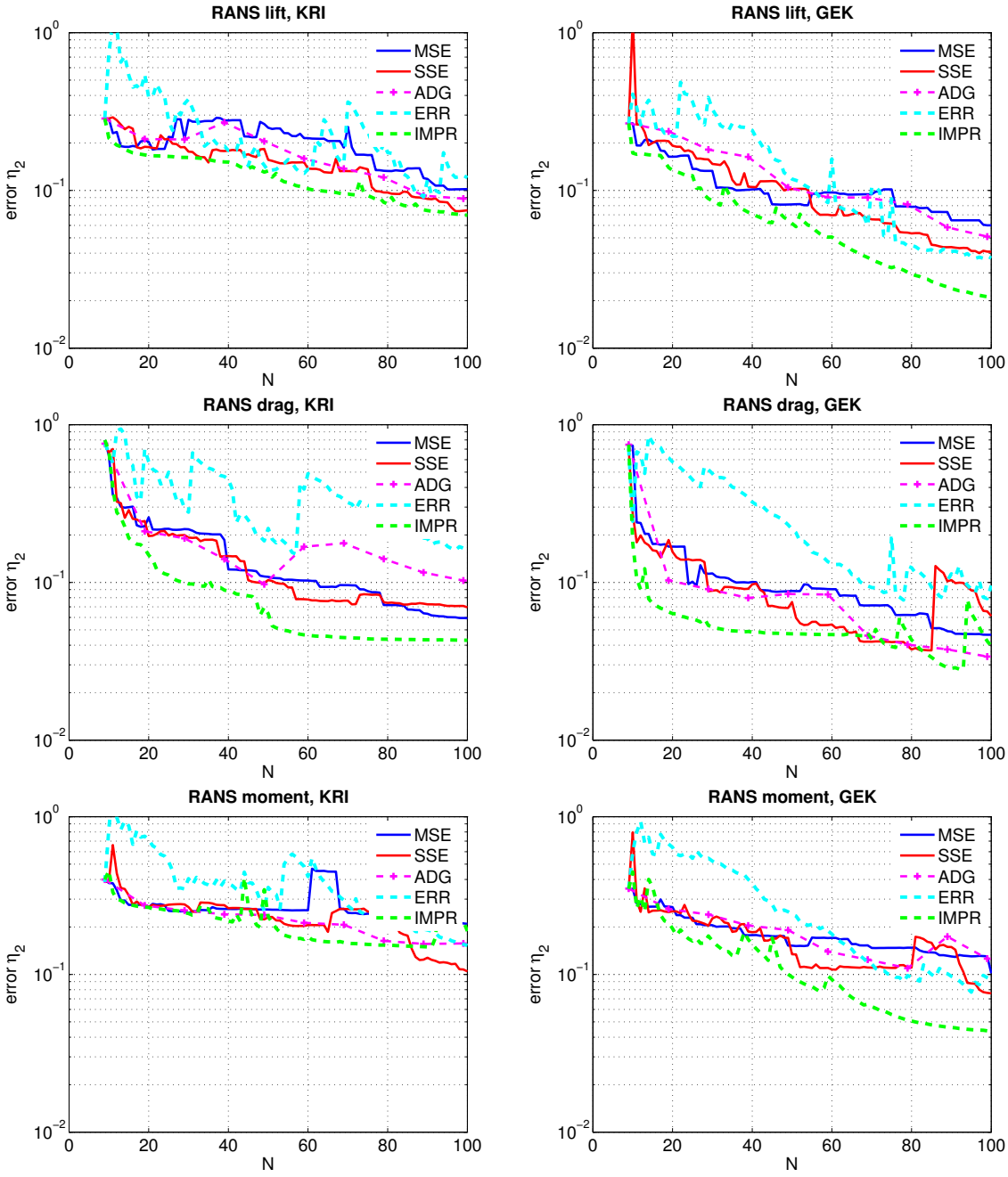

Figure 4.18: Root mean squared error η_2 for interpolation of TAU RANS responses.

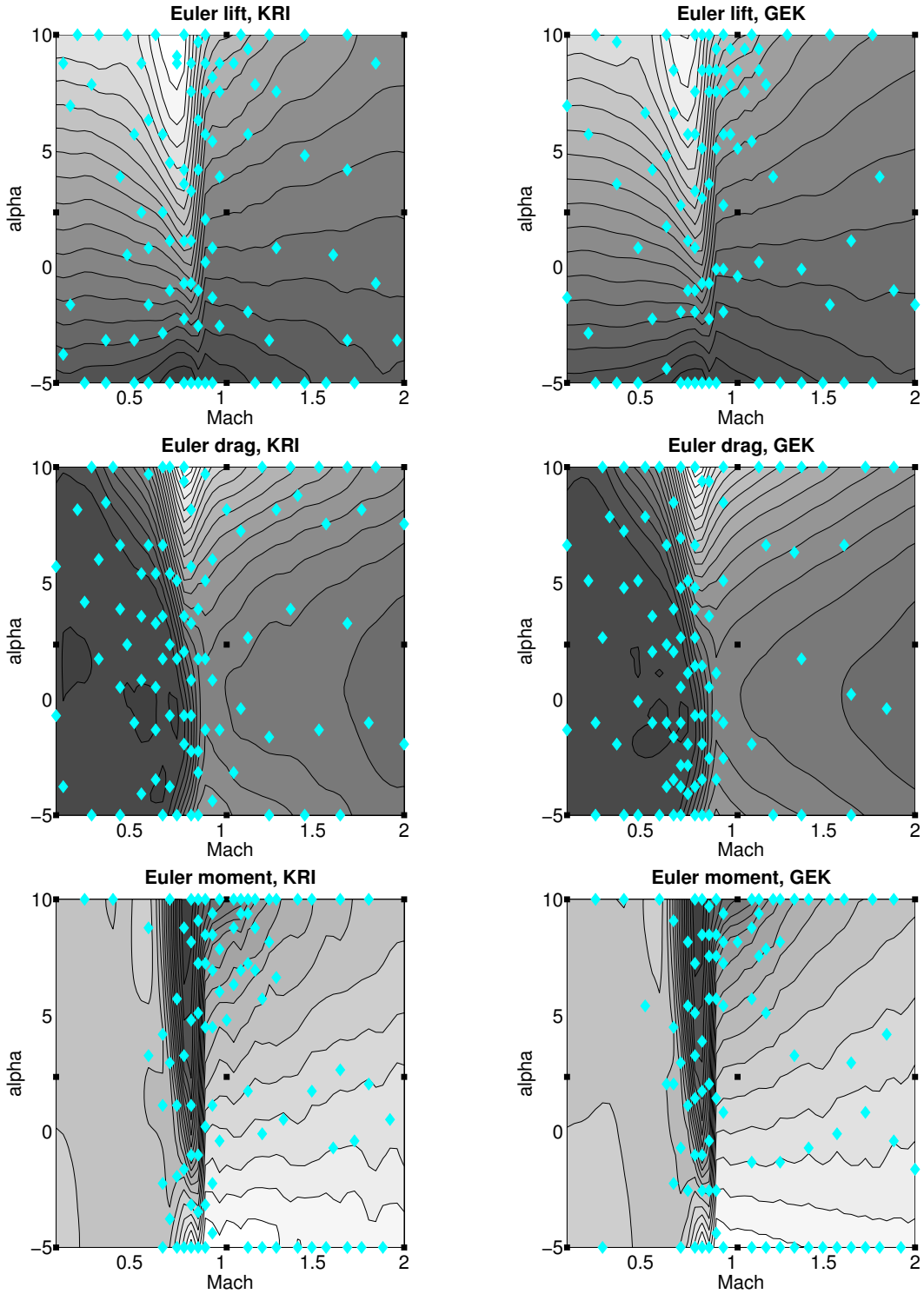

Figure 4.19: Sample distributions for ERR strategy, FLOWer Euler responses.

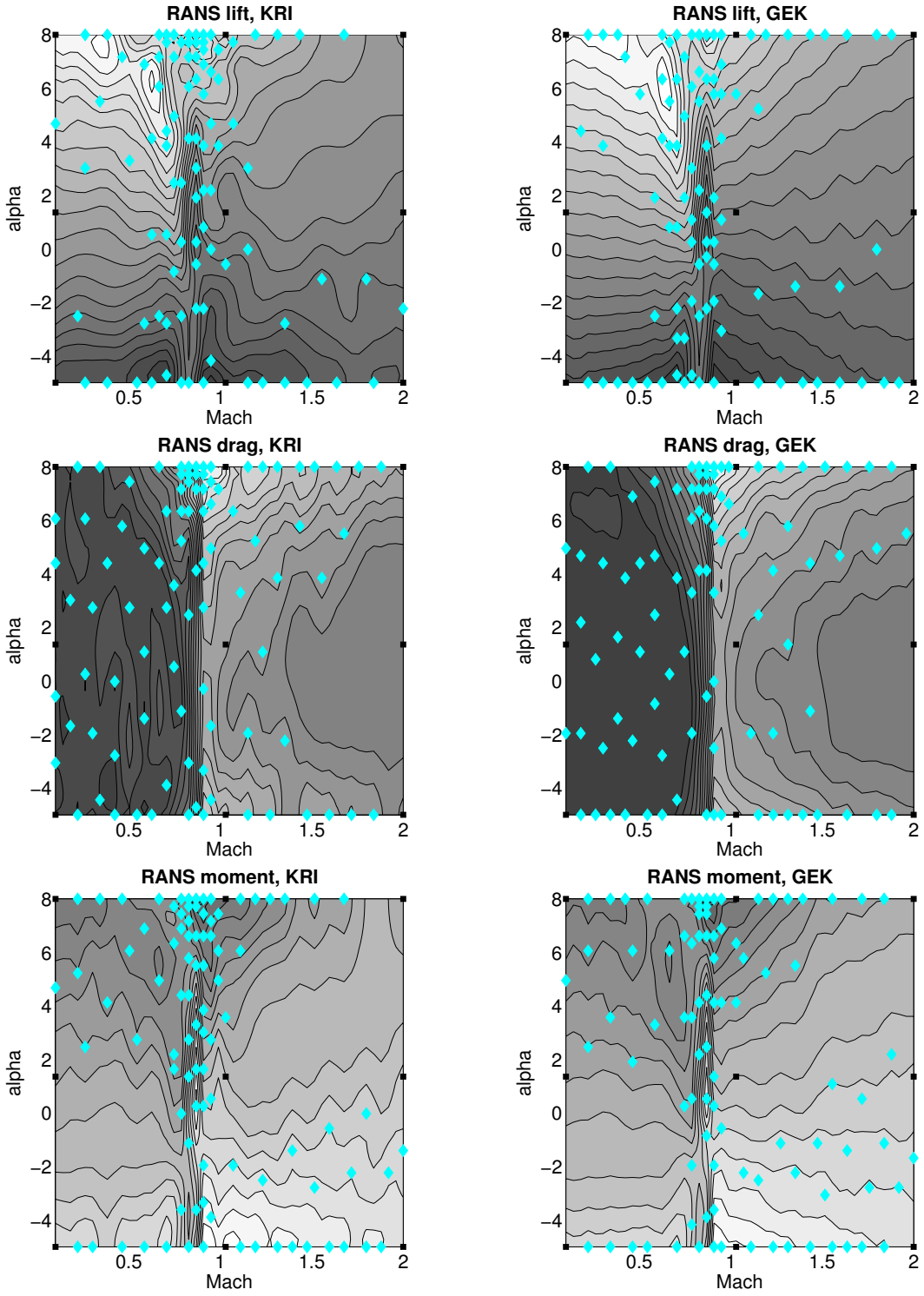

Figure 4.20: Sample distributions for ERR strategy, TAU RANS responses.

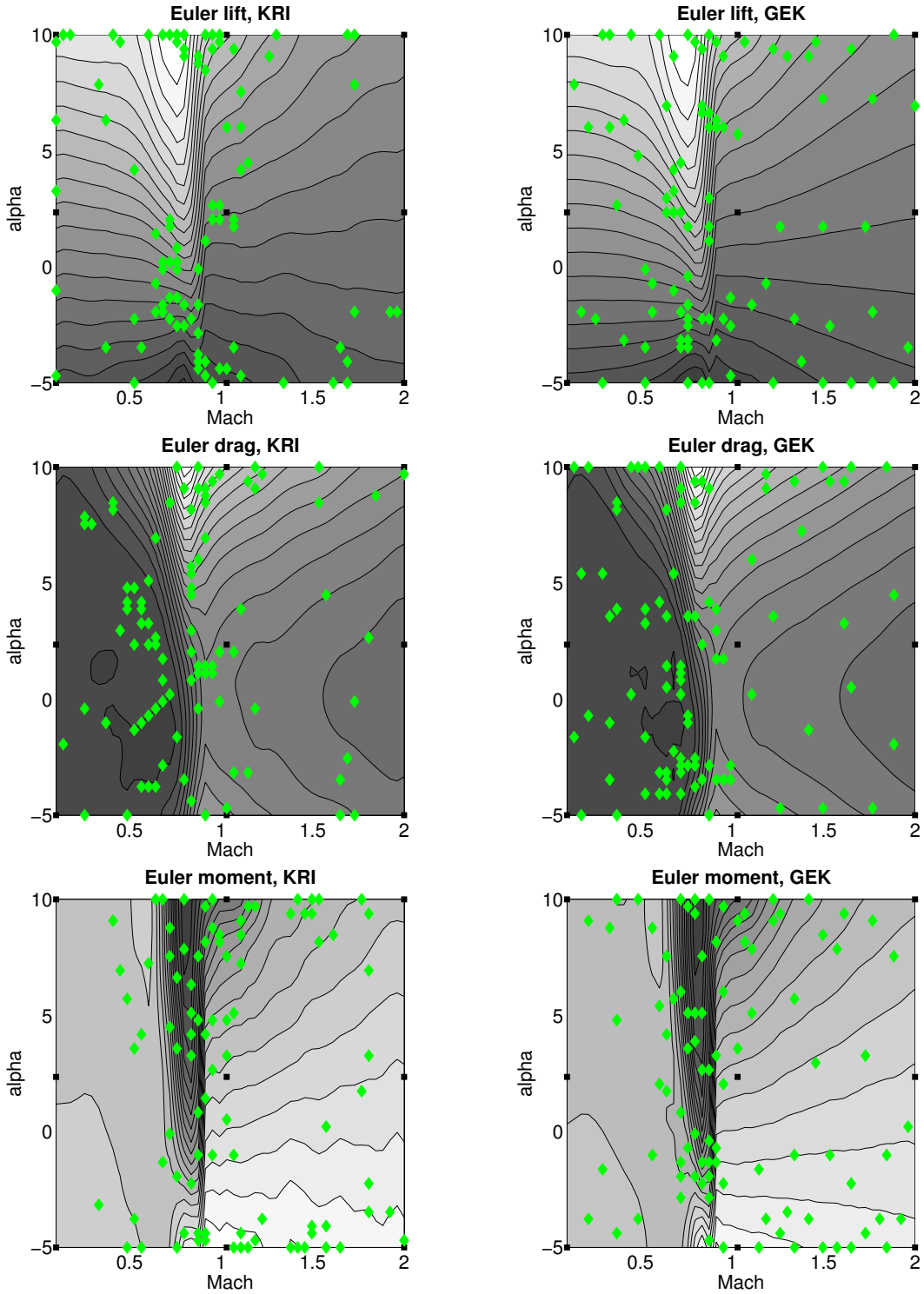

Figure 4.21: Sample distributions for IMPR strategy, FLOWer Euler responses.

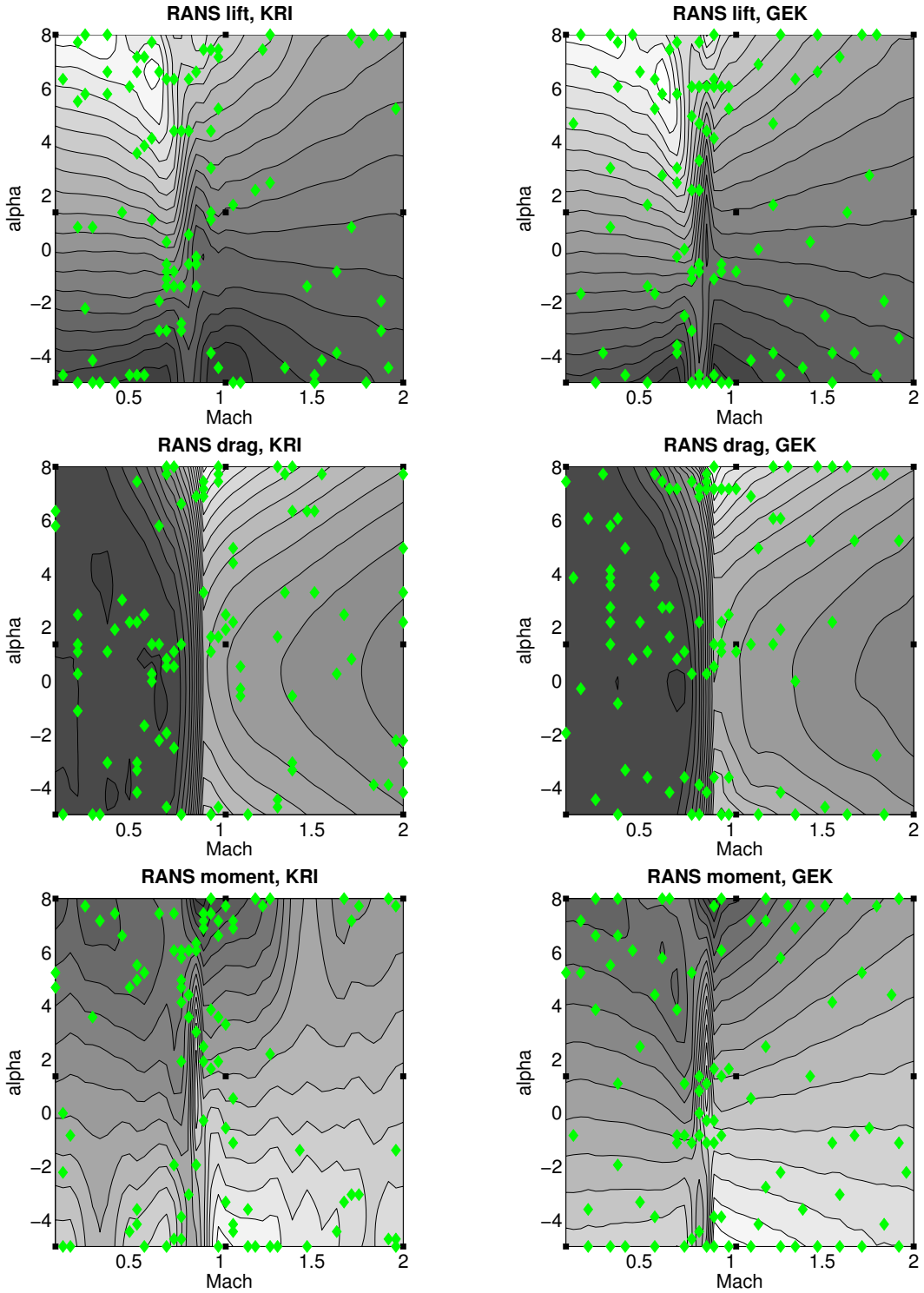

Figure 4.22: Sample distributions for IMPR strategy, TAU RANS responses.

4.5 A strategy for computer experiments with multiple responses

The previously investigated adaptive designs were developed for scalar-valued responses $y : \mathbb{R}^d \to \mathbb{R}$. But often a computer experiment yields a vector-valued response $y : \mathbb{R}^d \to \mathbb{R}^n$ of dimension n. For each component $y_1, \ldots, y_n$ of the response a surrogate $\widehat{y}_1, \ldots, \widehat{y}_n$ is sought. In the aerodynamic test cases for instance, one evaluation of the flow solver yields three aerodynamic coefficients $y(x) = (c_l(x), c_d(x), c_m(x))^\top \in \mathbb{R}^3$, with $x = (\mathrm{Ma}, \alpha)^\top \in \mathbb{R}^2$. Being able to compute all n quantities at once without additional costs, having n different designs for $\widehat{y}_1, \ldots, \widehat{y}_n$ certainly is not efficient. A naive approach would just use a space-filling one-stage approach like a Halton design, which does neither depend on the responses $y_1, \ldots, y_n$ nor on the surrogates $\widehat{y}_1, \ldots, \widehat{y}_n$. The previous sections demonstrated that adaptive sampling strategies (mixed exploration & exploitation) for scalar-valued responses are capable of outperforming one-stage approaches, meaning more accurate surrogates using less samples. In the following, an attempt is made to use such a mixed adaptive sampling strategy for vector-valued responses as well. Little can be found in the literature on the topic of sequential designs for multiple responses problems. In [85] for instance, a modified maximum entropy design, which is input- and output-based by adapting the correlation matrix entries to a predicted error, was extended to vector-valued responses. An application of a combined, MSE-based selection criterion for multiple responses was presented in [49]. In [23], a Voronoi-tesselation of the input parameter domain was combined with a nonlinearity measure and the maximum of these n measures for all responses was used as a selection criterion.

The sample sensitivity error strategy (SSE) produced the overall best surrogates for the scalar-valued test cases and it will now be extended to the vector-valued case. Each response y_k is now assigned a separate surrogate model $\widehat{y}_k$, which for instance means that each surrogate has its own hyperparameters, but they all share one design X ($k = 1, \ldots, n$). Let $e_k(x)$ denote the cross-validation error as in (3.18) for the k-th surrogate and $\mathrm{RMSE}[\widehat{y}_k(x)]$ be the root mean squared error predictor for $\widehat{y}_k$. Having distinct hyperparameters, the RMSE is in fact a distinct function for each $\widehat{y}_k$, otherwise they would only differ in constant factors σ_k, cf. (3.10). For single-output responses, $e(x)\,\mathrm{RMSE}[\widehat{y}(x)]$ is used as a selection criterion for a new sample $x^{(N+1)}$ (3.20). It is nonnegative and equal to zero in each $x^{(i)} \in X^{(N)}$. For multiple-output responses, the geometric mean of $e_k(x)\,\mathrm{RMSE}[\widehat{y}_k(x)]$ ($k = 1, \ldots, n$) can be used as a combined selection criterion

$$x^{(N+1)} = \arg\max_{x \in \Omega} \left\{ \prod_{k=1}^{n} \left(e_k(x)\,\mathrm{RMSE}\left[\widehat{y}_k(x)\right] \right)^{\frac{1}{n}} \right\}. \tag{4.8}$$

The geometric mean has the advantage of being independent of the particular scales of each y_k. The selection criterion represents a tradeoff between the different error indicators $e_k(x) \cdot \mathrm{RMSE}[\widehat{y}_k(x)]$ ($k = 1, \ldots, n$) and chooses a new sample which is possibly not the best choice for each surrogate alone, but a compromise for increasing the accuracy of all $\widehat{y}_k$ collectively. In the investigated test cases, the responses c_l, c_d and c_m share a mutual critical region, the transsonic flow regime, due to the underlying physics of the flow equations. If this is not the case for a vector-valued computer experiment, the geometrical mean might be inappropriate. If $\widehat{y}_k$'s error indicator is close to zero in some part of the input parameter

domain, but $\widehat{y}_l$'s error indicator has its maximum in this part for some $l \neq k$, still no sample might be placed there in the whole adaptive process because the product stays close to zero. Instead, a weighted arithmetic mean of the error indicators

$$x^{(N+1)} = \arg\max_{x \in \Omega} \left\{ \frac{1}{n} \sum_{k=1}^{n} \left(\frac{1}{\sigma_k} e_k(x) \, \mathrm{RMSE}\left[\widehat{y}_k(x)\right] \right) \right\} \tag{4.9}$$

might be more suitable as a selection criterion, where σ_k^2 denote y_k's process variances (2.79).

Strategy (4.8) is applied to the aerodynamic test cases, generating a mutual adaptive design X for the three responses c_l, c_d and c_m. In the Euler case and in the RANS case, Kriging and GEK interpolations are investigated. Figures 4.23 and 4.24 depict the error η_2 (4.2) for the *multiple responses strategy* against the performance of the SSE strategy applied to each response individually (same data as in Sect. 4.3). Both strategies perform comparable for the most part. Often, the combined strategy yields even more accurate surrogates than the individual SSE strategies. Figs. 4.25 and 4.26 show the shared designs in each column. Comparing the distributions of the samples, it can be observed that they basically have the same features as the individually determined designs of Sect. 4.3. Many samples are located on the boundaries of the α-axis, while in the interior of Ω, most samples are concentrated in the transsonic region without totally neglecting the sub- and supersonic region. The proposed mutual sequential sampling strategy (4.8) proves to work well in the aerodynamic test cases. It can compete with the designs adapted to each single response individually and therefore also outperforms the naive approach for vector-valued responses of a nonadaptive one-stage Halton design.

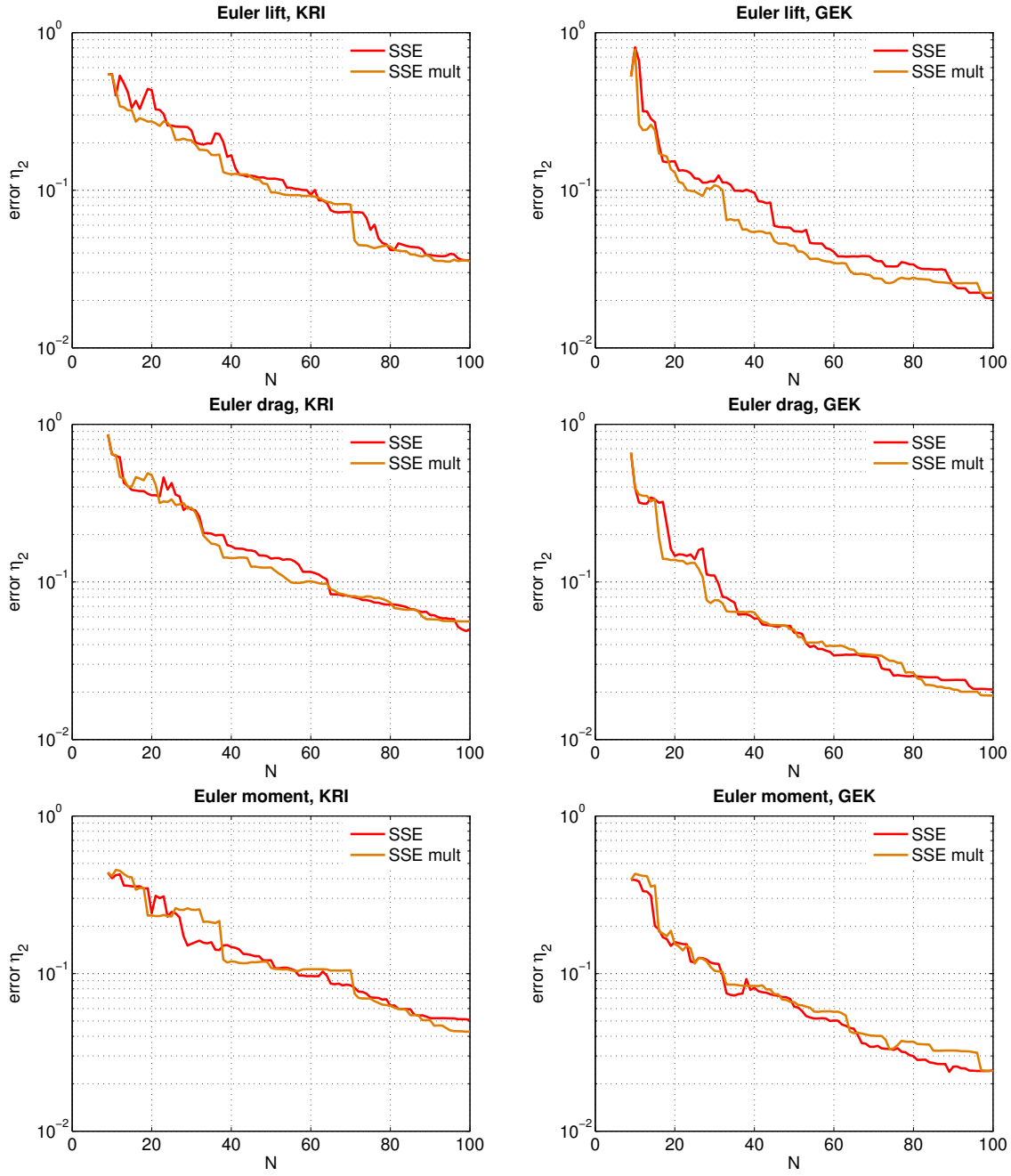

Figure 4.23: Root mean squared error η_2 for interpolation of FLOWer Euler responses.

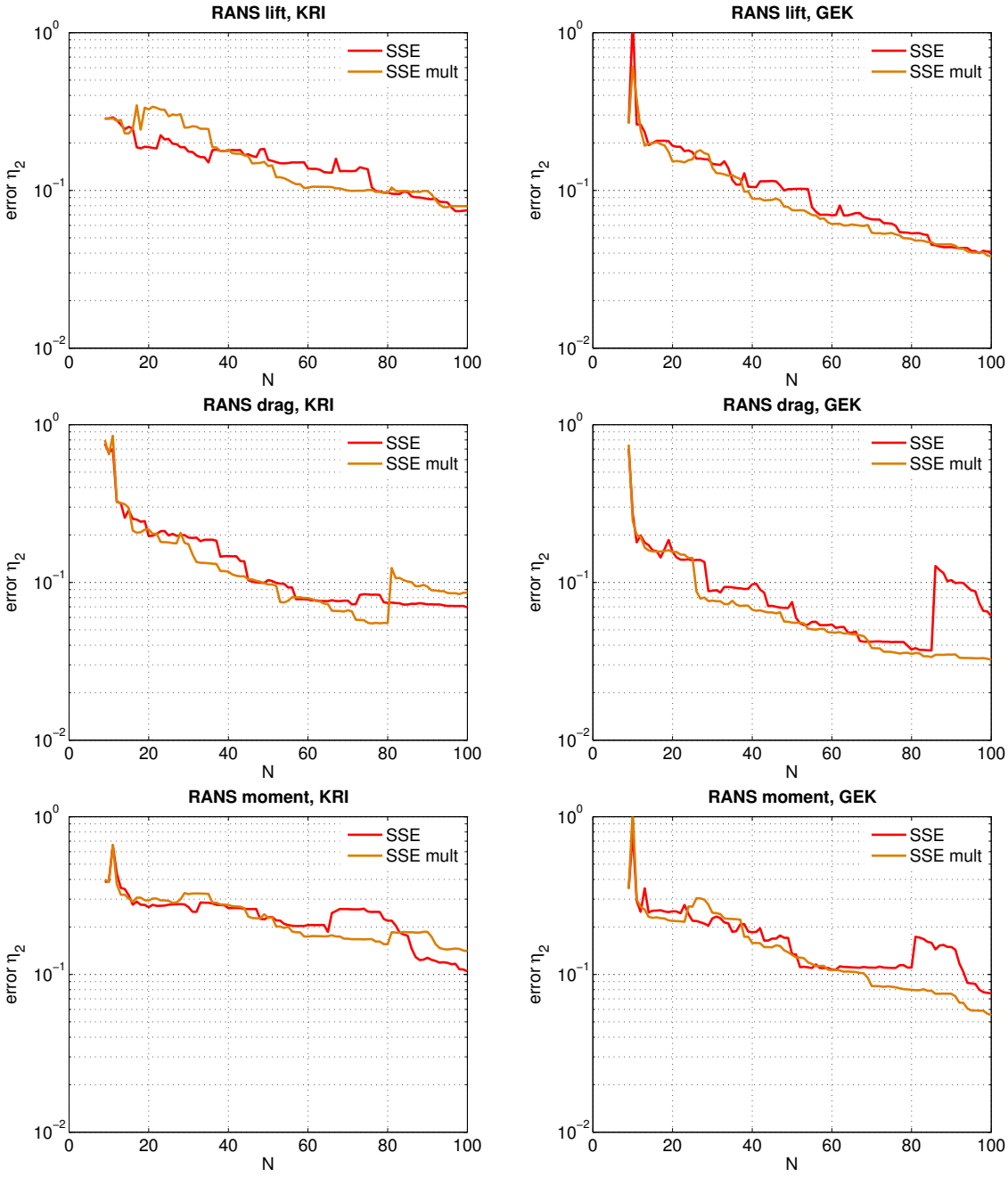

Figure 4.24: Root mean squared error η_2 for interpolation of TAU RANS responses.

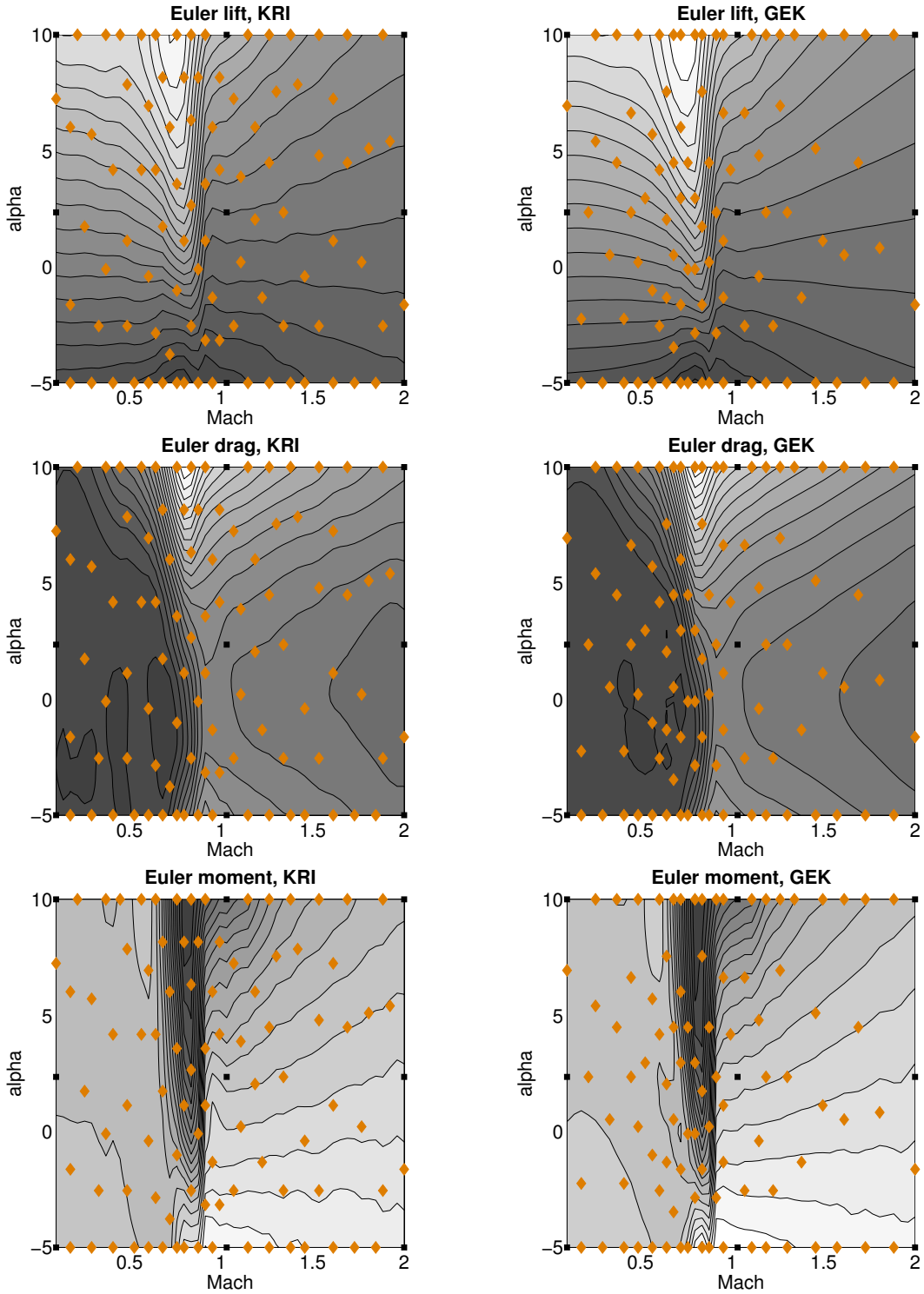

Figure 4.25: Sample distributions for the multiple responses strategy, FLOWer Euler.

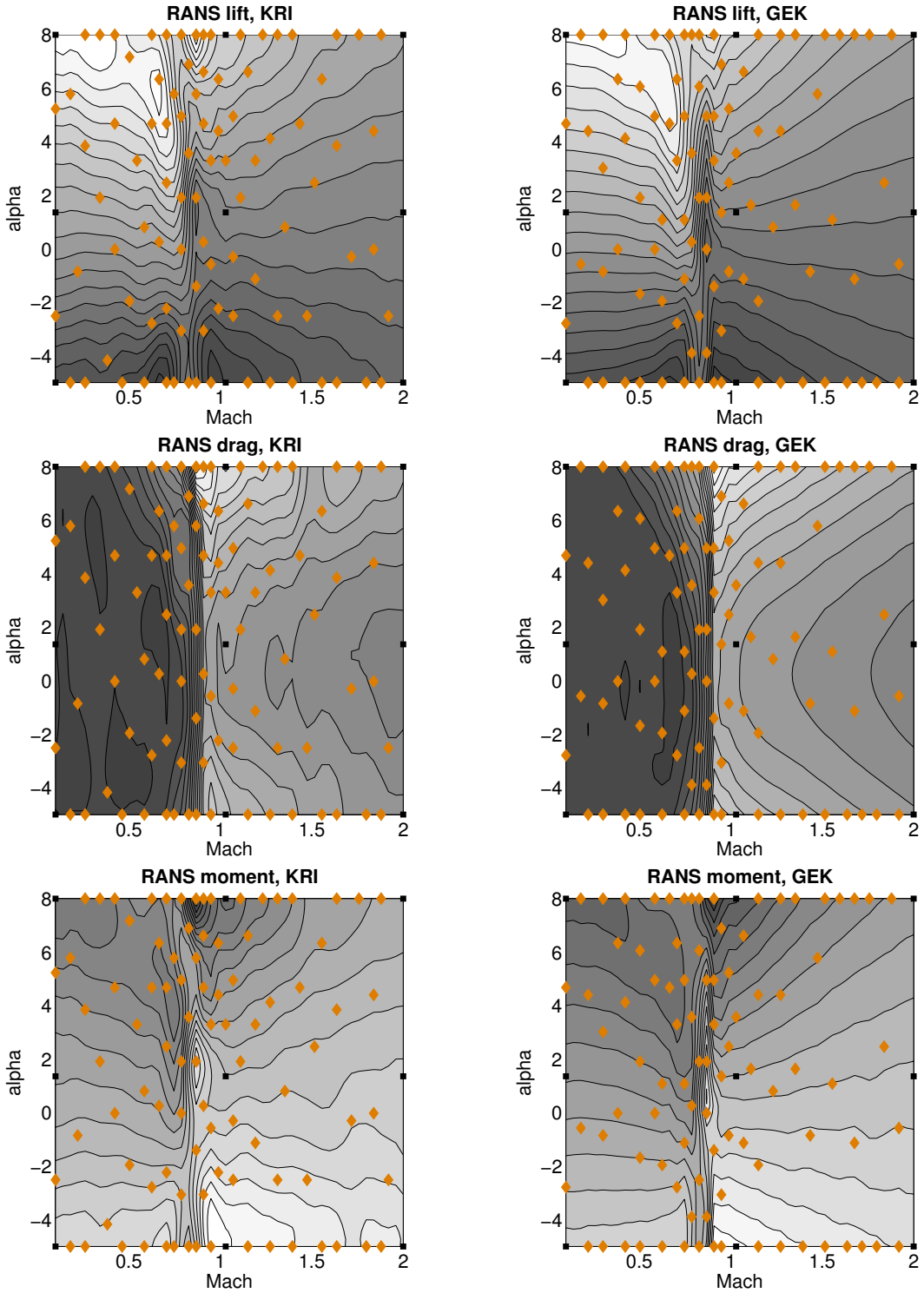

Figure 4.26: Sample distributions for the multiple responses strategy, TAU RANS.

4.6 Summary

Designs for efficient and globally accurate Kriging and gradient-enhanced Kriging surrogate models were validated in this chapter. Aerodynamic responses based on Euler and RANS computations were identified as test cases of sufficient complexity. Space-filling one-stage approaches were investigated as well as adaptive sampling strategies.

Comparing one-stage approaches, designs based on a low-discrepancy Halton sequence could clearly outperform factorial designs and optimized Latin hypercube designs regarding global accuracy. This result was also observed in many test cases in the existing literature [68, 131, 81, 60]. The Halton sequence's accuracy then served as a benchmark for adaptive designs.

As adaptive designs, especially mixed strategies containing an exploration (space-filling) and an exploitation (error-based) component were focussed. Two mixed exploration & exploitation sampling strategies (SSE and ADG) were generally able to produce Kriging and GEK surrogates of higher accuracy than the one-stage designs and a pure exploration strategy (MSE). They provided a more efficient usage of the expensive evaluations of the response and reduced the number of required samples. A comparison of the distributions of the samples in the input parameter domain backed up the results. The mean squared error prediction was identified as a useful exploration criterion, while cross-validation methods served as a viable exploitation component. Critical regions in the input parameter domain were identified automatically and were provided with a higher concentration of samples, while other regions were sampled coarser but were not neglected either. Other comparative studies came to similar conclusions. In [46, 47], combined criteria were found to be a more reliable indicator for the source of approximation error than single. In [23, 25], hybrid exploration & exploitation methods generally outperformed pure exploration or exploitation methods regarding global accuracy.

For gaining a deeper insight into optimal designs and for a more profound assessment of the approximation error, also two theoretical sampling strategies were investigated which make use of a dense set of validation data. They are therefore merely of academic interest. The first one was a pure exploitation method, adding a sample where the true error is highest in every stage. In the Euler case, it showed overall good results, but the hybrid SSE strategy was able to keep up with its performance. In the RANS case however, the method produced inaccurate approximations. The concentration of the samples in the input parameter domain became too unbalanced, demonstrating that pure exploitation strategies are generally not recommended for global approximation. A second strategy chose the next sample by minimizing the approximation error of the next stage's surrogate. In this best-case study, an experimental lower error boundary could be established. Both mixed strategies (SSE and ADG) could compete at least in some cases, proving that they are indeed powerful sampling strategies.

Regarding efficiency, the SSE strategy was also enhanced for the use in test cases with multiple responses. Distinct designs for each response are barely desired, if one evaluation of the computer experiment yields all responses simultaneously. A mixed selection criterion was proposed, which represents a balanced choice of the error indicators for the different responses. It generally produced results similar to the SSE strategy for single-output responses, both regarding global accuracy and sample distributions, establishing the method

as an efficient and powerful adaptive sampling strategy.

5 Generic surrogate modeling

In this chapter, a new surrogate modeling framework is introduced. Its purpose is to generate a globally accurate surrogate with low computational effort, which can be accomplished by using only few expensive function evaluations of $y(x)$. Adaptive sampling strategies have been investigated in the last two chapters. They reduce the number N of required samples for a targeted accuracy by an intelligent distribution of the samples $X = \{x^{(1)}, \dots, x^{(N)}\}$ in the input parameter domain Ω. However, the response $y(x)$ is treated as a black-box and common surrogate modeling techniques can only use the information contained in the data pairs $\{(x^{(i)}, y(x^{(i)}))\}_{i=1}^{N}$. The new approach is motivated by the assumption that for a predefined problem class, the behavior of the response $y(x)$ is not arbitrary, but rather related to other instances $\phi_1(x), \phi_2(x), \phi_3(x), \dots$ of a mutual problem class. For example in CFD, responses of aerodynamic coefficients, depending on the input parameters Mach number and angle of attack, share structural similarities for different airfoil geometries. To identify these structures, the concept of *statistical shape models* is applied, which uses a principal component analysis to quantify modes of variation of a previously computed training database $\{\phi_1, \dots, \phi_M\}$. If functions of a problem class are accessible in form of such a database, a new test case y of this function family can be approximated using only few costly evaluations. Instead of directly interpolating the samples, first the principal components are fitted to the data which then act as a *generic surrogate model* (GSM). Based on this model, interpolation of the samples is performed with a *variable-fidelity modeling* method (VFM), which combines the generic surrogate model with the actual evaluations in a data fusion framework, cf. Fig. 5.1.

Concerning the computational effort, one has to differentiate between *online costs* and *offline costs*. In Fig. 5.1, all steps on the left side are considered offline costs. They are dominated by the numerical effort of generating a training database, whose elements are globally accurate surrogate models themselves. This task can be seen as a general preprocessing, meaning after the database has been generated and stored once, it can be used for every new test case of the mutual problem class. The new approach transfers online costs to offline costs by using the information contained in the training database for reproducing the behavior of $y(x)$ without the excessive use of expensive (online) response evaluations. The global, generally highly nonlinear trend of the unknown response $y(x)$ is captured by the generic surrogate model. Still, the new surrogate modeling technique is flexible enough to interpolate possible deviations of the evaluations $y(x^{(i)})$ from the generic model.

This new concept combines and enhances different techniques emerging from a broad field of disciplines, which are now presented with an overview on the existing literature. Using secondary data for improving the approximation quality of a primary variable is state-of-the-art in surrogate modeling for expensive simulations, both in global approximation and in optimization. For instance, only few highly accurate but expensive (primary) evaluations are augmented by many less accurate but cheaper (secondary) computations, which despite

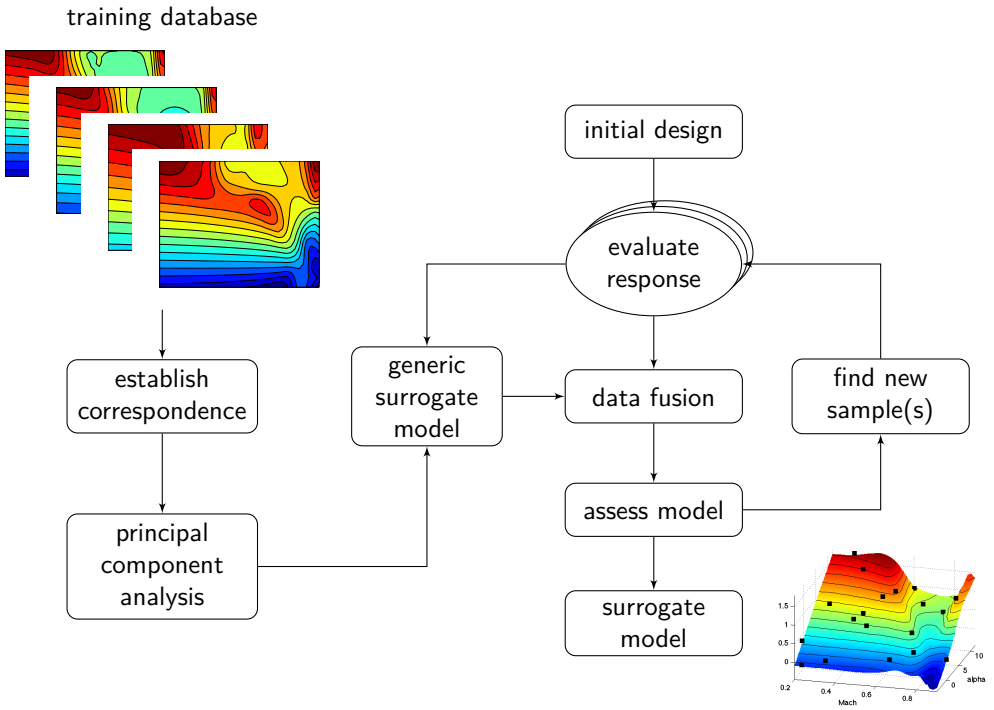

Figure 5.1: Generic surrogate modeling framework.

their inaccuracy are still able to reproduce the global trend of the primary response. *Cokriging* originates from geostatistics [22, 146] and enables the incorporation of any auxiliary variable, which is correlated with the primary one. For instance, gradient-enhanced Kriging (see Sect. 2.4) can be interpreted as a special case of Cokriging, where derivative information serves as secondary data. In the last decade, Cokriging was also used for computer experiments [69, 40, 39], see also [56, 55] for recent applications in aerodynamics. Other *variable-fidelity modeling* methods (VFM) use bridge functions to correct the discrepancy between data of low and high fidelity and date back almost as long as surrogate modeling for computer experiments itself, see e.g. [12]. Nowadays they are also used in aerodynamics [137, 116, 52, 82]. Recently, a new VFM method was developed [53], which overcomes the difficulties of model building and robustness in Cokriging as well as the problems of accuracy and missing mean squared error prediction in the correction-based VFM methods. While statistical shape models [19, 29, 27] are popular in the fields of computer vision and (medical) image processing [43, 78, 128], to the author's knowledge the generic surrogate modeling approach represents the first attempt to use this technique in response surface methods for global approximation of expensive computer experiments. Statistical shape analysis mainly consists of a *proper orthogonal decomposition* (POD) of the training database, which is also called principal component analysis (PCA) or Karhunen-Loève transformation, depending on the context, and is used in various fields of research. POD was initially used for the identification of coherent structures in turbulent flows, cf. [5, 133]. It has become a standard method in reduced-order modeling (ROM) for dynamical systems [76, 145, 77, 120, 153].

Another application of POD is the gappy POD method, which can be used for the reconstruction of incomplete data. First introduced in [32] for partial image data, it is nowadays also applied in CFD [10, 141]. In [51], gappy POD was compared to Kriging interpolation for the reconstruction of incomplete flow fields. The generic surrogate modeling framework was introduced by this thesis' author in [118]. This chapter contains a detailed presentation of the method including illustrative examples, while the subsequent one is devoted to a thorough numerical investigation involving new test cases.

The outline of this chapter is as follows. A database of surrogates for a predefined problem class from aerodynamic simulation is introduced in Sect. 5.1 and the problem of establishing a correspondence between the database elements is discussed. The concept of statistical shape models as well as the POD method for L^2-functions are presented in Sect. 5.2. In Sect. 5.3, the gappy POD method is described and extended to the continuous case. It is also explained how the gappy POD fit of the POD basis elements to some sample data establishes a generic surrogate model. In Sect. 5.4, it is discussed how the generic surrogate model approximation and the sample data are combined by a variable-fidelity modeling interpolation.

5.1 Training database

5.1.1 Plain database

The test cases are now concretized for better understanding without loss of generality. The goal is to generate surrogate models for scalar aerodynamic coefficients lift (c_l), drag (c_d) and pitching moment (c_m) for two-dimensional airfoil geometries, which depend on the input parameters Mach number (Ma) and angle of attack (α), cf. Sect. 4.1. For every input parameter configuration, the response can be evaluated by running a simulation with a (computationally intensive) CFD solver. The responses of different airfoils are assumed to constitute a mutual problem class and to share structural similarities, as illustrated in Figs. 6.3–6.5. When already having generated surrogate models for several airfoil geometries, theses similarities can be used to improve the quality of a surrogate model for a new instance of this problem class.

A database of surrogate models $\phi_1(x), \ldots, \phi_M(x)$ for the same problem class is considered and the superscripts are omitted for simplicity ($\phi_i := \widehat{\phi}_i$). In these test cases, each $\phi_i \in L^2(\Omega)$ is the response of an aerodynamic coefficient (e.g. lift) depending on the input parameters $x = (\mathrm{Ma}, \alpha)$ for a particular airfoil geometry. So the database consists of previously computed response surface functions corresponding to M different airfoils. Furthermore, each response function is assumed to have a sufficient accuracy, i.e. it represents a globally valid surrogate model in the domain of interest $\Omega \subset \mathbb{R}^2$.

5.1.2 Aligned database and correspondence problem

In the fields of computer vision and image recognition, statistical shape models built from a dataset of examples have been widely used, cf. [27]. After establishing correspondence between the database functions by admissible transformations, also called alignment, a principal component analysis is performed to identify the most important modes of variation.

Often, Euclidean or similarity transformations are used to establish correspondence between the images, see e.g. [19]. For an overview on image registration techniques and possible transformations see [38, 135, 154].

As an *admissible transformation* the following one is defined, which is an affine transformation in each dimension of $\Omega \subset \mathbb{R}^2$ and in the image space $\phi_j(\Omega) \subset \mathbb{R}$:

$$\overline{x}(q^j) := \begin{pmatrix} x_1(1+q_1^j)+q_2^j \\ x_2(1+q_3^j)+q_4^j \end{pmatrix}, \tag{5.1}$$

$$\overline{\phi}_j\left(\overline{x}(q^j), q^j\right) := \phi_j\left(\overline{x}(q^j)\right)(1+q_5^j)+q_6^j, \tag{5.2}$$

parametrized by $q^j \in \mathbb{R}^6$ $(j = 1, \ldots, M)$. The transformation is chosen such that, for $q^j = 0$, $\overline{\phi}_j\left(\overline{x}(0), 0\right) = \phi_j(x)$ holds. Note that this transformation was found to be suitable for the investigated test cases, other problem classes could require other admissible transformations. One function $\phi_1(x)$ is defined as a reference, meaning no transformation is applied $(q^1 := 0)$. For the other transformation parameters $q^2, \ldots, q^M \in \mathbb{R}^6$ an optimization problem with $6(M-1)$ unknowns has to be solved:

$$\min_{q^2,\ldots,q^M} \left\{ \frac{1}{M(M-1)} \sum_{j=1}^{M} \sum_{k>j} \int_\Omega \left(\overline{\phi}_j(\overline{x}(q^j), q^j) - \overline{\phi}_k(\overline{x}(q^k), q^k)\right)^2 dx + \frac{\delta}{2} \sum_{j=2}^{M} {q^j}^\top q^j \right\}. \tag{5.3}$$

The solution minimizes the overall sum of squared differences, while for robustness a regularization term is included, which guarantees that the transformation does not become too large. The regularization parameter $\delta > 0$ can be chosen in relation to the initial overall sum of squared differences for $q^j = 0$ $(j = 1, \ldots, M)$, e.g.

$$\delta := 10^{-1} \cdot \frac{1}{M(M-1)} \sum_{j=1}^{M} \sum_{k>j} \int_\Omega \left(\overline{\phi}_j(\overline{x}(0), 0) - \overline{\phi}_k(\overline{x}(0), 0)\right)^2 dx. \tag{5.4}$$

In this way the regularization is independent of the problem-specific level of the correspondence error.

In Fig. 5.2, the alignment is illustrated with an example of two lift responses (cf. Sect. 6.1). Structural similarities like the linear domain in the left half of Ω, the ridge on the right or the local minimum on the lower right are aligned by the transformation, while structural discrepancies like the shape of the local maximum on the upper left are preserved. The nonlinear least squares problem (5.3) is solved by a Gauß-Newton algorithm [103]. Some computational issues will now be addressed briefly. The integral in (5.3) is approximated by a numerical quadrature

$$\int_\Omega f(x)dx \approx \sum_{i=1}^{n} w_i f(x^{(i)}), \tag{5.5}$$

e.g. with the $x^{(i)}$ as elements of a $\sqrt{n} \times \sqrt{n}$-tensorgrid and positive weights w_i. With the

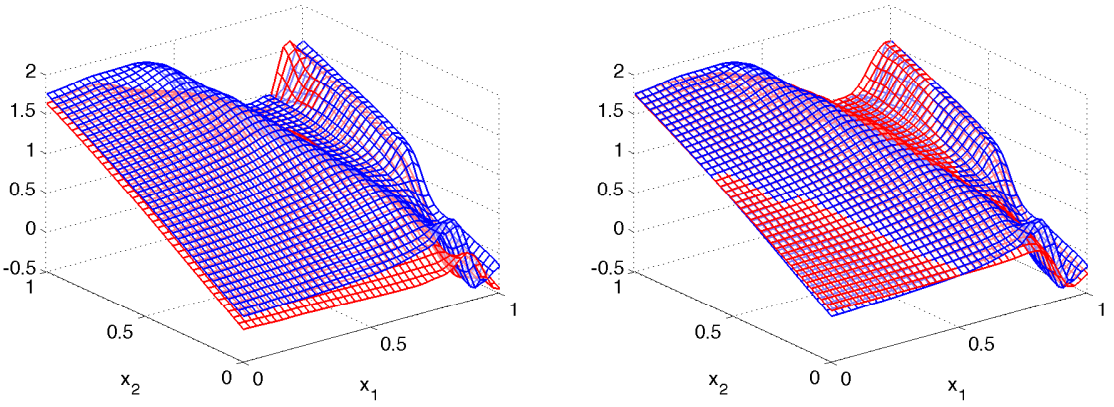

Figure 5.2: Two lift responses ϕ_1 (reference response, blue) and ϕ_2 (red). Left: before alignment, right: after alignment.

term

$$e_{i,j,k} := \sqrt{w_i}\left(\overline{\phi}_j(\overline{x}(x^{(i)}, q^j), q^j) - \overline{\phi}_k(\overline{x}(x^{(i)}, q^k), q^k)\right), \tag{5.6}$$

$$e \in \mathbb{R}^{n\frac{M(M-1)}{2}}, \tag{5.7}$$

the first summand of (5.3) can be written in discretized form as a sum of squared differences

$$\mathrm{SSD}(q) = \frac{1}{2}\sum_{k=1}^{M}\sum_{j>k}\sum_{i=1}^{n} e_{i,j,k}^2. \tag{5.8}$$

Then using

$$J := \frac{\partial e(q)}{\partial q} \in \mathbb{R}^{n\frac{M(M-1)}{2}\times 6(M-1)}, \tag{5.9}$$

the Gauß-Newton algorithm requires the gradient and an approximation to the Hessian

$$\nabla_q \mathrm{SSD}(q) = J^\top e \in \mathbb{R}^{6(M-1)} \tag{5.10}$$

$$\nabla_q^2 \mathrm{SSD}(q) \approx J^\top J \in \mathbb{R}^{6(M-1)\times 6(M-1)}. \tag{5.11}$$

Each entry of the symmetric matrix $J^\top J$ is a dot product of length $n\frac{M(M-1)}{2}$ and there are $\frac{6(M-1)(6(M-1)+1)}{2}$ entries to compute, so the algorithm can really benefit from parallelization. Furthermore, the memory usage by storing the matrix J is $\mathcal{O}(nM^3)$. Derivation of $e_{i,j,k}(q)$ in (5.9) includes application of the chain rule

$$\frac{d}{dq^j}\overline{\phi}_j(\overline{x}(x^{(i)}, q^j), q^j) = \frac{\partial}{\partial \overline{x}}\overline{\phi}_j(\overline{x}(x^{(i)}, q^j), q^j) \cdot \frac{\partial}{\partial q^j}\overline{x}(x^{(i)}, q^j) + \frac{\partial}{\partial q^j}\overline{\phi}_j(\overline{x}(x^{(i)}, q^j), q^j) \tag{5.12}$$

and first-order derivatives $\frac{\partial}{\partial \overline{x}_l}\overline{\phi}_j(\overline{x}(x, q^j), q^j)$ of the database responses appear. Therefore, the database functions have to satisfy $\phi_j \in C^1(\Omega)$. They are surrogate models themselves,

so this can be guaranteed by an appropriate choice of their basis- or correlation functions, cf. Sect. 2.3.3.

5.2 Proper orthogonal decomposition

Let $L^2(\Omega, \mathbb{R})$ denote the Hilbert space of square-integrable, real-valued functions defined on a bounded $\Omega \subset \mathbb{R}^d$. The inner product $\langle f, g \rangle_{L^2(\Omega)} = \int_\Omega f(x)g(x)dx$ and the norm $\|f\|^2_{L^2(\Omega)} = \int_\Omega f(x)^2 dx$ are defined as usual for all $f, g \in L^2(\Omega)$. For the dataset $\{\phi_1, \ldots, \phi_M\} \subset L^2(\Omega)$, a proper orthogonal decomposition (POD) is sought, which is also called principal component analysis (PCA) or Karhunen-Loève transformation, see also [5] or [133]. Again, the superscripts are omitted for simplicity ($\phi_i := \overline{\widehat{\phi}}_i$, see (5.2)). Nonetheless, keep in mind that in this chapter, all functions are surrogate models which have been aligned by admissible transformations, while the POD method applies for any $\{\phi_i\}_{i=1}^M \subset L^2(\Omega)$ as well. In PCA, significant structures of the dataset are identified, which is realized by an orthogonal decomposition of the covariance matrix

$$C^{(\phi)} := \left[\langle \phi_i, \phi_j \rangle_{L^2(\Omega)}\right]_{i,j=1}^{M,M} \in \mathbb{R}^{M \times M}. \tag{5.13}$$

Often, mean-centered functions are considered, i.e. the proper orthogonal decomposition is performed on $\{\check{\phi}_1, \ldots, \check{\phi}_M\}$, $\check{\phi}_i := \phi_i - \frac{1}{M}\sum_{k=1}^M \phi_k$ [27]. This leads to a decomposition of the space of variations from the mean instead of the space spanned by the functions themselves. In the statistical shape model

$$\frac{1}{M}\sum_{k=1}^{M} \phi_k + \sum_{j=1}^{L} a_j \psi_j, \tag{5.14}$$

the principal components of variation $\psi_j \in L^2(\Omega)$ from the mean are controlled by parameters $a_j \in \mathbb{R}$ ($j = 1, \ldots, L$). Proper orthogonal decompositions of both mean-centered $\{\check{\phi}_1, \ldots, \check{\phi}_M\}$ and plain datasets $\{\phi_1, \ldots, \phi_M\}$ have been investigated. In this thesis' test cases, mean-centering did not produce better results while increasing the complexity of algorithmic implementation, thus in the remainder of this chapter the plain formulation is used.

The POD method is now introduced for functions in Hilbert space, mainly following the discussion of [145]. The M functions $\phi_i \in L^2(\Omega)$ ($i = 1, \ldots, M$) are assumed to be linear independent and $\phi(x) := (\phi_1(x), \ldots, \phi_M(x)) \in \mathbb{R}^{1\times M}$ is defined for all $x \in \Omega$. For the POD method, consider $\mathcal{Y}^M := \operatorname{span}\{\phi_1, \ldots, \phi_M\} \subset L^2(\Omega)$ of dimension M. Let $\{\psi_1, \ldots, \psi_M\}$ be an orthonormal basis of $\mathcal{Y}^M$. Clearly, when using the complete basis, every ϕ_i can be represented by a linear combination of its elements

$$\phi_i = \sum_{j=1}^{M} \langle \phi_i, \psi_j \rangle_{L^2(\Omega)} \psi_j \quad (i = 1, \ldots, M). \tag{5.15}$$

But now a set of orthonormal functions $\{\psi_1, \ldots, \psi_L\} \subset \mathcal{Y}^M$ of lower dimension $L \leq M$ is sought, which describes $\mathcal{Y}^M$ as good as possible, i.e. every $\phi_i \in \mathcal{Y}^M$ can be approximated by a linear combination of $\{\psi_1, \ldots, \psi_L\}$ in an optimal way.

Definition 5.1 (POD basis). *For any $L \leq M$, the solution $\{\psi_1, \ldots, \psi_L\} \subset \mathcal{Y}^M$ of the optimization problem*

$$\begin{aligned} \min_{\psi_1,\ldots,\psi_L \in \mathcal{Y}^M} & \sum_{i=1}^{M} \left\| \phi_i - \sum_{j=1}^{L} \langle \phi_i, \psi_j \rangle_{L^2(\Omega)} \psi_j \right\|^2_{L^2(\Omega)} \\ \text{s.t.} \quad & \langle \psi_i, \psi_j \rangle_{L^2(\Omega)} = \delta_{ij} \quad (i,j = 1, \ldots, L) \end{aligned} \tag{5.16}$$

is called POD basis of rank L.

Here, δ_{ij} denotes the Kronecker delta. Expanding the norm in (5.16) and exploiting $\langle \psi_i, \psi_j \rangle_{L^2(\Omega)} = \delta_{ij}$ yields an equivalent formulation:

$$\begin{aligned} \max_{\psi_1,\ldots,\psi_L \in \mathcal{Y}^M} & \sum_{i=1}^{M} \sum_{j=1}^{L} \langle \phi_i, \psi_j \rangle^2_{L^2(\Omega)} \\ \text{s.t.} \quad & \langle \psi_i, \psi_j \rangle_{L^2(\Omega)} = \delta_{ij} \quad (i,j = 1, \ldots, L). \end{aligned} \tag{5.17}$$

For a given $L \leq M$, the amount of $\mathcal{Y}^M$'s information contained in the POD basis is maximized.

In the following three theorems, the solution of (5.16) is derived. It can be expressed as the solution of the eigenvalue problem for an operator $\mathcal{C} : L^2(\Omega) \to \mathcal{Y}^M$ which is defined as follows

$$\mathcal{C}\psi := \sum_{i=1}^{M} \langle \psi, \phi_i \rangle_{L^2(\Omega)} \phi_i. \tag{5.18}$$

The existence of an orthonormal basis is guaranteed by the Hilbert-Schmidt theorem.

Theorem 5.2 (Hilbert-Schmidt). *Let $\mathcal{C} : H \to H$ be a bounded, compact, self-adjoint operator on a real, separable Hilbert space H. Then there exists a complete orthonormal basis $\{\psi_i\}_{i=1}^{\infty}$ for H satisfying*

$$\mathcal{C}\psi_i = \omega_i \psi_i \quad \textit{and} \quad \omega_i \xrightarrow[i\to\infty]{} 0. \tag{5.19}$$

Proof. A proof can be found in [113], Chap. VI, Sect. 5. □

It can easily be shown that $\mathcal{C}$ in (5.18) satisfies the assumptions of Theorem 5.2, cf. [145]. By the construction of $\mathcal{C}$, which is a mapping to the finite-dimensional $\mathcal{Y}^M \subset L^2(\Omega)$, there exist M orthonormal eigenfunctions $\{\psi_1, \ldots, \psi_M\} \subset \mathcal{Y}^M$ corresponding to the real and non-negative eigenvalues $\omega_1, \ldots, \omega_M$. Also, $\operatorname{span}\{\psi_1, \ldots, \psi_M\} = \mathcal{Y}^M$ which implies that $\{\psi_1, \ldots, \psi_M\}$ is an orthonormal basis for $\mathcal{Y}^M$. Furthermore, $\omega_i = 0$ for all $i > M$, cf. [77].

Theorem 5.3. *For the operator $\mathcal{C} : L^2(\Omega) \to \mathcal{Y}^M$ as defined in (5.18) exist an orthonormal basis $\{\psi_i\}_{i=1}^{\infty}$ and a sequence of real, non-negative numbers $\{\omega_i\}_{i=1}^{\infty}$ satisfying*

$$\mathcal{C}\psi_i = \omega_i \psi_i, \tag{5.20}$$

$$\omega_1 \geq \ldots \geq \omega_M > 0, \tag{5.21}$$

$$\omega_i = 0 \quad (i > M). \tag{5.22}$$

Then for any $L \leq M$, $\{\psi_1, \ldots, \psi_L\} \subset \mathcal{Y}^M$ is the solution of (5.16).

Proof. The proof is given in [145]. □

Thus the POD basis of rank L is given by the first L eigenfunctions ψ_i of $\mathcal{C}$ corresponding to the L largest eigenvalues ω_i $(i = 1, \ldots, L)$. A finite approach for solving the eigenvalue problem (5.20) is now deduced. A matrix

$$C := \left[\langle \phi_i, \phi_j \rangle_{L^2(\Omega)}\right]_{i,j=1}^{M,M} \in \mathbb{R}^{M \times M} \tag{5.23}$$

is defined (cf. (5.13)), for which the eigenvalue problem can be solved easily.

Theorem 5.4. *The orthonormal eigenfunctions $\{\psi_i\}_{i=1}^M$ and the corresponding decreasing eigenvalues $\{\omega_i\}_{i=1}^M$ of $\mathcal{C}$ can be computed via the solution of the finite-dimensional eigenvalue problem*

$$Cv^k = \omega_k v^k, \quad v^k \in \mathbb{R}^M \quad (k = 1, \ldots, M). \tag{5.24}$$

The eigenfunctions are given by

$$\psi_k = \frac{1}{\sqrt{\omega_k}} \sum_{i=1}^{M} v_i^k \phi_i \quad (k = 1, \ldots, M). \tag{5.25}$$

Proof. The proof follows the approach in [145] . Because $\psi_k \in \mathcal{Y}^M = \operatorname{span}\{\phi_1, \ldots, \phi_M\}$, the following ansatz is made using coefficients $\kappa_k \in \mathbb{R} \setminus \{0\}$, $v_i^k \in \mathbb{R}$ $(k = 1, \ldots, M, i = 1, \ldots, M)$

$$\psi_k = \kappa_k \sum_{i=1}^{M} v_i^k \phi_i \quad (k = 1, \ldots, M). \tag{5.26}$$

Inserting (5.26) into the eigensystem (5.20) yields

$$\kappa_k \sum_{j=1}^{M} \left(\sum_{i=1}^{M} \langle \phi_i, \phi_j \rangle_{L^2(\Omega)} v_i^k \right) \phi_j = \kappa_k \sum_{j=1}^{M} \left(\omega_k v_j^k \right) \phi_j \quad (k = 1, \ldots, M). \tag{5.27}$$

Exploiting the linear independence of $\{\phi_1, \ldots, \phi_M\}$, this is equivalent to

$$\sum_{i=1}^{M} \langle \phi_i, \phi_j \rangle_{L^2(\Omega)} v_i^k = \omega_k v_j^k \quad (k = 1, \ldots, M,\ j = 1, \ldots, M). \tag{5.28}$$

This is an eigenvalue problem for the matrix $C = [\langle \phi_i, \phi_j \rangle_{L^2(\Omega)}]_{i,j=1}^{M,M} \in \mathbb{R}^{M \times M}$

$$Cv^k = \omega_k v^k, \quad v^k \in \mathbb{R}^M \quad (k = 1, \ldots, M). \tag{5.29}$$

Thus $\mathcal{C} : L^2(\Omega) \to \mathcal{Y}^M$ and $C \in \mathbb{R}^{M \times M}$ have the same eigenvalues $\{\omega_i\}_{i=1}^M$. Setting $\kappa_k := \frac{1}{\sqrt{\omega_k}}$ for normalization in (5.26), the orthonormal eigenfunctions $\{\psi_i\}_{i=1}^M$ of $\mathcal{C}$ can be computed

using the entries of the orthonormal eigenvectors of C as linear coefficients

$$\psi_k = \frac{1}{\sqrt{\omega_k}} \sum_{i=1}^{M} v_i^k \phi_i \quad (k = 1, \dots, M), \tag{5.30}$$

from which the claim follows. □

So for determining the POD basis $\{\psi_1, \dots, \psi_L\}$ of rank L, only the L eigenvectors $v^k \in \mathbb{R}^M$ of C corresponding to the L largest eigenvalues ω_k have to be computed. With notations $\phi(x) := (\phi_1(x), \dots, \phi_M(x)) \in \mathbb{R}^{1\times M}$, $V_L := [v_i^k]_{i,k} \in \mathbb{R}^{M\times L}$ and $W_L := \operatorname{diag}(\sqrt{\omega_1}, \dots, \sqrt{\omega_L}) \in \mathbb{R}^{L\times L}$ one can also write

$$\psi_k(x) = \frac{1}{\sqrt{\omega_k}} \phi(x) v^k \quad (k = 1, \dots, L), \tag{5.31}$$

$$\psi(x) := (\psi_1(x), \dots, \psi_L(x)) = \phi(x) V_L W_L^{-1} \in \mathbb{R}^{1\times L}. \tag{5.32}$$

When the database elements

$$\phi_i = \sum_{j=1}^{M} \langle \phi_i, \psi_j \rangle_{L^2(\Omega)} \psi_j \in \mathcal{Y}^M \tag{5.33}$$

are not represented with the full basis $\{\psi_1, \dots, \psi_M\}$, but an approximation

$$\phi_i \approx \sum_{j=1}^{L} \langle \phi_i, \psi_j \rangle_{L^2(\Omega)} \psi_j \in \mathcal{Y}^L \tag{5.34}$$

is used which is a projection onto the subspace $\mathcal{Y}^L := \operatorname{span}\{\psi_1, \dots, \psi_L\} \subset \mathcal{Y}^M$, an approximation error is made. The total error can be expressed as the sum of the $M - L$ remaining eigenvalues of the truncated POD basis of rank L (cf. [145])

$$\sum_{i=1}^{M} \left\| \phi_i - \sum_{j=1}^{L} \langle \phi_i, \psi_j \rangle_{L^2(\Omega)} \psi_j \right\|_{L^2(\Omega)}^2 = \sum_{j=L+1}^{M} \omega_j. \tag{5.35}$$

This can be used for a heuristic determination of the required number L of POD basis elements. With $\sum_{j=1}^{M} \omega_j$ often being referred to as energy of the system $\mathcal{Y}^M$ [76], $L < M$ is chosen such that the relative amount of energy contained in $\mathcal{Y}^L$

$$\frac{\sum_{j=1}^{L} \omega_j}{\sum_{j=1}^{M} \omega_j} \tag{5.36}$$

satisfies at least a certain ratio. In reduced-order modeling (ROM), POD is used for reducing the dimension L of the space of basis functions for computational reasons. In the statistical shape analysis framework however, POD is applied for extracting only the most important structures $\{\psi_1, \dots, \psi_L\}$ of the database $\{\phi_1, \dots, \phi_M\}$, while the lesser important structures $\{\psi_{L+1}, \dots, \psi_M\}$ can be neglected.

5.3 Gappy POD in Hilbert spaces

The gappy POD method was initially introduced in [32] for reconstructing images of faces from incomplete data. A mask function was used which set the greyscale of every pixel to zero which was not part of the incomplete dataset. In [10], the gappy POD method was applied to CFD problems for the first time and a selection vector was used to set the missing flow solution vector's entries to zero. A straightforward approach for applying this methodology to $L^2(\Omega)$ is now introduced, where the gappy data is given by function evaluations on a discrete subset of Ω. Two cases are discussed, depending on which database is used for POD: the gappy POD method for the plain database (cf. Sect. 5.1.1) and, based on this, a gappy POD method for an aligned database (cf. Sect. 5.1.2).

5.3.1 Using the plain database

First, it is briefly illustrated how a "non gappy" function is approximated by the POD basis. When approximating an arbitrary function $y \in L^2(\Omega)$ (not necessarily $y \in \mathcal{Y}^M$) with the POD basis, the solution of the optimization problem

$$\min_{a_1^{(\psi)},\dots,a_L^{(\psi)} \in \mathbb{R}} \frac{1}{2} \left\| y(x) - \sum_{j=1}^{L} a_j^{(\psi)} \psi_j(x) \right\|_{L^2(\Omega)}^2 \tag{5.37}$$

is sought. This is a linear least squares problem and using optimality conditions and exploiting $\langle \psi_i, \psi_j \rangle_{L^2(\Omega)} = \delta_{ij}$,

$$a_j^{(\psi)} = \langle y, \psi_j \rangle_{L^2(\Omega)} \quad (j = 1, \dots, L) \tag{5.38}$$

can be derived. The approximating function $\widetilde{y}(x) := \sum_{j=1}^{L} a_j^{(\psi)} \psi_j(x)$ is a projection of $y(x)$ onto the subspace $\mathcal{Y}^L \subset \mathcal{Y}^M \subset L^2(\Omega)$.

Now for the gappy POD method, suppose $y(x)$ itself is unknown again and only a set of N data pairs of evaluations

$$\begin{gathered} \left\{ (x^{(i)}, y(x^{(i)})) \right\}_{i=1}^{N}, \\ x^{(i)} \in \Omega \subset \mathbb{R}^d,\ y(x^{(i)}) \in \mathbb{R},\ x^{(i)} \neq x^{(j)}\ (i \neq j), \quad (i = 1, \dots, N) \end{gathered} \tag{5.39}$$

is available. Based on the data, a reconstruction of the unknown function $y \in L^2(\Omega)$ is sought. Assuming there is a strong relation between y and $\mathcal{Y}^M$, y is approximated by a $\widetilde{y} \in \mathcal{Y}^L$. Similar to (5.37), an optimization problem is posed

$$\min_{a_1^{(\psi)},\dots,a_L^{(\psi)} \in \mathbb{R}} \frac{1}{2} \sum_{i=1}^{N} \left(y(x^{(i)}) - \sum_{j=1}^{L} a_j^{(\psi)} \psi_j(x^{(i)}) \right)^2, \tag{5.40}$$

which again is a linear least squares problem. Clearly, though the ψ_j are orthonormal in $L^2(\Omega)$, meaning $\langle \psi_i, \psi_j \rangle_{L^2(\Omega)} = \delta_{ij}$, this is not the case anymore on a subset of Ω. Particularly, $\sum_{k=1}^{N} \psi_i(x^{(k)}) \psi_j(x^{(k)})$ $(i \neq j)$ is generally not equal to zero. Introducing the design

matrix

$$\Psi := \left[\psi_j(x^{(i)})\right]_{i=1,j=1}^{N,L} \in \mathbb{R}^{N\times L},\ N \geq L,\ \operatorname{rank}\Psi = L. \tag{5.41}$$

and $Y := (y(x^{(i)}), \ldots, y(x^{(N)}))^\top \in \mathbb{R}^N$, the solution of (5.40) is given by the linear equation

$$\Psi^\top \Psi a^{(\psi)} = \Psi^\top Y, \tag{5.42}$$

cf. Theorem 2.1.

Since every ψ_j $(j = 1, \ldots, L)$ is a linear combination of $\{\phi_1, \ldots, \phi_M\}$, a framework is now derived for avoiding multiple redundant evaluations of $\phi_i(x)$ both while solving (5.42) and also for evaluating $\widetilde{y}(x) = \sum_{j=1}^{L} a_j^{(\psi)}\psi_j(x)$. Even if each $\phi_j(x)$ is a surrogate model, depending on its complexity and L, M and N, the required evaluations can be a bottleneck in the model generation. Defining

$$\Phi := \left[\phi_j(x^{(i)})\right]_{i=1,j=1}^{N,M} \in \mathbb{R}^{N\times M}, \tag{5.43}$$

analogously to (5.32) one obtains

$$\Psi = \Phi V_L W_L^{-1}. \tag{5.44}$$

Also, with (5.32) the approximating function reads

$$\begin{aligned}\widetilde{y}(x) &= \sum_{j=1}^{L} a_j^{(\psi)}\psi_j(x)\\ &= \psi(x)a^{(\psi)}\\ &= \phi(x)\underbrace{V_L W_L^{-1} a^{(\psi)}}_{:=a^{(\phi)}\in\mathbb{R}^{M\times 1}},\end{aligned} \tag{5.45}$$

such that a closed form of $\widetilde{y}(x) = \phi(x)a^{(\phi)} = \sum_{j=1}^{M} a_j^{(\phi)}\phi_j(x)$ as a linear combination of $\{\phi_1, \ldots, \phi_M\}$ is derived.

5.3.2 Using an aligned database

If the database elements $\{\phi_1, \ldots, \phi_M\}$ have been transformed by solving the correspondence problem (5.3), the unknown function $y(x)$ (i.e. its known evaluations $\{(x^{(i)}, y(x^{(i)}))\}_{i=1}^{N}$, respectively) has to be allowed a transformation of the same class. At this point the task is not finding an overall alignment, but fitting the POD basis to a fixed set of data pairs $\{(x^{(i)}, y(x^{(i)}))\}_{i=1}^{N}$. So rather than transforming $y(x)$, equivalently a transformation is applied to the approximating function $\widetilde{y}(x)$, parametrized by p. The double bar indicates that a second transformation is applied after the already known initial transformation of the

database functions by alignment (parametrized by q):

$$\overline{\overline{x}}(p) := \begin{pmatrix} \overline{x}_1(q^j)(1+p_1)+p_2 \\ \overline{x}_2(q^j)(1+p_3)+p_4 \end{pmatrix} \tag{5.46}$$

$$= \begin{pmatrix} \left(x_1(1+q_1^j)+q_2^j\right)(1+p_1)+p_2 \\ \left(x_2(1+q_3^j)+q_4^j\right)(1+p_3)+p_4 \end{pmatrix}, \tag{5.47}$$

$$\overline{\phi}_j\left(\overline{\overline{x}}(p), q^j\right) := \phi_j\left(\overline{\overline{x}}(p)\right)(1+q_5^j)+q_6^j, \tag{5.48}$$

$$\widetilde{\overline{y}}\left(\overline{\overline{x}}(p), p, a^{(\psi)}\right) := \widetilde{y}\left(\overline{\overline{x}}(p), a^{(\psi)}\right)+p_5 \tag{5.49}$$

$$= \overline{\phi}\left(\overline{\overline{x}}(p), q\right) V_L W_L^{-1} a^{(\psi)} + p_5. \tag{5.50}$$

Here p does not contain a scaling parameter like q^j in (5.2), because scaling is already performed via $a^{(\psi)}$. The linear least squares problem (5.40) is then augmented by the transformation parameters $p \in \mathbb{R}^5$ and again a regularization term is included:

$$\min_{a^{(\psi)}\in\mathbb{R}^L, p\in\mathbb{R}^5} \left\{ \frac{1}{2}\sum_{i=1}^{N}\left(y(x^{(i)}) - \widetilde{\overline{y}}\left(\overline{\overline{x}}^{(i)}(p), p, a^{(\psi)}\right)\right)^2 + \frac{\delta}{2}p^\top p \right\}. \tag{5.51}$$

Together with the alignment of the database, (5.51) is a nonlinear enhancement of the common gappy POD method. This nonlinear least squares problem is solved by a Gauß-Newton algorithm for a and p simultaneously. Unlike (5.3), this can be accomplished without computational issues. Intuitively, an initial value for the algorithm is $p = 0$ (no transformation) and its corresponding linear least squares solution $a^{(\psi)} = (\Psi^\top\Psi)^{-1}\Psi^\top Y$ (5.42). The approximation $\widetilde{\overline{y}}(\overline{\overline{x}}(p), p, a^{(\psi)})$ to the unknown function $y(x) \in L^2(\Omega)$ based on the data pairs of evaluations $\{(x^{(i)}, y(x^{(i)}))\}_{i=1}^N$ and the function database $\{\phi_1, \ldots, \phi_M\}$ is called *generic surrogate model* (GSM).

It is now briefly clarified how this new approach is different from the usual application of POD in numerical simulation and optimization, e.g. in CFD. In reduced-order modeling (ROM) [76, 145], the structure of the underlying (PDE-) system is investigated and POD is applied to a set of snapshots, which represent for instance spatial solutions for particular points in time. The aim is to reduce the numerical effort by solving the system on a lower-dimensional space spanned by the POD basis. Here, on the contrary, POD is used for identifying the most important structures of the data set $\{\phi_1, \ldots, \phi_M\}$ based on an approach used in the field of statistical shape models. The database functions ϕ_i do not emerge from different snapshots in time or another known parametrization. They are responses based on RANS computations for different airfoil geometries, whose connection among each other is unclear, except for the assumption that they share structural similarities. For the new unknown response $y \in L^2(\Omega) \supset \mathcal{Y}^{\mathcal{M}} = \operatorname{span}\{\phi_1, \ldots, \phi_M\}$ only this assumption holds. Therefore, classical error analysis [77, 13, 14] does not apply in this framework. Furthermore, the only available information on y is given by the actual evaluations $\{(x^{(i)}, y(x^{(i)}))\}_{i=1}^N$. The gappy POD approximation minimizes the error in the evaluations $x^{(i)}$ (5.51), while the behavior of $y(x)$ in $\Omega \setminus \{x^{(1)}, \ldots, x^{(N)}\}$ is generally unknown, except for the assumed structural similarities with $\mathcal{Y}^M$.

5.4 Hierarchical Kriging

In this section, the generic surrogate model will be extended from a least squares approximation to an interpolation method. The information contained in the data points is considered as extremely valuable, especially since the evaluations of $y(x^{(i)})$ are assumed computationally very expensive. So a surrogate model should be as accurate as possible, particularly in the proximity of any $x^{(i)}$, which can be realized by interpolation rather than approximation. Therefore, a Kriging type data fusion method is introduced to generate an interpolation model which uses the generic surrogate model as a global trend.

Variable-fidelity modeling (VFM) comprises methods for improving the approximation quality of interpolating only few computationally expensive evaluations (high-fidelity) when having access to secondary data. This secondary data may consist of cheaper computations of a less accurate model (low-fidelity) or a second variable which is assumed to be correlated with the primary variable. For instance in CFD, computations with a Navier-Stokes code are regarded as high-fidelity data, while computations of the same problem with an Euler code serve as low-fidelity data. In this approach, the generic surrogate model $\widetilde{\overline{y}}$ (5.50) is used as the low-fidelity model to improve the interpolation quality of the (high-fidelity) evaluations $\{(x^{(i)}, y(x^{(i)}))\}_{i=1}^{N}$. So far, two major VFM frameworks could be distinguished. *Cokriging*, originally developed in geostatistics, establishes a relation between primary and auxiliary variable by cross correlation [146]. Other VFM methods use an (additive, multiplicative or hybrid) *bridge function*, which corrects the discrepancy between a lo-fi and a hi-fi surrogate model [52]. Recently, a new robust VFM method was introduced [53], whose implementation and computational complexity does not exceed the common Kriging method. It is called *hierarchical Kriging* and the ansatz is a straightforward extension of the Kriging method (Sect. 2.3).

In the Kriging model (2.27), the regression term is replaced by the lo-fi model (in this case the generic surrogate model $\widetilde{\overline{y}}$)

$$y(x) = \beta\widetilde{\overline{y}}\,(\overline{\overline{x}}(p), p, a) + z(x), \quad x \in \Omega \subset \mathbb{R}^d, \tag{5.52}$$

with $\beta \in \mathbb{R}$, $z(x)$ capturing the lack of fit like in (2.28). Analogously to (2.26), the hierarchical Kriging predictor

$$\widehat{y}(x) = \lambda(x)^\top Y \tag{5.53}$$

is a weighted sum of $Y = (y(x^{(1)}), \ldots, y(x^{(N)}))^\top$. Again, the weights $\lambda_i(x)$ are determined by solving a linear equation

$$\begin{bmatrix} R & \widetilde{F} \\ \widetilde{F}^\top & 0 \end{bmatrix} \begin{pmatrix} \lambda(x) \\ \mu(x) \end{pmatrix} = \begin{pmatrix} r(x) \\ \widetilde{f}(x) \end{pmatrix} \tag{5.54}$$

which minimizes the mean squared error subject to the unbiasedness constraint, see Theorem 2.4. R, $r(x)$, $\lambda(x)$ and $\mu(x)$ are the same as in Theorem 2.4, while

$$\widetilde{F} := \left(\widetilde{\overline{y}}(\overline{\overline{x}}^{(i)}(p), p, a)\right)_{i=1}^{N} \in \mathbb{R}^N, \tag{5.55}$$

$$\widetilde{f}(x) := \widetilde{\overline{y}}\,(\overline{\overline{x}}(p), p, a) \in \mathbb{R} \tag{5.56}$$

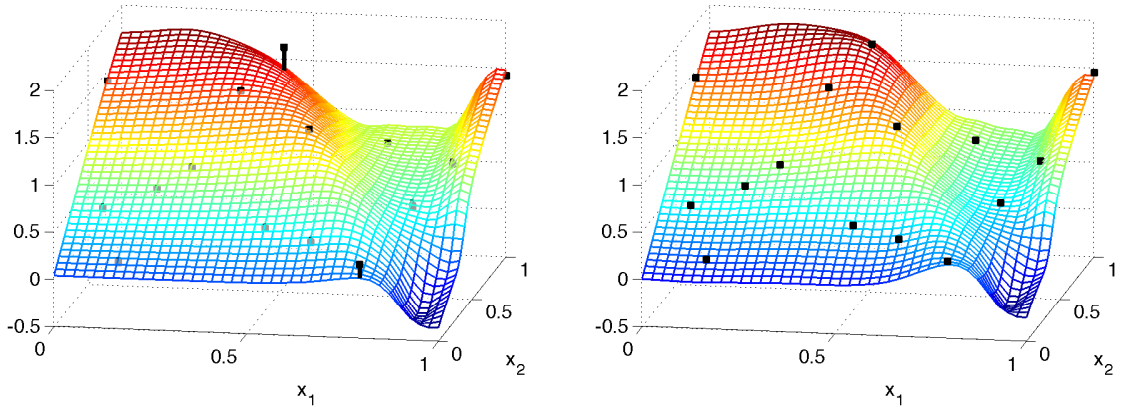

Figure 5.3: The generic surrogate model $\widetilde{\widetilde{y}}$ (left) and the hierarchical Kriging predictor $\widehat{y}$ (right) for $N = 15$ lift evaluations based on a LHC design.

replace the linear regression design matrix and $\widetilde{f}(x)$ contains the evaluation of the generic surrogate model in x. Analogously, a closed form of the hierarchical Kriging predictor is derived

$$\widehat{y}(x) = \lambda(x)^\top Y = \begin{pmatrix} r(x) \\ \widetilde{f}(x) \end{pmatrix}^\top \begin{bmatrix} R & \widetilde{F} \\ \widetilde{F}^\top & 0 \end{bmatrix}^{-1} \begin{pmatrix} Y \\ 0 \end{pmatrix} \tag{5.57}$$

or equivalently with $\beta = \left(\widetilde{F}^\top R^{-1} \widetilde{F}\right)^{-1} \widetilde{F}^\top R^{-1} Y$

$$\widehat{y}(x) = \beta \widetilde{f}(x) + r(x)^\top R^{-1} \left(Y - \beta \widetilde{F}\right). \tag{5.58}$$

Formulation (5.57) is more suitable for the implementation, because β does not need to be computed explicitly and the solution of the linear system $\begin{bmatrix} R & \widetilde{F} \\ \widetilde{F}^\top & 0 \end{bmatrix}^{-1} \begin{pmatrix} Y \\ 0 \end{pmatrix}$ only has to be computed once since it is independent of x. It also turns out that, inserting any $x^{(i)}$ $(i = 1, \ldots, N)$, the hierarchical Kriging predictor is an interpolator $(\widehat{y}(x^{(i)}) = y(x^{(i)}))$, cf. Sect. 2.3.2. Formulation (5.58) demonstrates how the predictor works. Assuming that the generic surrogate model fit $\widetilde{\widetilde{y}}$ approximates the data $y(x^{(i)})$, β will be close to 1. The second summand is a weighted sum of correlation functions which "pulls" the response towards the exact evaluations $y(x^{(i)})$. An example is given in Fig. 5.3. The generic surrogate model $\widetilde{\widetilde{y}}$ already is a suitable approximation of the given data Y, with two exceptions in the upper middle and the lower right, where the behavior of $y(x)$ can not be reproduced exactly. The hierarchical Kriging predictor however corrects these discrepancies and interpolates the data correctly.

6 Numerical investigation of generic surrogate modeling

The generic surrogate modeling framework is now validated with numerical examples. In Sect. 6.1, a training database for aerodynamic simulation is proposed and the alignment of the database elements is discussed. The results of its proper orthogonal decomposition are examined in Sect. 6.2. Three distinct test cases are introduced in Sect. 6.3. The performance of hierarchical Kriging based on generic surrogate models is compared to common Kriging and gradient-enhanced Kriging interpolations. In Sect. 6.4, an error indicator for generic surrogate models is introduced, which can assess the suitability of GSM in the particular test cases. A concluding summary is given in Sect. 6.5.

6.1 Training database for aerodynamic simulation

As in Chap. 4, the goal is to generate globally accurate surrogate models for aerodynamic coefficients depending on the input parameters Mach number and angle of attack α. For two-dimensional airfoil geometries, the 2d RANS equations can be solved numerically with the TAU code to evaluate the responses lift, drag and pitching moment, cf. Sect. 4.1. Each coefficient $c_l(\mathrm{Ma}, \alpha)$, $c_d(\mathrm{Ma}, \alpha)$, $c_m(\mathrm{Ma}, \alpha)$ constitutes a distinct problem class. Within each of these problem classes, responses based on different airfoil geometries are assumed to share structural similarities. For the numerical investigation of the new generic surrogate modeling approach the following database is considered. 23 airfoils of the NACA four-digit series are chosen, where the first digit varies within $[3.0, 6.0]$, the second within $[1.0, 6.0]$ and the last two digits which determine the thickness are fixed at 12, see Fig. 6.1. In a previous study [118], also the RAE2822 airfoil was included in the database. It is excluded now and instead, the RAE2822 responses are considered an additional test case for investigating the performance of GSM approximations in Sect. 6.3.

For each airfoil geometry, an accurate surrogate model for each problem class ($c_l(\mathrm{Ma}, \alpha)$, $c_d(\mathrm{Ma}, \alpha)$, $c_m(\mathrm{Ma}, \alpha)$) is computed in the reference domain $\Omega := [0.2, 0.9] \times [-4°, +12°]$. Note that if alignment of the database is considered (Sect. 5.1.2), the surrogates should be accurate in a domain larger than the reference domain (e.g. $\widetilde{\Omega} := [0.1, 1.0] \times [-6°, +14°]$) due to possible translations (5.1) in the alignment process, see Fig. 6.2. For each airfoil, 400 CFD solutions are computed on a 20×20 factorial design which covers the domain $\widetilde{\Omega}$. This sums up to 9200 CFD simulations for generating the database. Depending on the input parameters (Ma, α), one flow solution takes from 30 minutes to over 3 hours of CPU time. All computations can be performed independently from each other in parallel and using two AMD Opteron 6100 architectures with 48 2.3GHz cores each, the generation of the databases took approximately two weeks.

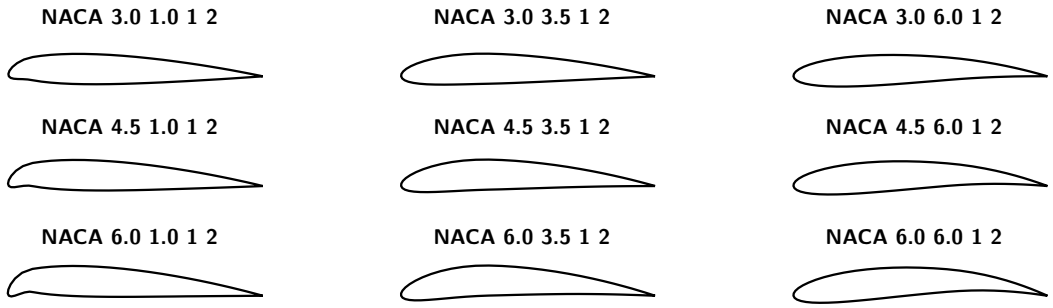

Figure 6.1: Airfoil database, 9 examples from the NACA four-digit series.

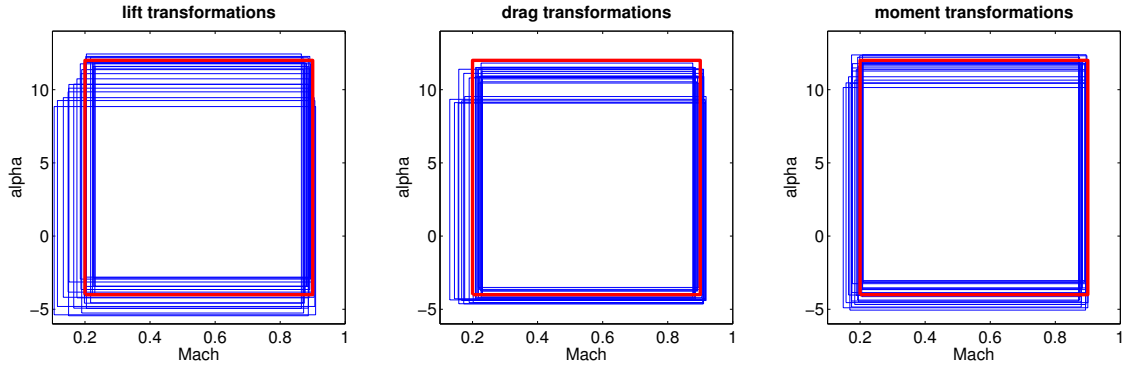

Figure 6.2: Reference domain Ω (red) and preimages of Ω (blue) for the transformations (5.1).

The plain responses $\{\phi_1, \ldots, \phi_{23}\}$ (without alignment) are depicted in Figs. 6.3–6.5. Though emerging from different airfoil geometries, it is evident that there are mutual characteristics. In the lift case, all responses show a linear behaviour for low Mach numbers, there is a ridge reaching from the lower right to the upper left, followed by a broad valley in the right half and also a local maximum and a local minimum appear near the right corners. All drag responses show a constant behavior for low Mach numbers, a slow ascend in the upper right half and a local maximum near the upper right corner. In the moment case, the responses share a linear domain in the left part, a broad ridge in the upper middle, a steep descent with a narrow valley in the upper right and a local maximum near the lower right corner. But these shared structural similarities do not coincide exactly in their position and level, and they also vary in shape. Given that, automatically identifying these mutual characteristics by POD can be problematic, because the dataset $\{\phi_1, \ldots, \phi_{23}\}$ contains too much variance. POD is a linear decomposition. If, for instance, two responses share a local bump whose maximum does however not have the same coordinates, POD will detect it as two distinct modes of variation, which is not desired. Therefore, the correspondence problem (5.3) is solved and the mutual characteristics are aligned as well as possible by admissible transformations (5.1)–(5.2), both in Ω and the image space $\phi_i(\Omega)$. Since the transformations merely are affine transformations in every dimension, the individual shapes of the respones ϕ_i are preserved. The total variance of the dataset $\{\phi_1, \ldots, \phi_{23}\}$ is clearly reduced, cf. Fig. 6.6.

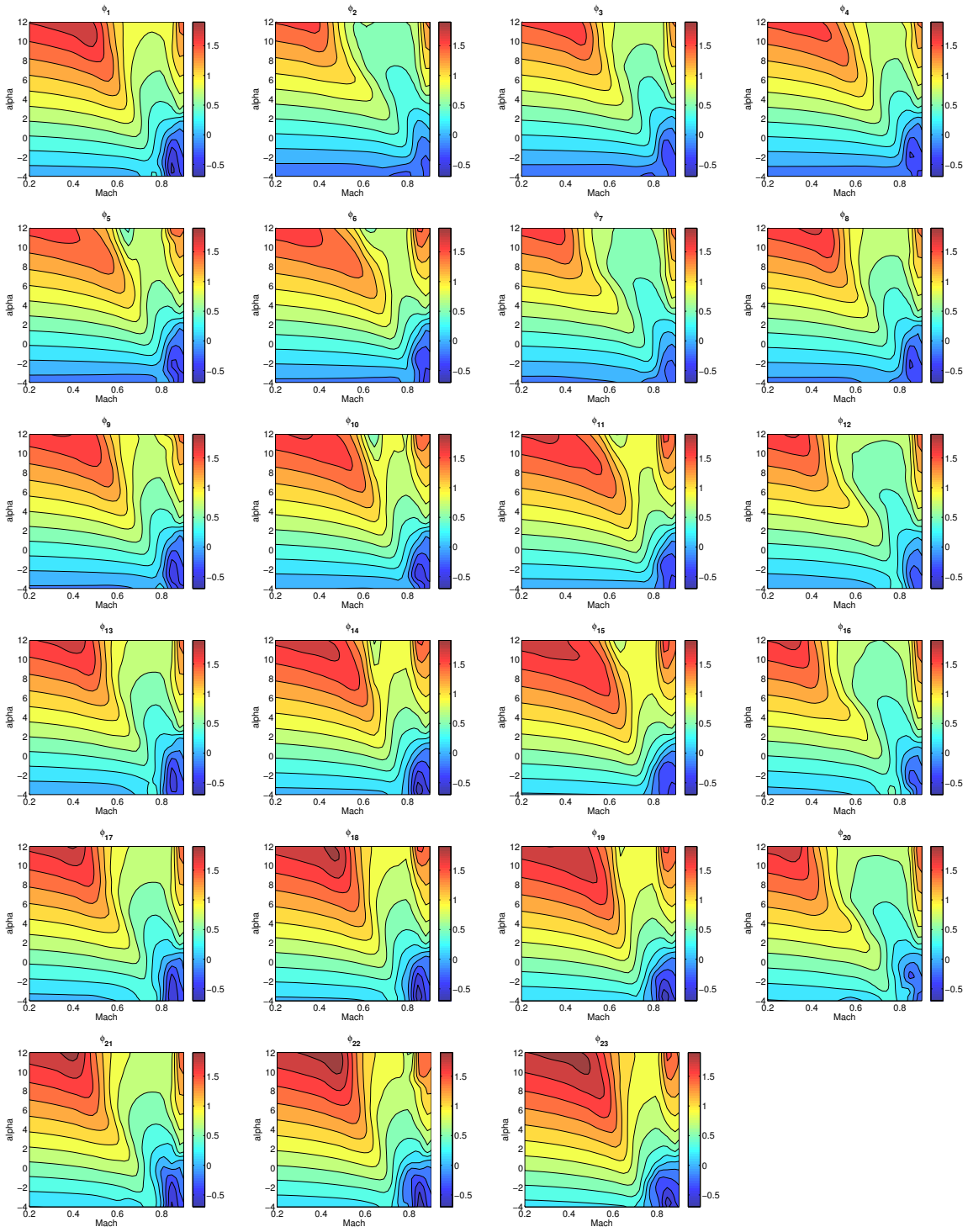

Figure 6.3: Function database $\{\phi_1, \ldots, \phi_{23}\}$ for the c_l case, no alignment.

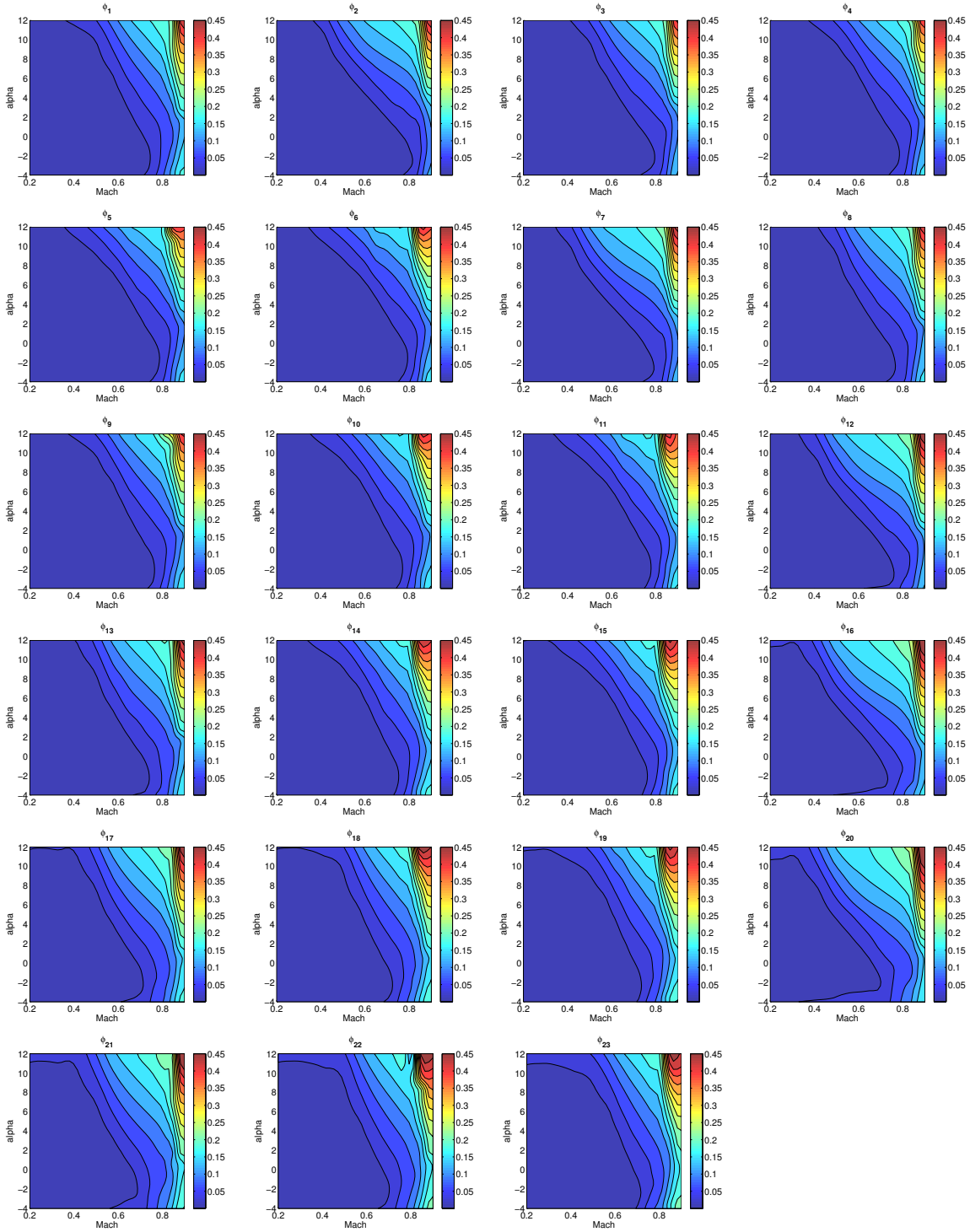

Figure 6.4: Function database $\{\phi_1, \ldots, \phi_{23}\}$ for the c_d case, no alignment.

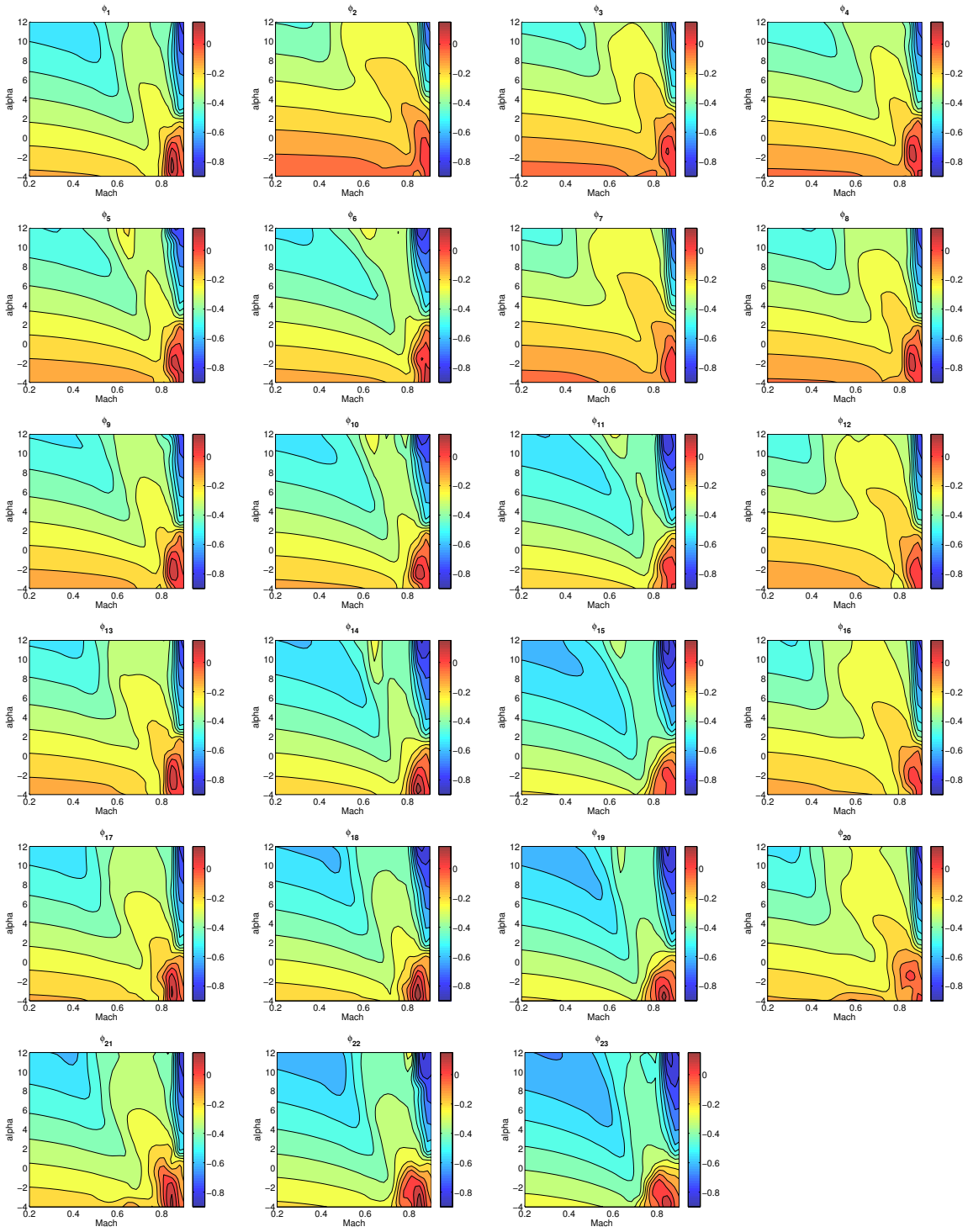

Figure 6.5: Function database $\{\phi_1, \ldots, \phi_{23}\}$ for the c_m case, no alignment.

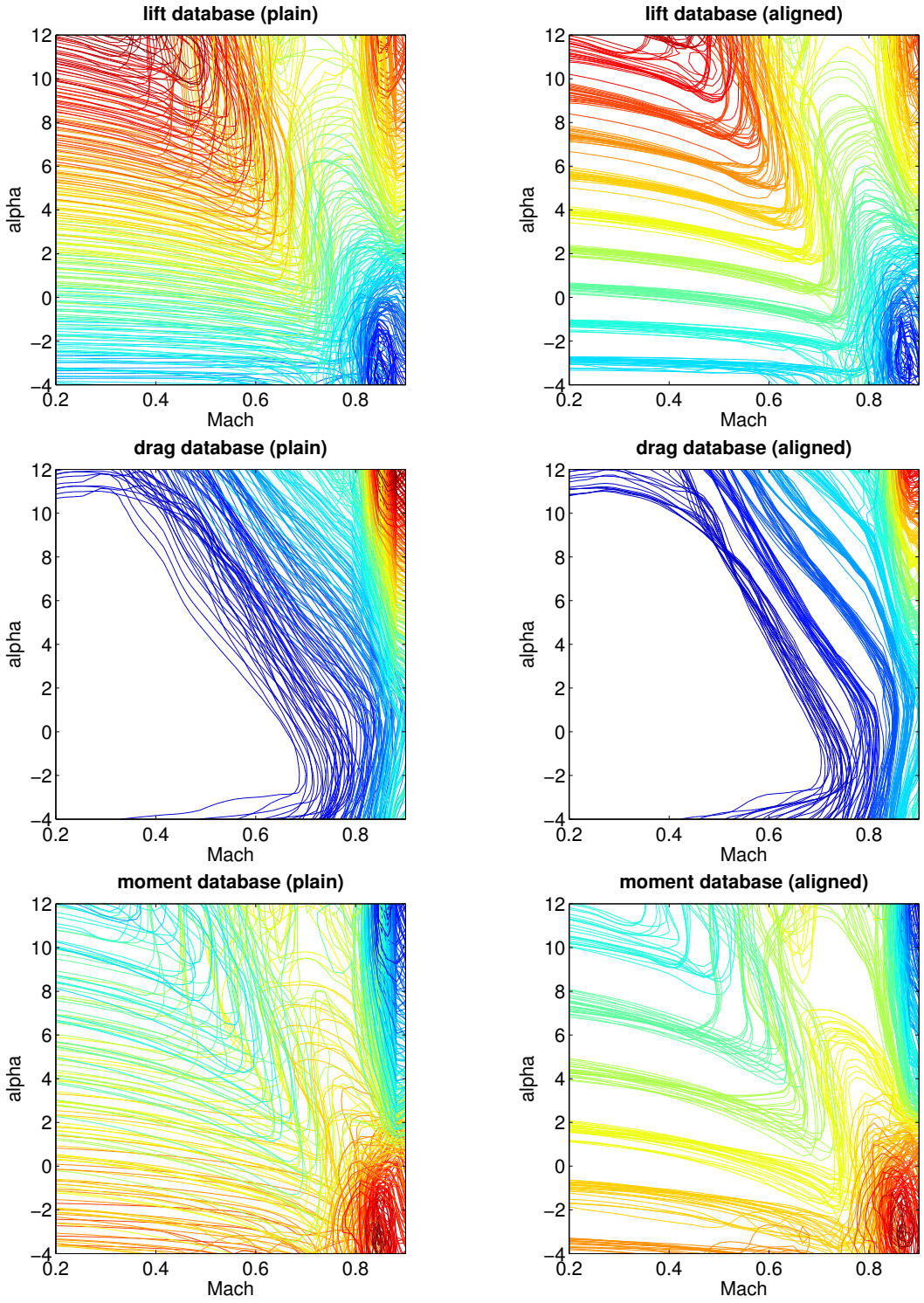

Figure 6.6: Contour plots for all $M = 23$ responses before (left) and after alignment (5.1)–(5.2) (right).

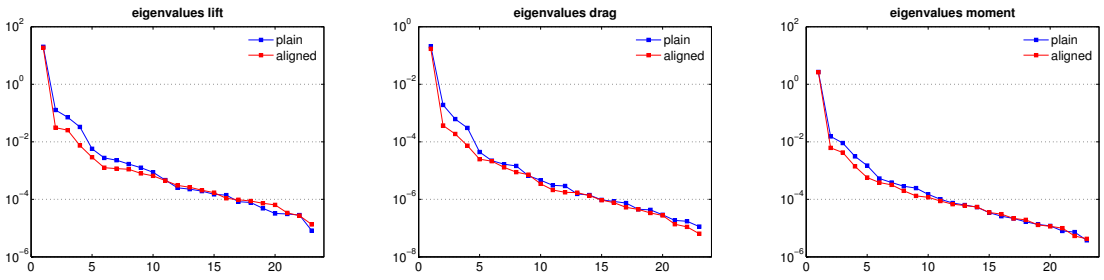

Figure 6.7: Distributions of the $M = 23$ eigenvalues ω_i for the plain and the aligned databases.

6.2 POD basis

Having generated the database of $M = 23$ response functions $\phi_1, \ldots, \phi_M$, the proper orthogonal decomposition of $\mathcal{Y} = \text{span}\{\phi_1, \ldots, \phi_M\}$ is performed. The most influential structures of the database are thus identified. The POD basis $\{\psi_1, \ldots, \psi_L\} \subset \mathcal{Y}^M$ is computed by solving the eigenvalue problem for the matrix $C = [\langle \phi_i, \phi_j \rangle_{L^2(\Omega)}]_{i,j=1}^{M,M} \in \mathbb{R}^{M \times M}$ and defining $\psi_k = \frac{1}{\sqrt{\omega_k}} \sum_{i=1}^{M} v_i^k \phi_i$, where ω_k and v^k are the eigenvalues and orthonormal eigenvectors $(k = 1, \ldots, L)$. The POD basis' rank L, i.e. the number of POD basis elements, is determined automatically. Using the approximation error formula (5.35), L is chosen as the smallest integer that satisfies

$$\frac{\sum_{j=1}^{L} \omega_j}{\sum_{j=1}^{M} \omega_j} \geq 0.999. \tag{6.1}$$

In this way, it is guaranteed that the POD basis $\{\psi_1, \ldots, \psi_L\}$ contains at least 99.9% of $\mathcal{Y}^M$'s total energy, cf. (5.36). Table 6.1 shows the number of required POD basis elements for the three problem classes. Using an aligned database reduces the number by one in each case, meaning that by aligning the characteristic features of the database functions the total variation is reduced (cf. also Fig. 6.6). Figure 6.7 illustrates the rapid decay of the eigenvalues, therefore already three to five POD basis elements are sufficient for (6.1) in these test cases.

	c_l	c_d	c_m
plain database	4	4	5
aligned database	3	3	4

Table 6.1: Rank L of the POD basis.

Remember that approximative reconstructions of the database elements are given by linear combinations $\sum_{j=1}^{L} a_j \psi_j \in \mathcal{Y}^L$ (plus additional nonlinear transformations, if alignment is used). The properties of ψ_1, ψ_2 and ψ_3 for aligned databases are depicted in Fig. 6.8. Clearly, in all cases ψ_1 is approximately a mean response of the database $\{\phi_1, \ldots, \phi_M\}$. In the lift case, ψ_2 controls the curvature from the lower right to the upper middle as well as the shape of the ridge on the right boundary. ψ_3 also affects the curvature in the middle and

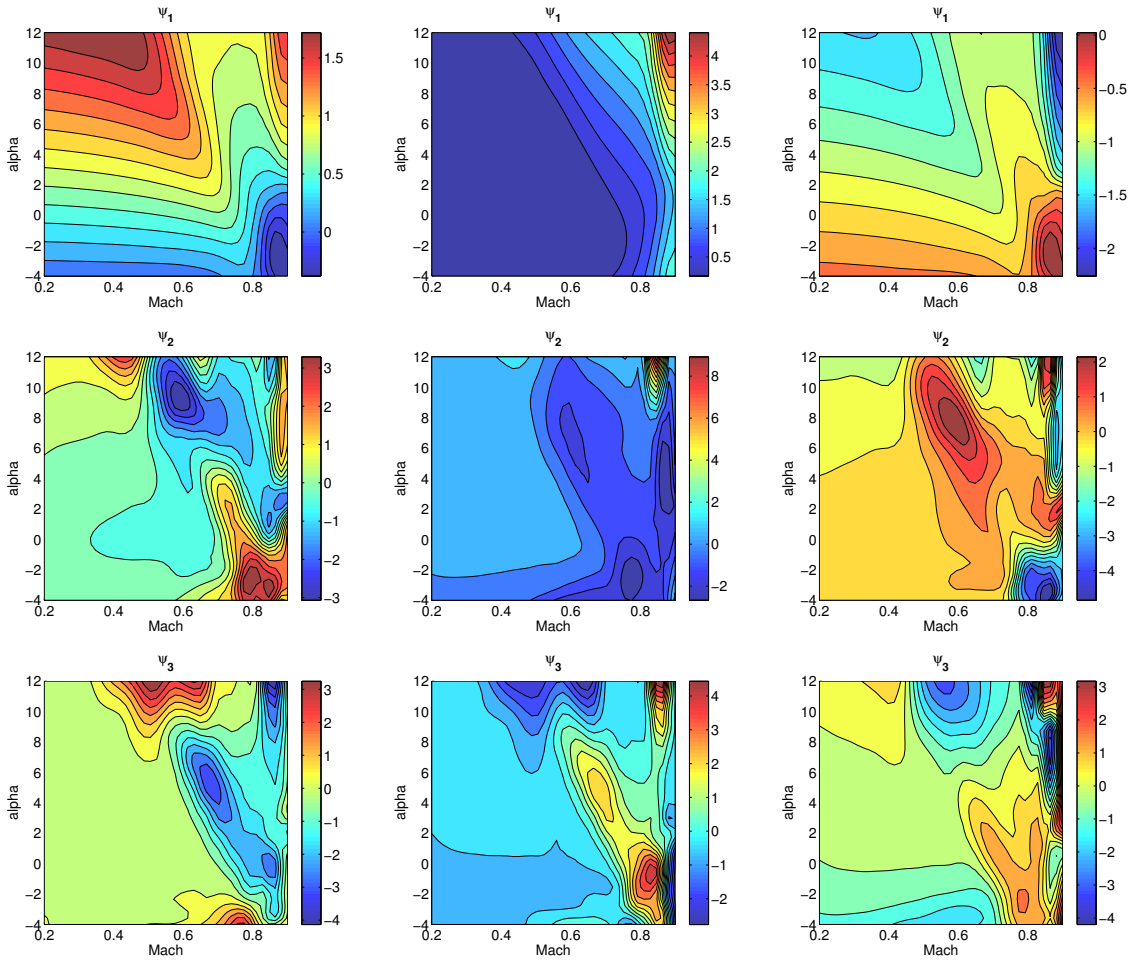

Figure 6.8: POD bases of rank $L = 3$ for aligned databases. Lift (left), drag (middle) and pitching moment (right).

additionally has influence on the position of the local maximum in the upper right corner. In the drag case, ψ_2 essentially shapes the local maximum in the upper right corner and ψ_3 additionally molds the ascend in the right half. For the moment responses, ψ_2 and ψ_3 both form the ridge in the middle as well as the valley and the local extrema near the right boundary.

6.3 Comparison of surrogate modeling approaches

With the POD basis $\{\psi_1, \ldots, \psi_L\}$ it is now possible to compute a generic surrogate model $\widetilde{f}(x)$ (5.56) and a hierarchical Kriging interpolation $\widehat{y}_{\mathrm{HK}}(x)$ (5.57) for new test cases, i.e. for responses $y(x)$ of new airfoil geometries. The following test cases are investigated, which will be referred to as test cases A, B and C.

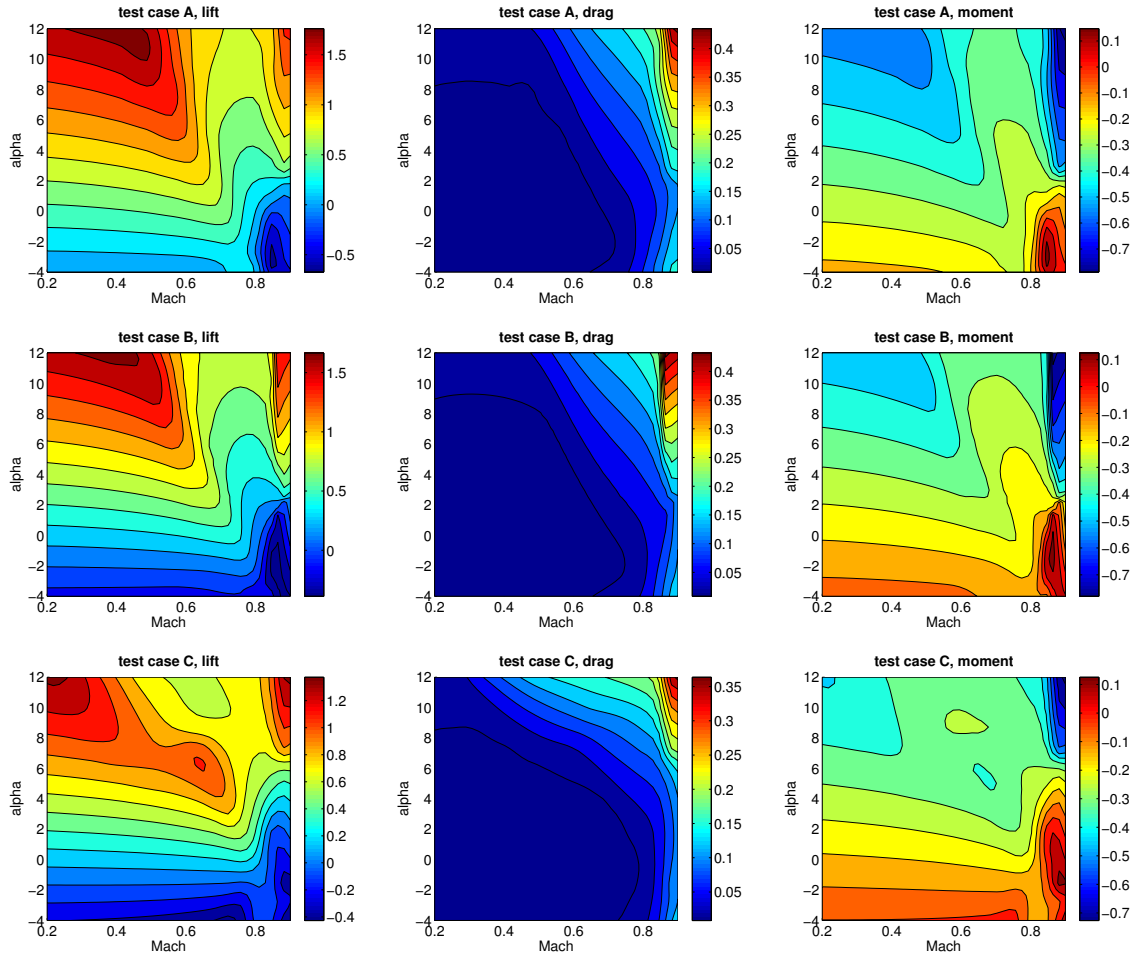

Figure 6.9: Validation data for test cases A, B and C.

A: NACA 4.500 3.500 1 2, an airfoil already contained in the database. The responses $y(x)$ based on this airfoil satisfy $y \in \mathcal{Y}^M = \text{span}\{\phi_1, \ldots, \phi_M\}$.

B: NACA 3.375 2.875 1 2, an airfoil not contained in the database. However, due to the mutual parametrization of the NACA airfoil geometries it is assumed that the responses y are strongly related to $\mathcal{Y}^M$.

C: RAE2822, not contained in the database. Neither does it belong to the family of the database airfoils, so the connection of the responses y and $\mathcal{Y}^M$ remains unclear.

The RANS equations are solved numerically with the DLR TAU solver for each test case. A set of validation data for the responses $c_l(\text{Ma}, \alpha)$, $c_d(\text{Ma}, \alpha)$, $c_m(\text{Ma}, \alpha)$ is generated on a 40×40 factorial design $\Omega^{\text{val}} \subset \Omega = [0.2, 0.9] \times [-4°, +12°]$, see Fig. 6.9. With the validation data, the relative root mean squared error η_2 (4.2) of any surrogate model can be evaluated.

In each test case, hierarchical Kriging interpolations based on generic surrogate models are computed both for plain databases (HK plain) and aligned databases (HK aligned). They are compared to common Kriging interpolations (KRI) and gradient-enhanced Kriging interpolations (GEK). Note that in the GEK case, triple the amount of online information of the actual response $y(x)$ is contained in the surrogate (evaluations in the N samples $X \subset \Omega$ plus partial derivatives w.r.t. each dimension). Hierarchical Kriging and common Kriging use the same amount of online information, while in HK the global trend is assumed to be captured by the offline information contained in the POD basis via the GSM. The four surrogate models are compared regarding global accuracy using the error measure η_2, based both on one-stage designs as well as adaptive designs. As one-stage designs, Halton designs of size $N = 5, 10, 15, 20, \ldots, 100$ are used, cf. Sect. 3.1.4. When using adaptive sampling, the SSE strategy is used for Kriging and GEK, cf. Sect. 3.2.2. It generally performed best in the comparative study of sampling strategies in Chap. 4. Basically, any adaptive strategy for Kriging can be applied in hierarchical Kriging as well. But it is possible to specify a custom adaptive sampling strategy for variable-fidelity methods as proposed in [52]. In the following, it is distinguished between the more accurate hierarchical Kriging $\widehat{y}_{\mathrm{HK}}(x)$ and the common Kriging interpolation $\widehat{y}_{\mathrm{KRI}}(x)$. Assuming that both surrogate models will converge to the true function $y(x)$ with growing (maybe very large) number N of samples, a new sample is added where the difference between both models is highest, i.e.

$$x^{(N+1)} = \arg\max_{x \in \Omega} \left| \widehat{y}_{\mathrm{HK}}(x) - \widehat{y}_{\mathrm{KRI}}(x) \right| . \tag{6.2}$$

Note that the difference is equal to zero in all existing samples $x^{(i)}$, because both surrogates are interpolations.

The sample distributions of this strategy are depicted in Fig. 6.10 for HK based on aligned databases, combined with contour line plots of the surrogates $\widehat{y}_{\mathrm{HK}}(x)$ for $N = 100$. The samples are highly concentrated in critical regions of the input parameter domain Ω where the response has large gradients or curvature, i.e. around $\mathrm{Ma} = 0.9$ for all responses or along the ridge in the middle for the c_l responses. Less critical regions like the almost linear region for low Mach numbers are sampled coarser, but are not neglected either. Apparently, strategy (6.2) is a mixed exploitation and exploration strategy and therefore uses the expensive evaluations $y(x^{(i)})$ efficiently without sacrificing accuracy, cf. Sect. 4.6. The results of this comparative study depend on the investigated test case, see Figs. 6.11–6.13. Therefore they will now be elaborated separately.

In test case A (Fig. 6.11), both HK surrogates clearly outperform KRI and GEK. The benefit is largest for small sample sizes. If N is small, the global structural information of the surrogate $\widehat{y}_{\mathrm{HK}}(x)$ is mainly given by the POD basis and not by $Y = (y(x^{(1)}), \ldots, y(x^{(N)}))^{\top}$. With the true response y being an element of the database, apparently only few samples are necessary such that it can be globally reproduced by the POD basis with adequate accuracy. Despite the good level of accuracy already for small N, there is no further drastic improvement with a larger number of samples as in KRI and GEK, especially for the non-adaptive Halton designs. By adding more data points to the interpolation, only few information on $y(x)$ is added to the surrogate $\widehat{y}_{\mathrm{HK}}(x)$ which is not already contained in the POD basis. Comparing HK based on plain and on aligned databases, it can be observed that alignment

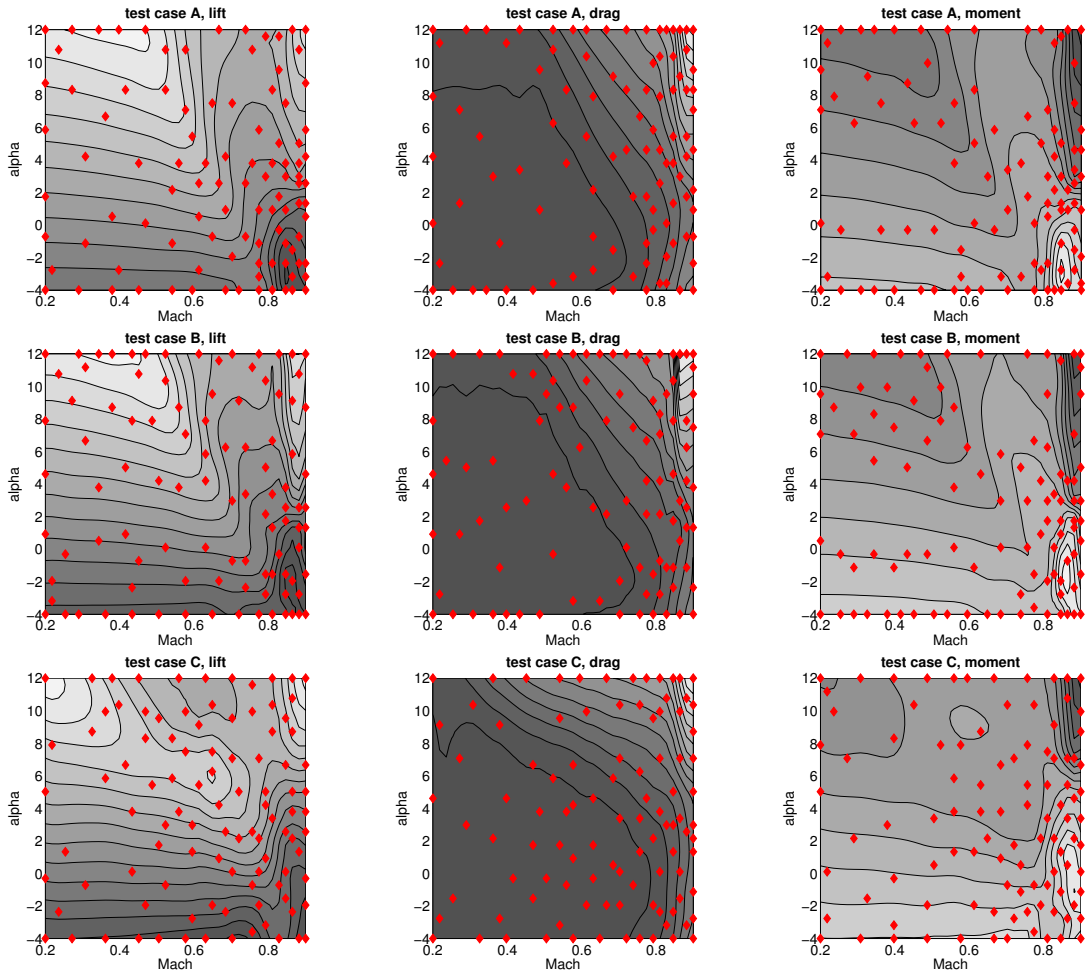

Figure 6.10: Sample distributions for strategy (6.2) for HK based on aligned databases.

is redundant in test case A. Using an aligned database instead of the plain one does not necessarily improve the approximation quality. In the drag case, it is even poorer. The reason is that with the alignment, additional degrees of (nonlinear) freedom are added to the surrogate modeling framework, while y actually is an element of the non-aligned $\mathcal{Y}^M$.

In test case B (Fig. 6.12), the new HK approaches are also able to outperform the common KRI interpolations. Only the GEK interpolations (which contain triple the amount of online information) are more accurate for large N when using adaptive sampling strategies. As already observed in test case A, HK interpolations are accurate already for moderate designs, but increasing N does not improve the approximation quality significantly (e.g. $N > 40$). This indicates that y is indeed strongly related to $\mathcal{Y}^M$ and the global trend can be accurately reproduced by elements of $\mathcal{Y}^L$, even with only a small number of samples. However, y is not and element of the database as in test case A, but a new test case. So the aligned HK interpolations generally perform better than HK based on the plain databases. By using transformations in the gappy POD fitting, the generic surrogate model can be adapted to the

data Y far better than by using merely a linear fit and the global behavior is approximated even more precisely. Nevertheless, if the design is too small (e.g. $N < 20$ samples in adaptive designs), HK based on aligned databases can produce inaccurate results. This is due to the gappy POD problem (5.51) having too many degrees of freedom ($= L + 5$) compared to the number of conditions ($= N$). It results in the few evaluations Y being approximated very closely by the GSM at the cost of an unfavorable behavior in the rest of the domain $\Omega \setminus X$, also known as overfitting (cf. Sect. 2.1.3).

In test case C (Fig. 6.13), the results are a little different. The common KRI surrogates can only be outperformed by HK for moderate sample sizes ($N < 30$–40). For larger sample sizes, the accuracy of HK based on aligned databases is still comparable to the KRI performance, while the HK based on plain databases is less accurate in most cases. GEK is generally most accurate (using triple the amount of online information). Especially when applying adaptive sampling, it performs significantly better than all other three surrogate modeling strategies. Contrary to test cases A and B, the responses $y \notin \mathcal{Y}^M$ cannot be reproduced adequately with the POD basis. The global trend is not approximated as exact as in the previous test cases, such that HK only has minor benefits compared to KRI (e.g. for small N). Using an aligned database reduces the error significantly compared to HK with plain databases. In this case, the additional degrees of freedom in the gappy POD fitting (5.51) to the data Y help to improve the global accuracy. Also, contrary to the previous test cases, the approximation quality of HK interpolations is improved greatly with growing sample size. The global behavior is not captured merely by the POD basis, but also by the evaluations Y, such that more samples add viable information to the surrogate $\widehat{y}_{\mathrm{HK}}(x)$.

The advantage of hierarchical Kriging over common Kriging interpolations is exemplarily depicted in Fig. 6.14 for test case B. Both Kriging and hierachical Kriging based on aligned databases are shown interpolating a Halton design of moderate sample size $N = 25$. The Kriging interpolations deviate from the validation data e.g. in the upper middle because not enough samples are concentrated there. On the right boundary, they completely fail to approximate the correct response. For the GSM-based HK interpolations, on the contrary, the shape of the responses is reproduced exactly. Even in a region with no samples in proximity like the upper right corner, the global trend is accurately captured by the database information contained in the GSM.

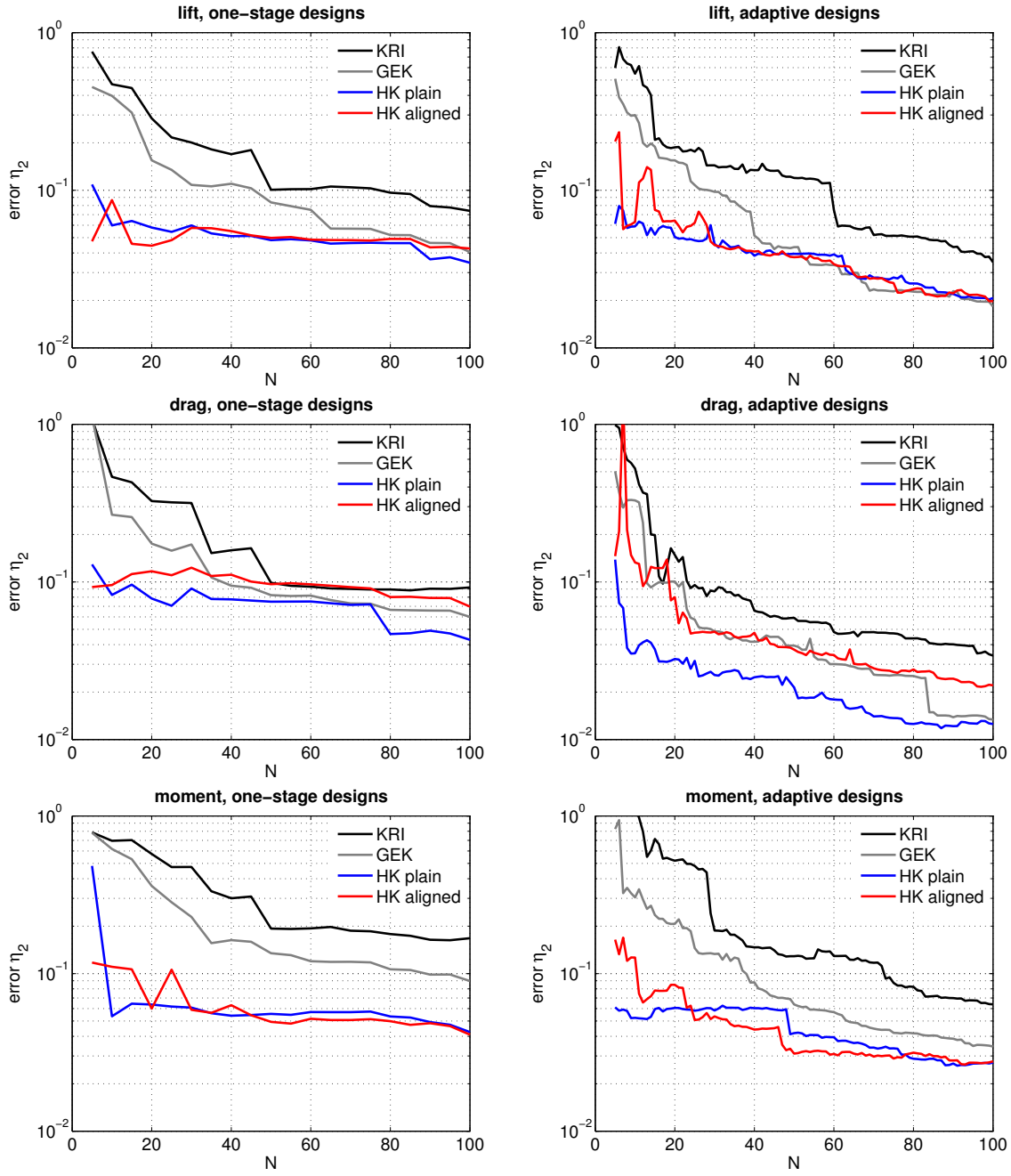

Figure 6.11: Root mean squared errors η_2 for surrogate models in test case A.

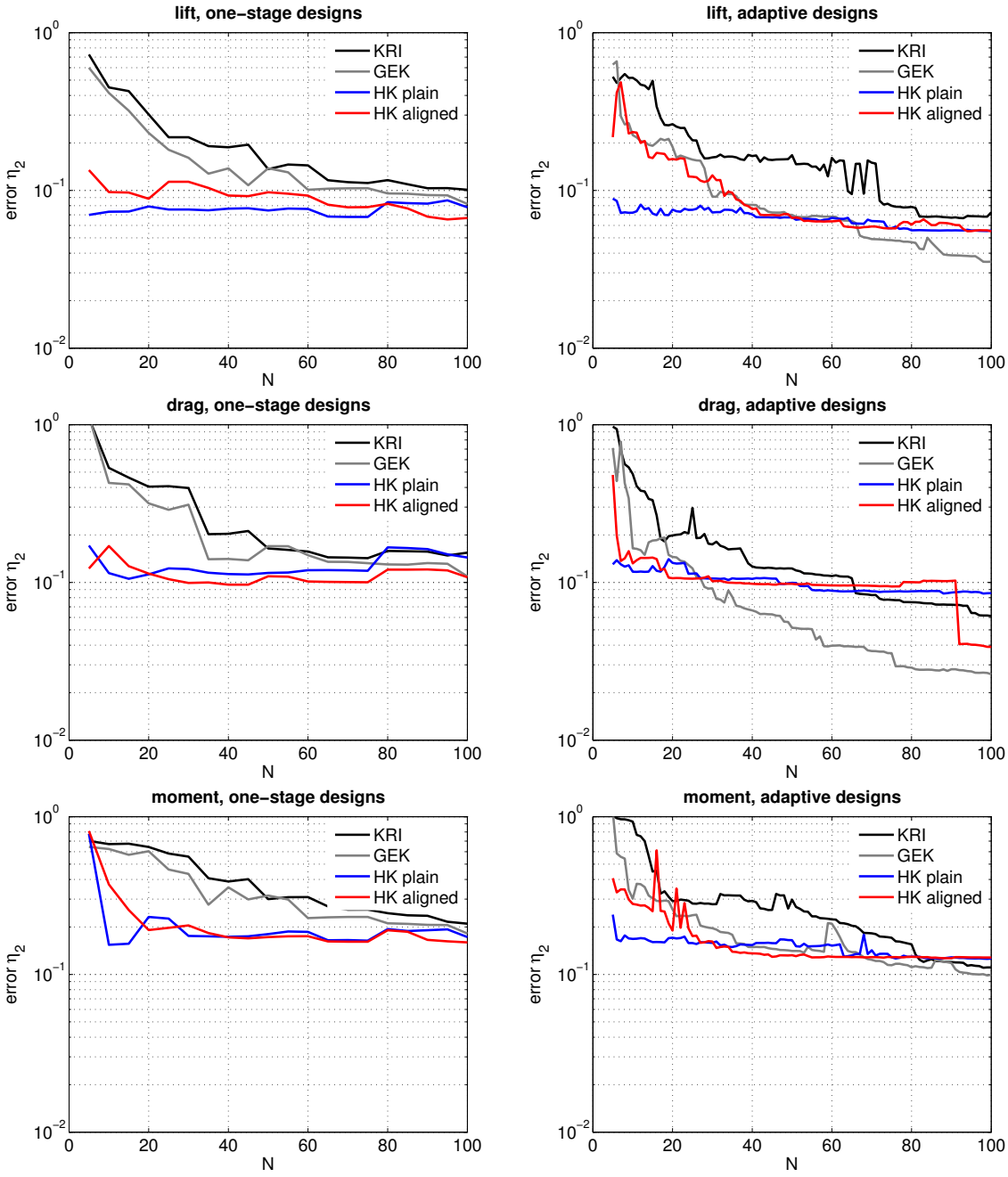

Figure 6.12: Root mean squared errors η_2 for surrogate models in test case B.

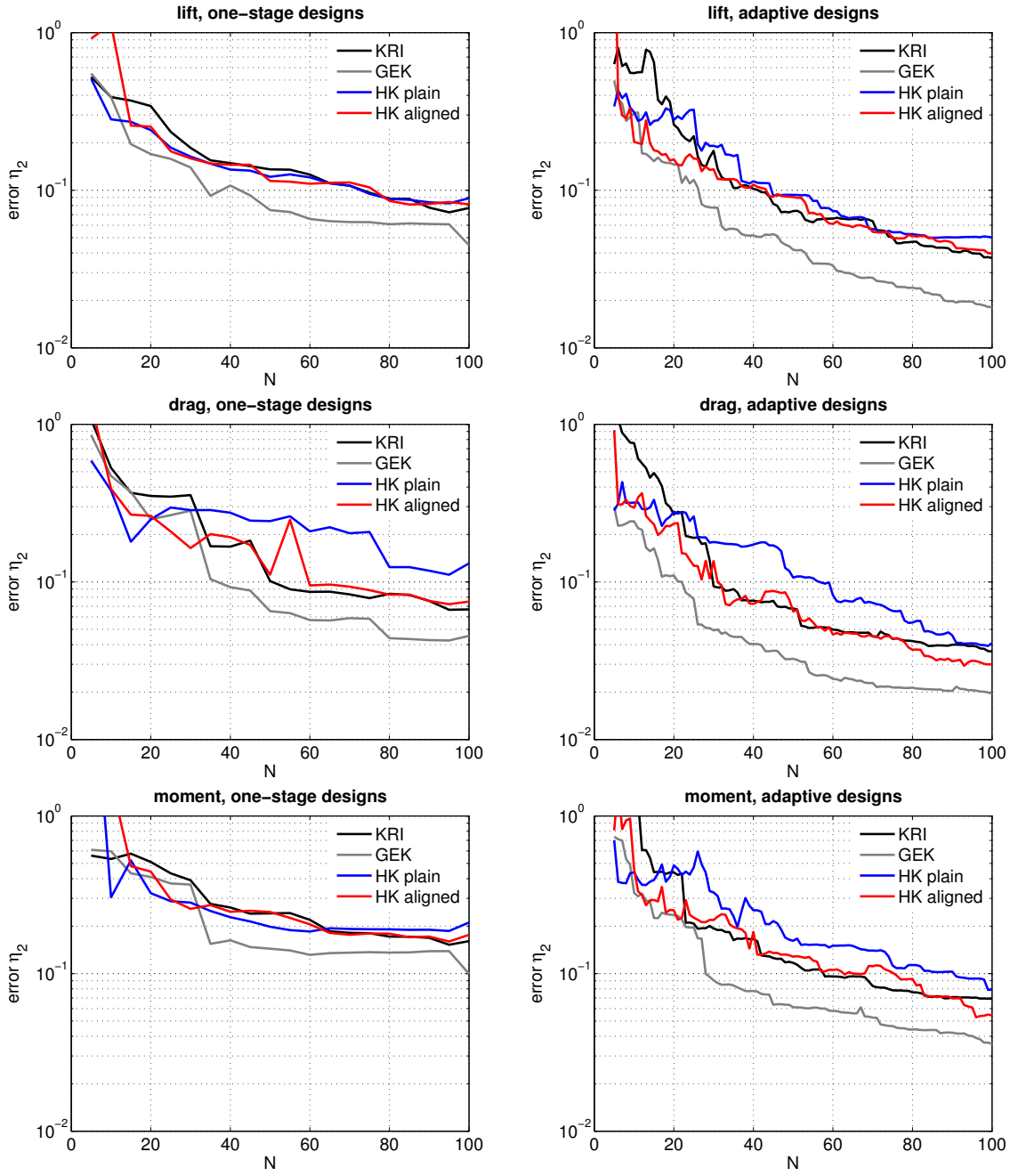

Figure 6.13: Root mean squared errors η_2 for surrogate models in test case C.

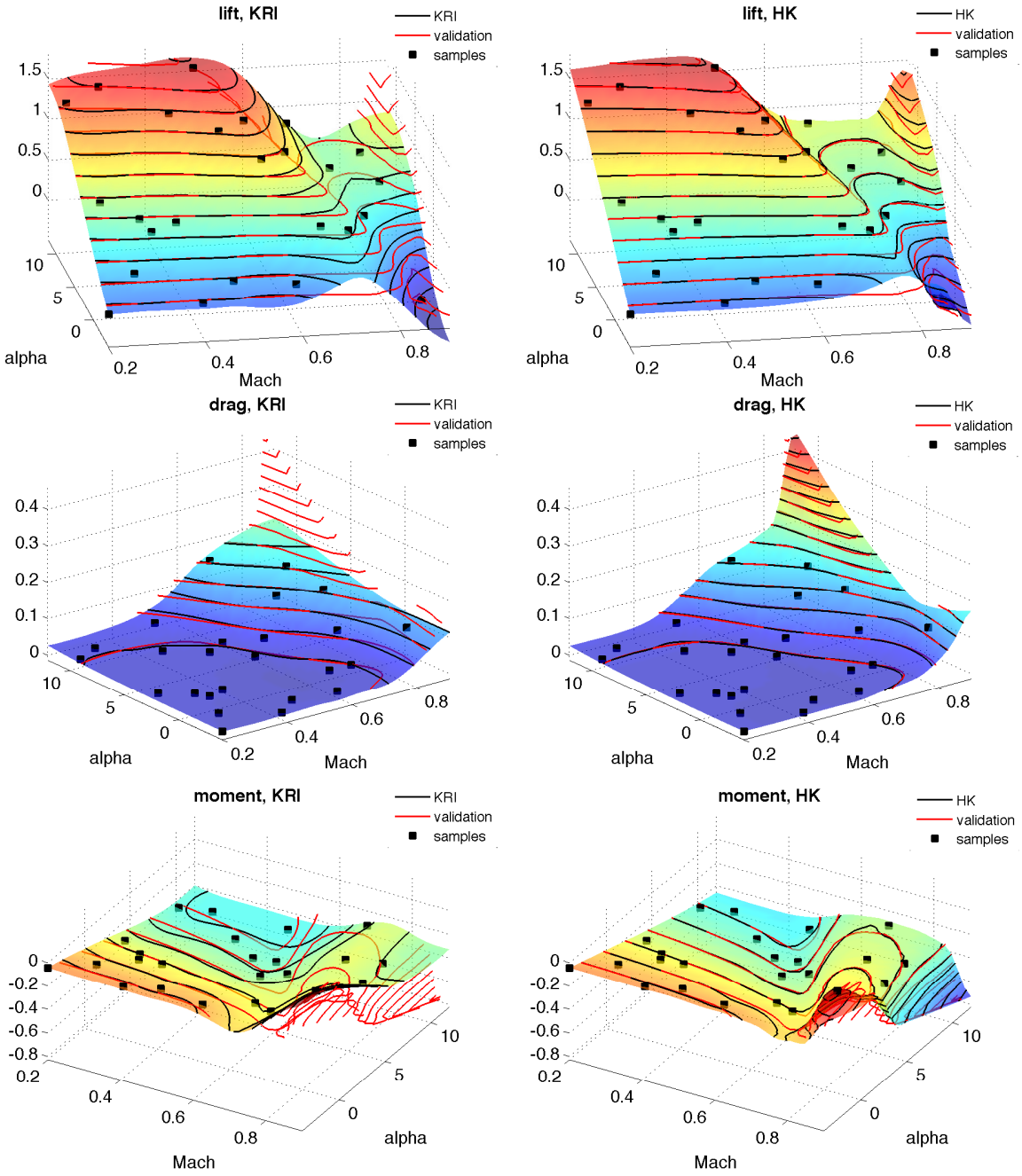

Figure 6.14: Kriging (left) and hierarchical Kriging (right) surrogate models for test case B based on a Halton design, $N = 25$. Red: validation data.

6.4 Assessment of the GSM approximation quality

As discussed in the previous section, the performance of the hierarchical Kriging interpolations apparently depends on the test case. The higher the relation of the unknown response $y(x)$ and the space $\mathcal{Y}^M$ spanned by the database functions, the better GSM can approximate y's global trend and the more accurate HK can interpolate y. The amount of this relation, i.e. the approximation quality of the GSM $\widetilde{f} \in \mathcal{Y}^L$ for y has to be quantified. This information is needed online, in other words it has to be possible to decide whether GSM-based HK is an appropriate choice of surrogate modeling for a certain test case during the process of model generation. Therefore, only $Y = (y(x^{(1)}), \ldots, y(x^{(N)}))^\top$ can be used for the assessment without validation data. The quantity

$$\widetilde{\eta}_2 := \left(\frac{\sum_{i=1}^{N} \left(y(x^{(i)}) - \widetilde{f}(x^{(i)}) \right)^2}{\sum_{i=1}^{N} \left(y(x^{(i)}) - \bar{y} \right)^2} \right)^{\frac{1}{2}} \tag{6.3}$$

is defined as an error indicator for the *goodness of fit*, where the mean $\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y(x^{(i)})$ of the entries of $Y \in \mathbb{R}^N$ is defined as usual. (6.3) is the ratio of a root mean squared error of the GSM $\widetilde{f}$ to the standard deviation σ of Y's entries. The lower $\widetilde{\eta}_2$, the better the approximation. The quantity is closely connected to the *coefficient of determination* $R^2 := 1 - \widetilde{\eta}_2^2$ in regression modeling (cf. [101], Sect. 4.2). In linear regression for instance, partitioning of

$$\sum_{i=1}^{N} \left(y(x^{(i)}) - \bar{y} \right)^2 = \sum_{i=1}^{N} \left(\widetilde{f}(x^{(i)}) - y(x^{(i)}) \right)^2 + \sum_{i=1}^{N} \left(\widetilde{f}(x^{(i)}) - \bar{y} \right)^2 \tag{6.4}$$

yields

$$R^2 = \frac{\sum_{i=1}^{N} \left(\widetilde{f}(x^{(i)}) - \bar{y} \right)^2}{\sum_{i=1}^{N} \left(y(x^{(i)}) - \bar{y} \right)^2}, \tag{6.5}$$

so R^2 describes the ratio of explained variance to the total variance.

The error indicator $\widetilde{\eta}_2$ is depicted for all HK interpolations in Figs. 6.15–6.17. It can be observed that the goodness of fit is generally best in test case A, second best in test case B and worst in test case C. In test case C for instance, the root mean squared approximation error of the GSM is larger than 10% of the standard deviation σ ($\widetilde{\eta}_2 > 0.1$) for $N > 20$. Comparing all error levels, $\widetilde{\eta}_2$ indeed turns out to be a measure for the practicability of GSM for the particular test cases. Generally speaking, the GSM-based HK interpolation are able to outperform common Kriging interpolations significantly when $\widetilde{\eta}_2 < 0.1$ is satisfied, cf. Figs. 6.11–6.13. Furthermore, the values converge to a particular limit with growing N for each surrogate in each test case ($\widetilde{\eta}_2 \to \widetilde{\eta}_2^*$ $(N \to \infty)$). The limit $\widetilde{\eta}_2^*$ can be interpreted as an overall relative root mean squared error (in relation to the standard deviation) for the approximation of y with a GSM $\widetilde{f} \in \mathcal{Y}^L$. Note that even in test case A, where $y \in \mathcal{Y}^M$ holds, the error does not tend to zero with growing sample size. By truncating the POD basis and using only $L < M$ basis elements, an approximation error is produced, cf. (5.35). For small

N however, the error indicator has no explanatory power. In all cases, $\widetilde{\eta}_2$ is close to zero while the actual error η_2 is large, when N is small compared to the degrees of freedom in the generic surrogate model ($= L$ for plain, $= L + 5$ for aligned databases). The few evaluations $y(x^{(i)})$ $(i = 1, \ldots, N)$ are approximated closely by $\widetilde{f}$ (causing a small $\widetilde{\eta}_2$), while in $\Omega \setminus X$ overfitting can be expected (causing a large η_2), cf. Sect. 2.1.3.

Summing up, with $\widetilde{\eta}_2$ a reliable error indicator for the goodness of the GSM fit $\widetilde{f}$ has been presented. After a certain starting phase for small N, it rapidly converges to an overall measure of the appropriateness of GSM for the function y. Error bounds can be defined in terms of standard deviations and $\widetilde{\eta}_2$ can monitor if the bounds are maintained, while only accessing online information Y without any validation data.

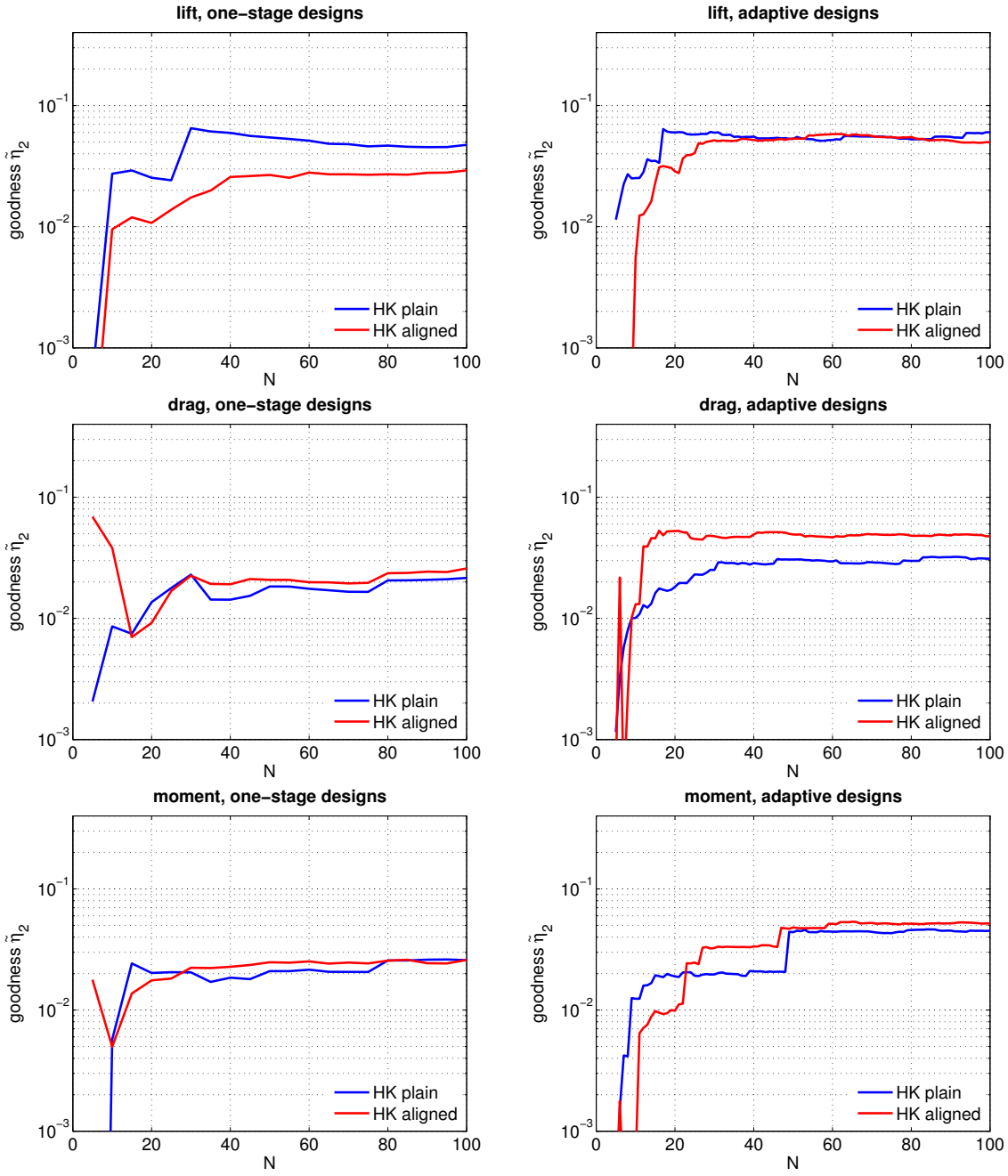

Figure 6.15: Goodness of fit $\widetilde{\eta}_2$ for GSM in test case A.

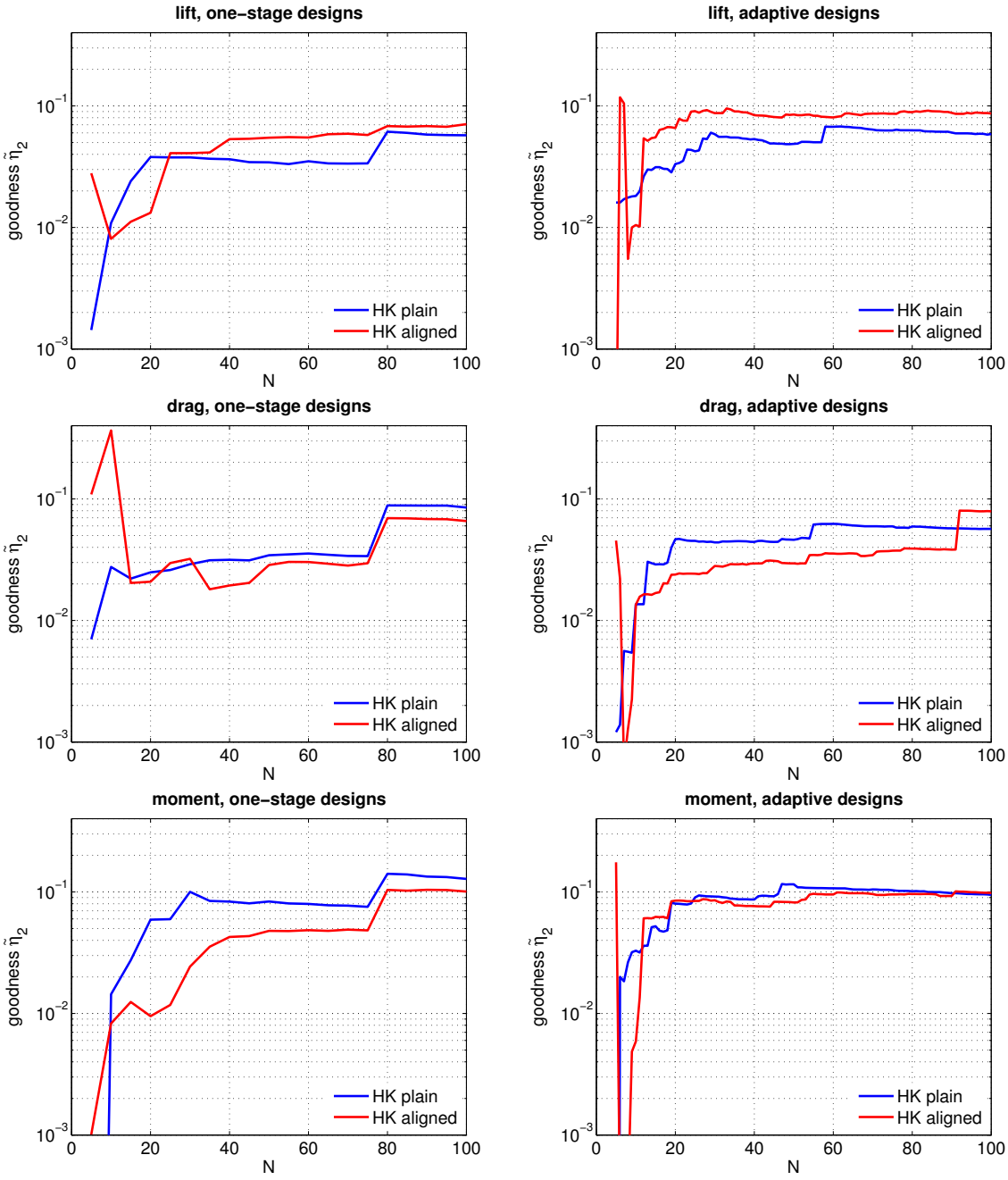

Figure 6.16: Goodness of fit $\widetilde{\eta}_2$ for GSM in test case B.

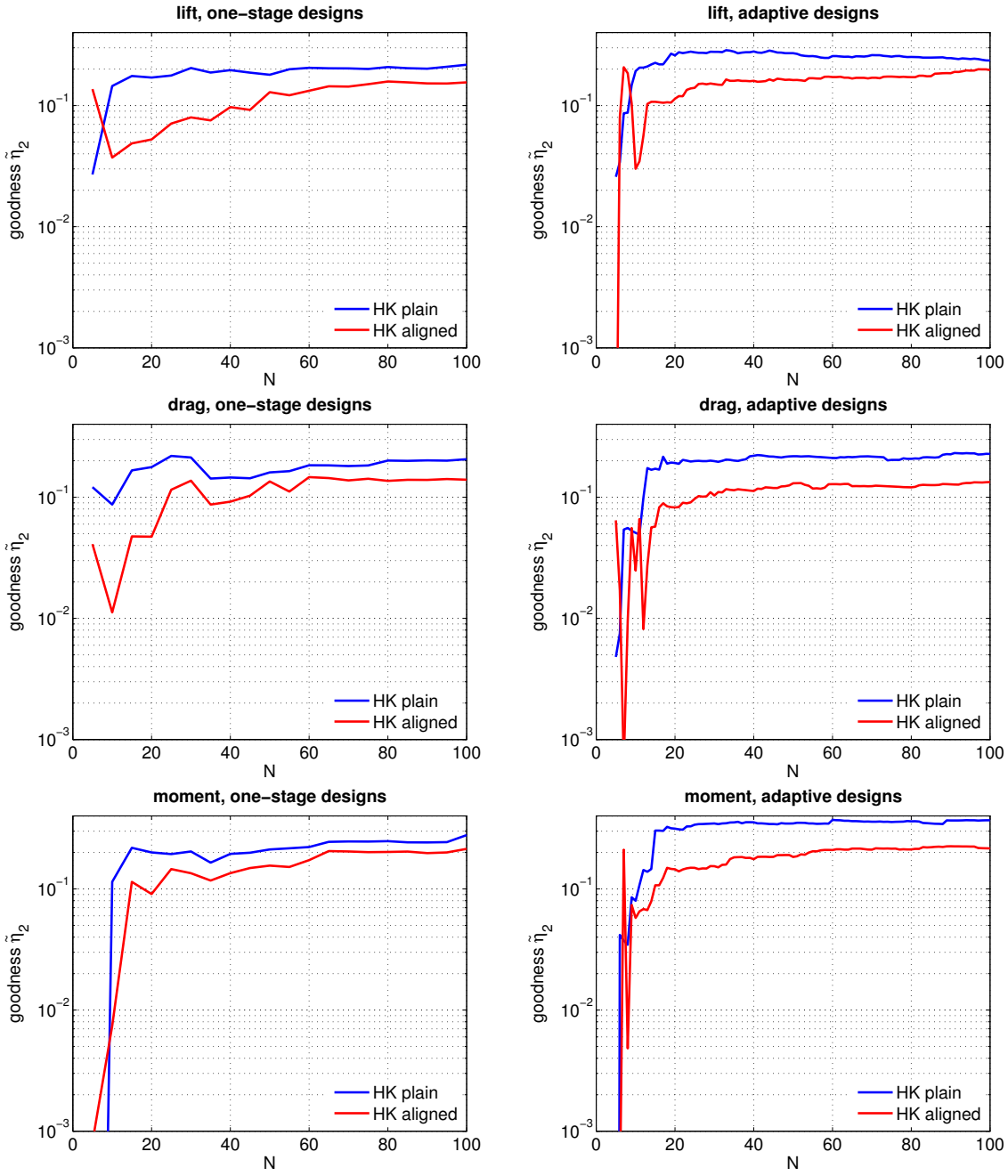

Figure 6.17: Goodness of fit $\widetilde{\eta}_2$ for GSM in test case C.

6.5 Summary

In this chapter, hierarchical Kriging based on generic surrogate models was validated with numerical test cases. It was demonstrated that the aerodynamic coefficient responses based on RANS solutions for two-dimensional airfoil geometries, depending on the Mach number and the angle of attack α, each constitute a distinct problem class. Inside these classes lift, drag and pitching moment, responses based on different airfoil geometries featured strong structural similarities. The mutual structures were extracted by proper orthogonal decomposition of the space spanned by a database of response functions. Prior alignment of the database elements reduced the total variation while characteristic features were preserved. In this way, the similarities could be identified even if their position in the input parameter domain Ω did not exactly coincide. The number of POD basis elements, which contain the most influential structures, was determined using an error formula for the truncated POD basis. In the investigated test cases, only three to five elements were sufficient for accurately describing the global behavior of all database functions.

The performance of hierarchical Kriging using the generic surrogate model as a global trend depended on the test case, i.e. on the airfoil geometry, for which a surrogate model of the aerodynamic coefficients is sought. Three distinct test cases were investigated. In test case A, the airfoil was already contained in the database, therefore it was merely of academic interest. A new airfoil not contained in the database was considered in test case B, which came from the same family of parametrization as the database airfoils. The airfoil in test case C was neither contained in the database nor did it come from the same parametrization family. It turned out that the more a response y was related to the space $\mathcal{Y}^M$ spanned by the database responses, the better the global trend could be approximated by the GSM and the more HK interpolations could outperform common Kriging. Also, the more y deviated from $\mathcal{Y}^M$, the more important the alignment became, because it introduced additional degrees of freedom to the gappy POD GSM fitting. While alignment was redundant in test case A, it was necessary in test case C for competing with common Kriging interpolations.

The GSM framework was proven to be a successful preprocessing for the hierarchical Kriging interpolations. In test cases A and B, already small sample sizes N were sufficient to reproduce the global trend. Not only the Kriging interpolations were outperformed, but also gradient-enhanced Kriging, which contained triple the amount of online information. In test case C on the contrary, Kriging could only be outperformed by HK for moderate sample sizes ($N < 30$–40). For larger designs however, it could at least compete. For assessing the quality of the GSM approximations, an error indicator $\widetilde{\eta}_2$ was introduced. An admissible error boundary could be defined in relation to the standard deviation σ of y. Both could be evaluated online using only the data Y. In the investigated test cases, GSM-based HK was able to outperform common Kriging significantly as long as the error indicator did not exceed a threshold of 10%. Care had to be taken regarding the degrees of freedom of the gappy POD approximation. If they were not significantly outnumbered by the samples, overfitting could be expected (involving a small $\widetilde{\eta}_2$ while actually η_2 was large).

Combined with the error indicator, the GSM framework was shown to be a reliable tool for improving the Kriging approximation quality. It was demonstrated that globally accurate surrogate models can be generated by reproducing the global trend of HK interpolations via the GSM. In this way, a large amount of online costs, dominated by expensive evaluations

of $y(x)$, were transferred to offline costs. Less evaluations were necessary for achieving a desired accuracy by using the information stored in the POD basis. For test cases with more than $d = 2$ input parameters, the framework is expected to be even more beneficial. Traditional surrogate models typically suffer from the curse of dimensionality. Even when using adaptive sampling strategies, the designs have to be at least rudimentary space-filling (exploration). For accurately describing the global behavior in higher dimensions, large N become unavoidable, yet computationally extremely demanding. While the offline costs of the GSM framework cannot escape the curse of dimensionality either (each database function is a surrogate model as well), its online costs for model generation are still negligible compared to the online costs of evaluating $y(x)$ numerously.

7 Conclusions

The aim of this thesis was the investigation of efficient and globally accurate surrogate models for expensive simulations. For the sake of efficiency, the number of costly evaluations has to be kept as small as possible without sacrificing accuracy. As test cases, aerodynamic coefficients of airfoil geometries, depending on the input parameters Mach number and angle of attack, were considered. The 2d Euler and RANS equations were solved numerically with the CFD codes FLOWer and TAU. Since the resulting responses are highly nonlinear, global surrogate modeling becomes challenging, even for a low-dimensional input parameter space. The main contributions of this work are as follows. First, profound reviews of the Kriging and the gradient-enhanced Kriging frameworks were given. Second, design of experiments strategies were investigated, compared, and numerically validated with the aerodynamic test cases. Third, a new approach was introduced, which uses information about the structural behavior, extracted from a previously computed database of responses, to improve the global approximation quality.

The background of Kriging and gradient-enhanced Kriging was investigated in detail. They are suitable for the interpolation of highly nonlinear, deterministic black-box functions such as the test cases. A special emphasis was put on the choice and the parameter estimation of the correlation function. For the investigated test cases, splines were used as ansatz functions for the interpolation, both in Kriging and in GEK. They were chosen because they met the requirements of being sufficiently smooth, having a local support, and producing well-conditioned positive definite correlation matrices. A gradient-based optimization routine was proposed to solve the maximum likelihood estimation for the hyperparameters. The exact computation of the hyperparameters was crucial for modeling the different sensitivities of the response with respect to each input parameter dimension correctly. If derivatives of the response are available, they can be used to prescribe the correct gradient of the surrogate model in the GEK framework and therefore to increase the accuracy of the approximating function. A comparison of Kriging and GEK regarding accuracy vs. effort revealed that GEK can only be used efficiently, if the gradient is computed independently of the number of input parameters, i.e. by adjoints or automatic differentiation.

Using a suitable design of experiments strategy could improve the approximation quality significantly for a given budget of samples. Among space-filling one-stage designs, the low-discrepancy Halton sequence could outperform other popular methods. Adaptive designs were reviewed regarding the concepts of exploration and exploitation. In Kriging and GEK, the MSE turned out to be a poor estimator of the true error, but a useful exploration component for adaptive selection criteria. Cross-validation, on the other hand, was identified as an appropriate exploitation component. Mixed exploration & exploitation strategies generally produced the most accurate surrogate models. Not only did they outperform the nonadaptive one-stage designs and a pure exploration-based adaptive strategy, they also produced more reliable surrogates than a theoretical pure exploitation method, which used

the true error based on validation data. They excelled in an efficient use of the costly samples by generating designs, which were more densely concentrated in automatically identified critical regions, while other regions of the input parameter domain were not neglected either. Regarding efficiency, two additional approaches were proposed. One of the two mixed strategies is suited for adding multiple samples per stage of the adaptive process, allowing the computation in parallel completely independent of each other. For the other one, an enhancement for problems with vector-valued responses was introduced. It uses a single, balanced criterion for all outputs, generating a mutual design.

A new surrogate modeling framework called generic surrogate modeling was introduced. It can be applied if a computer experiment is not treated as a single instance, but as an element of a larger problem class. It was demonstrated that the aerodynamic coefficients of different airfoil geometries constitute such a mutual problem class. Their responses of a CFD solver share structural similarities over the whole input parameter domain. The most important structures were extracted by proper orthogonal decomposition. For new test cases, i.e. for responses of new airfoil geometries, this information was used to predict the global trend in a variable-fidelity modeling framework. The new responses were interpolated accurately using only few expensive evaluations. The approximation quality was generally higher than in ordinary Kriging. A significant benefit was observed already for small sample sizes, where the information extracted from the database was sufficient to reproduce the global behavior. The higher the relation of the response and the training database was, the better the new framework succeeded. For test cases closely related to the database, even gradient-enhanced Kriging could be outperformed. To quantify the relationship of the new response and the database, an error indicator was introduced, which is evaluated using only the actual samples without validation data. While traditional methods required a large amount of online costs for resolving the nonlinear global behavior, this new approach transferred these costs to offline costs. For a predefined problem class, the database of surrogates for different responses had only to be computed and stored once, and this offline information could be used in each new test case.

A topic which lies beyond the scope of this thesis is (local) error estimation for surrogate models. It was briefly demonstrated that the MSE, which was used in numerous applications in the past, is actually not suited as a local error measure for the approximation of nonlinear responses of deterministic computer experiments. Further investigation of cross-validation or mixed approaches is required. In order to keep the number of evaluations as small as possible, or even below a maximum budgeted number, reliable error estimation becomes necessary to satisfy a certain target accuracy.

Regarding global surrogate modeling, some questions were raised about the appropriateness of the stationarity assumption in the Kriging model, i.e. if a single correlation model should be used for the whole input parameter domain. In the presence of different levels of smoothness throughout this domain, automatically locating them could save even more unnecessary expensive evaluations e.g. in linear regions, in addition to adaptive sampling.

Additional research can also be performed on the new approach of generic surrogate modeling. Since it was successfully demonstrated that gradient information of the response improves the approximation quality in GEK, it can be expected to enhance the hierarchical Kriging approximation based on generic surrogate models as well. The GSM is treated as the Kriging regression term in the hierarchical framework, and as long as it is continuously

differentiable, GEK can also be applied in hierarchical Kriging. Also, the framework has to be validated for other problem classes, especially for test cases with more than two input parameters. With the capability of capturing the global trend using offline information, it can be assumed to outperform traditional surrogate modeling techniques even more, the higher the dimension of the input parameter space.

Bibliography

[1] J. D. Anderson. *Computational Fluid Dynamics: The Basics with Applications.* McGraw-Hill, 1995.

[2] J. D. Anderson. *Fundamentals of Aerodynamics.* McGraw-Hill, 5th edition, 2010.

[3] A. B. Antognini and M. Zagoraiou. Exact optimal designs for computer experiments via Kriging metamodelling. *Journal of Statistical Planning and Inference*, 140(9):2607–2617, 2010.

[4] P. Bajorski. *Statistics for Imaging, Optics, and Photonics.* Wiley, 2011.

[5] G. Berkooz, P. Holmes, and J. L. Lumley. The proper orthogonal decomposition in the analysis of turbulent flows. *Annual Review of Fluid Mechanics*, 25(1):539–575, 1993.

[6] Å. Björck. *Numerical Methods for Least Squares Problems.* SIAM, 1996.

[7] G. Box, W. Hunter, and J. Hunter. *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building.* Wiley, 1978.

[8] G. E. P. Box and N. R. Draper. *Empirical Model-Building and Response Surfaces.* Wiley, 1987.

[9] M. D. Buhmann. *Radial Basis Functions.* Cambridge University Press, 2003.

[10] T. Bui-Thanh, M. Damodaran, and K. Willcox. Aerodynamic data reconstruction and inverse design using proper orthogonal decomposition. *AIAA Journal*, 42(8):1505–1516, 2004.

[11] D. Busby, C. L. Farmer, and A. Iske. Hierarchical nonlinear approximation for experimental design and statistical data fitting. *SIAM Journal on Scientific Computing*, 29(1):49–69, 2007.

[12] K. J. Chang, R. T. Haftka, G. L. Giles, and P.-J. Kao. Sensitivity-based scaling for approximating structural response. *Journal of Aircraft*, 30(2):283–288, 1993.

[13] S. Chaturantabut and D. Sorensen. Nonlinear model reduction via discrete empirical interpolation. *SIAM Journal on Scientific Computing*, 32(5):2737–2764, 2010.

[14] S. Chaturantabut and D. Sorensen. A state space error estimate for POD-DEIM nonlinear model reduction. *SIAM Journal on Numerical Analysis*, 50(1):46–63, 2012.

[15] J. Chilès and P. Delfiner. *Geostatistics: Modeling Spatial Uncertainty.* Wiley, 2012.

[16] J. A. Christen and B. Sansó. Advances in the sequential design of computer experiments based on active learning. *Communications in Statistics - Theory and Methods*, 40(24):4467–4483, 2011.

[17] H.-S. Chung and J. J. Alonso. Using gradients to construct cokriging approximation models for high-dimensional design optimization problems. In *40th AIAA Aerospace Sciences Meeting and Exhibit, AIAA 2002–0317*, 2002.

[18] W. Cochran and G. Cox. *Experimental Designs*. Wiley, 1957.

[19] T. Cootes, A. Hill, C. Taylor, and J. Haslam. Use of active shape models for locating structures in medical images. *Image and Vision Computing*, 12(6):355–365, 1994.

[20] H. Cramér. On the theory of stationary random processes. *Annals of Mathematics*, 41(1):215–230, 1940.

[21] N. Cressie. The origins of kriging. *Mathematical Geology*, 22:239–252, 1990.

[22] N. Cressie. *Statistics for Spatial Data*. Wiley-Interscience, 1993.

[23] K. Crombecq. *Surrogate Modelling of Computer Experiments with Sequential Experimental Design*. PhD thesis, University of Antwerp, 2011.

[24] K. Crombecq, L. De Tommasi, D. Gorissen, and T. Dhaene. A novel sequential design strategy for global surrogate modeling. In *Proceedings of the 2009 Winter Simulation Conference (WSC)*, pages 731–742, 2009.

[25] K. Crombecq, D. Gorissen, D. Deschrijver, and T. Dhaene. A novel hybrid sequential design strategy for global surrogate modeling of computer experiments. *SIAM Journal on Scientific Computing*, 33(4):1948–1974, 2011.

[26] C. Currin, T. Mitchell, M. Morris, and D. Ylvisaker. Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *Journal of the American Statistical Association*, 86(416):953–963, 1991.

[27] R. H. Davies, C. J. Twining, and C. J. Taylor. *Statistical Models of Shape: Optimisation and Evaluation*. Springer, 2008.

[28] N. Draper and H. Smith. *Applied Regression Analysis*. Wiley-Interscience, 1998.

[29] I. L. Dryden and K. Mardia. *Statistical Shape Analysis*. J. Wiley, 1998.

[30] V. Dubourg. *Adaptive Surrogate Models for Reliability Analysis and Reliability-based Design Optimization*. PhD thesis, Université Blaise Pascal, Clermont-Ferrand, France, 2011.

[31] R. Dwight and Z.-H. Han. Efficient uncertainty quantification using gradient-enhanced Kriging. In *Proceedings of 11th AIAA Conference on Non-Deterministic Approaches, Palm Springs CA. AIAA-2009-2276*. AIAA, 2009.

[32] R. Everson and L. Sirovich. Karhunen–Loève procedure for gappy data. *Journal of the Optical Society of America A*, 12(8):1657–1664, 1995.

[33] K. Fang, R. Li, and A. Sudjianto. *Design and Modeling for Computer Experiments*. Computer science and data analysis series. Chapman & Hall/CRC, 2006.

[34] A. Farhang-Mehr and S. Azarm. Bayesian meta-modelling of engineering design simulations: a sequential approach with adaptation to irregularities in the response behaviour. *International Journal for Numerical Methods in Engineering*, 62(15):2104–2126, 2005.

[35] G. Fasshauer. *Meshfree Approximation Methods With MATLAB*. Interdisciplinary mathematical sciences. World Scientific, 2007.

[36] V. Fedorov. *Theory of Optimal Experiments*. Probability and mathematical statistics. Academic Press, 1972.

[37] J. Ferziger and M. Perić. *Computational Methods for Fluid Dynamics*. Springer, 2002.

[38] B. Fischer and J. Modersitzki. Ill-posed medicine: an introduction to image registration. *Inverse Problems*, 24:034008, 2008.

[39] A. I. J. Forrester and A. J. Keane. Recent advances in surrogate-based optimization. *Progress in Aerospace Sciences*, 45(1–3):50–79, 2009.

[40] A. I. J. Forrester, A. Sóbester, and A. J. Keane. Multi-fidelity optimization via surrogate modelling. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*, 463(2088):3251–3269, 2007.

[41] A. I. J. Forrester, A. Sóbester, and A. J. Keane. *Engineering Design via Surrogate Modelling: A Practical Guide*. Wiley, 2008.

[42] J. Fox. *Applied Regression Analysis, Linear Models, and Related Methods*. Sage Publications, 1997.

[43] A. Frangi, D. Rueckert, J. Schnabel, and W. Niessen. Automatic construction of multiple-object three-dimensional statistical shape models: Application to cardiac modeling. *IEEE Transactions on Medical Imaging*, 21(9):1151–1166, 2002.

[44] I. Gikhman, A. Skorokhod, and S. Kotz. *The Theory of Stochastic Processes*, volume 1. Springer, 2004.

[45] M. Giles and N. Pierce. Adjoint equations in CFD: duality, boundary conditions and solution behaviour. *AIAA Paper*, 97-1850, 1997.

[46] T. Goel. *Multiple Surrogates and Error Modeling in Optimization of Liquid Rocket Propulsion Components*. PhD thesis, University of Florida, 2007.

[47] T. Goel, R. Hafkta, and W. Shyy. Comparing error estimation measures for polynomial and kriging approximation of noise-free functions. *Structural and Multidisciplinary Optimization*, 38:429–442, 2009.

[48] R. B. Gramacy and H. K. H. Lee. Bayesian treed gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, 103(483):1119–1130, 2008.

[49] R. B. Gramacy and H. K. H. Lee. Adaptive design and analysis of supercomputer experiments. *Technometrics*, 51(2):130–145, 2009.

[50] A. Griewank. On automatic differentiation. *Mathematical Programming: Recent Developments and Applications*, pages 83–108, 1989.

[51] H. Gunes, S. Sirisup, and G. E. Karniadakis. Gappy data: To krig or not to krig? *Journal of Computational Physics*, 212:358–382, 2006.

[52] Z. Han, S. Görtz, and R. Hain. A variable-fidelity modeling method for aero-loads prediction. In A. Dillmann, G. Heller, M. Klaas, H.-P. Kreplin, W. Nitsche, and W. Schröder, editors, *New Results in Numerical and Experimental Fluid Mechanics VII*, volume 112 of *Notes on Numerical Fluid Mechanics and Multidisciplinary Design*, pages 17–25. Springer Berlin / Heidelberg, 2010.

[53] Z.-H. Han and S. Görtz. Hierarchical kriging model for variable-fidelity surrogate modeling. *AIAA Journal*, 50(9):1885–1896, 2012.

[54] Z.-H. Han, S. Görtz, and R. Zimmermann. Improving variable-fidelity surrogate modeling via gradient-enhanced kriging and a generalized hybrid bridge function. *Aerospace Science and Technology*, 25(1):177–189, 2012.

[55] Z.-H. Han, R. Zimmermann, and S. Görtz. Alternative cokriging model for variable-fidelity surrogate modeling. *AIAA Journal*, 50(5):1205–1210, 2012.

[56] Z.-H. Han, R. Zimmermann, and S. Görtz. A new cokriging method for variable-fidelity surrogate modeling of aerodynamic data. In *48th AIAA Aerospace Sciences Meeting Including the New Horizons Forum and Aerospace Exposition, Orlando, Florida*, 4-7 Jan 2010.

[57] C. Huang, Y. Yao, N. Cressie, and T. Hsing. Multivariate intrinsic random functions for cokriging. *Mathematical Geosciences*, 41(8):887–904, 2009.

[58] D. Huang. *Experimental Planning and Sequential Kriging Optimization Using Variable Fidelity Data*. PhD thesis, The Ohio State University, 2005.

[59] S. D. Iaco, D. Myers, and D. Posa. On strict positive definiteness of product and product-sum covariance models. *Journal of Statistical Planning and Inference*, 141(3):1132–1140, 2011.

[60] B. Iooss, L. Boussouf, V. Feuillard, and A. Marrel. Numerical studies of the metamodel fitting and validation processes. *International Journal of Advances in Systems and Measurements*, 3:11–21, 2010.

[61] E. Isaaks and R. Srivastava. *An Introduction to Applied Geostatistics*. Oxford University Press, 1989.

[62] A. Iske. *Multiresolution Methods in Scattered Data Modelling*. Springer, 2004.

[63] R. Jin, W. Chen, and A. Sudjianto. On sequential sampling for global metamodeling in engineering design. In *Proceedings of DETC'02 ASME 2002 Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Montreal, Canada*, 2002.

[64] R. Jin, W. Chen, and A. Sudjianto. An efficient algorithm for constructing optimal design of computer experiments. *Journal of Statistical Planning and Inference*, 134(1):268–287, 2005.

[65] M. Johnson, L. Moore, and D. Ylvisaker. Minimax and maximin distance designs. *Journal of statistical planning and inference*, 26(2):131–148, 1990.

[66] B. Jones and R. T. Johnson. Design and analysis for the gaussian process model. *Quality and Reliability Engineering International*, 25(5):515–524, 2009.

[67] D. R. Jones, M. Schonlau, and W. J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13:455–492, 1998.

[68] J. R. Kalagnanam and U. M. Diwekar. An efficient sampling technique for off-line quality control. *Technometrics*, 39(3):308–319, 1997.

[69] M. Kennedy and A. O'Hagan. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13, 2000.

[70] J. P. C. Kleijnen. *Design and Analysis of Simulation Experiments*. Springer, 2007.

[71] J. P. C. Kleijnen and W. C. M. van Beers. Application-driven sequential designs for simulation experiments: Kriging metamodelling. *The Journal of the Operational Research Society*, 55(8):876–883, 2004.

[72] L. Kocis and W. J. Whiten. Computational investigations of low-discrepancy sequences. *ACM Transactions on Mathematical Software (TOMS)*, 23(2):266–294, 1997.

[73] J. Koehler and A. Owen. Computer experiments. In S. Ghosh and C. Rao, editors, *Handbook of Statistics, 13: Design and Analysis of Experiments*, pages 261–308. North-Holland, 1996.

[74] D. G. Krige. A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, 52(6):119–139, 1951.

[75] N. Kroll and J. Fassbender, editors. *MEGAFLOW: Numerical Flow Simulation for Aircraft Design*. Springer, 2005.

[76] K. Kunisch and S. Volkwein. Control of the burgers equation by a reduced-order approach using proper orthogonal decomposition. *Journal of Optimization Theory and Applications*, 102(2):345–371, 1999.

[77] K. Kunisch and S. Volkwein. Galerkin proper orthogonal decomposition methods for a general equation in fluid dynamics. *SIAM Journal on Numerical Analysis*, 40(2):492–515, 2003.

[78] H. Lamecker, M. Seebaß, H.-C. Hege, and P. Deuflhard. A 3D statistical shape model of the pelvic bone for segmentation. In J. Fitzpatrick and M. Sonka, editors, *Proceedings of SPIE - Volume 5370 Medical Imaging 2004: Image Processing*, pages 1341–1351, 2004.

[79] J. Laurenceau and M. Meaux. Comparison of gradient and response surface based optimization frameworks using adjoint method. *AIAA Paper*, 2008-1889, 2008.

[80] J. Laurenceau, M. Meaux, M. Montagnac, and P. Sagaut. Comparison of gradient-based and gradient-enhanced response-surface-based optimizers. *AIAA Journal*, 48(5):981–994, 2010.

[81] J. Laurenceau and P. Sagaut. Building efficient response surfaces of aerodynamic functions with kriging and cokriging. *AIAA Journal*, 46(2):498–507, 2008.

[82] L. Leifsson and S. Koziel. Variable-fidelity aerodynamic shape optimization. In S. Koziel and X.-S. Yang, editors, *Computational Optimization, Methods and Algorithms*, volume 356 of *Studies in Computational Intelligence*, pages 179–210. Springer Berlin Heidelberg, 2011.

[83] S. Levy and D. Steinberg. Computer experiments: a review. *AStA Advances in Statistical Analysis*, 94:311–324, 2010.

[84] R. M. Lewis. Using sensitivity information in the construction of kriging models for design optimization. In *Proceedings of 7th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, St. Louis, 1998.

[85] Y. Lin. *An Efficient Robust Concept Exploration Method and Sequential Exploratory Experimental Design.* PhD thesis, Georgia Institute of Technology, 2004.

[86] W. Liu. *Development of Gradient-Enhanced Kriging Approximations for Multidisciplinary Design Optimization.* PhD thesis, University of Notre Dame, 2003.

[87] B. Lockwood and M. Anitescu. Gradient-enhanced universal kriging for uncertainty propagation. *Nuclear Science and Engineering*, 170(2):168–195, 2012.

[88] S. N. Lophaven, H. B. Nielsen, and J. Søndergaard. Aspects of the MATLAB toolbox DACE. Technical Report IMM-REP-2002-13, Technical University of Denmark, Copenhagen, 2002.

[89] S. N. Lophaven, H. B. Nielsen, and J. Søndergaard. DACE: A MATLAB kriging toolbox. Technical Report IMM-REP-2002-12, Technical University of Denmark, Copenhagen, 2002.

[90] A. Lovison and E. Rigoni. Extracting optimal datasets for metamodelling and perspectives for incremental samplings. *Mathematics and Computers in Simulation*, 81(3):681–692, 2010.

[91] G. Marchuk, V. Agoshkov, and V. Shutyaev. *Adjoint Equations and Perturbation Algorithms in Nonlinear Problems*. CRC Press, 1996.

[92] K. V. Mardia and R. J. Marshall. Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71(1):135–146, 1984.

[93] J. Martin and T. Simpson. Use of adaptive metamodeling for design optimization. In *AIAA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*. AIAA, September 2002.

[94] G. Matheron. Principles of geostatistics. *Economic Geology*, 58(8):1246–1266, 1963.

[95] M. D. McKay, R. J. Beckman, and W. J. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2):239–245, 1979.

[96] M. Meckesheimer, A. J. Booker, R. R. Barton, and T. W. Simpson. Computationally inexpensive metamodel assessment strategies. *AIAA Journal*, 40:2053–2060, 2002.

[97] J. Melissen. *Packing and Covering with Circles*. PhD thesis, Utrecht University, 1997.

[98] M. D. Morris and T. J. Mitchell. Exploratory designs for computational experiments. *Journal of Statistical Planning and Inference*, 43(3):381–402, 1995.

[99] M. D. Morris, T. J. Mitchell, and D. Ylvisaker. Bayesian design and analysis of computer experiments: Use of derivatives in surface prediction. *Technometrics*, 35(3):243–255, 1993.

[100] D. E. Myers. Spatial interpolation: an overview. *Geoderma*, 62(1–3):17–28, 1994.

[101] R. Myers and D. Montgomery. *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. Wiley series in probability and statistics. Wiley, 1995.

[102] H. Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods*. CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, 1992.

[103] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, 2006.

[104] K. J. Nurmela and P. R. J. Östergård. Packing up to 50 equal circles in a square. *Discrete & Computational Geometry*, 18(1):111–120, 1997.

[105] K. J. Nurmela and P. R. J. Östergård. Covering a square with up to 30 equal circles. Research Report A62, Helsinki University of Technology, Laboratory for Theoretical Computer Science, Espoo, Finland, June 2000.

[106] M. Orr. Introduction to radial basis function networks. Technical report, Center for Cognitive Science, Scotland, 1996.

[107] I. G. Osio and C. H. Amon. An engineering design methodology with multistage bayesian surrogates and optimal sampling. *Research in Engineering Design*, 8:189–206, 1996.

[108] A. B. Owen. Orthogonal arrays for computer experiments, integration and visualization. *Statistica Sinica*, 2(2):439–452, 1992.

[109] E. Parzen. *Stochastic Processes*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 1999.

[110] V. Picheny, D. Ginsbourger, Y. Richet, and G. Caplin. Quantile-based optimization of noisy computer experiments with tunable precision. *Technometrics*, 55(1):2–13, 2013.

[111] N. Queipo, R. Haftka, W. Shyy, T. Goel, R. Vaidyanathan, and P. Kevin Tucker. Surrogate-based analysis and optimization. *Progress in Aerospace Sciences*, 41(1):1–28, 2005.

[112] C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[113] M. Reed and B. Simon. *Methods of Modern Mathematical Physics. I: Functional Analysis*. Academic Press, 1980.

[114] S. U. Rehman and A. Shapiro. An integral transform approach to cross-variograms modeling. *Computational Statistics & Data Analysis*, 22(3):213–233, 1996.

[115] G. Rennen. Subset selection from large datasets for kriging modeling. Discussion Paper 2008-26, Tilburg University, Center for Economic Research, 2008.

[116] T. Robinson, M. Eldred, K. Willcox, and R. Haimes. Strategies for multifidelity optimization with variable dimensional hierarchical models. In *Proceedings of the 47th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference (2nd AIAA Multidisciplinary Design Optimization Specialist Conference), Newport, RI*, 2006.

[117] B. Rosenbaum and V. Schulz. Comparing sampling strategies for aerodynamic Kriging surrogate models. *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik*, 92(11-12):852–868, 2012.

[118] B. Rosenbaum and V. Schulz. Efficient response surface methods based on generic surrogate models. *SIAM Journal on Scientific Computing*, 35(2):B529–B550, 2013.

[119] R. Rubinstein and D. Kroese. *Simulation and the Monte Carlo Method*. Wiley, 2008.

[120] E. Sachs and S. Volkwein. POD-Galerkin approximations in PDE-constrained optimization. *GAMM Reports*, 33(2):194–208, 2010.

[121] J. Sacks, S. B. Schiller, and W. J. Welch. Designs for computer experiments. *Technometrics*, 31(1):41–47, 1989.

[122] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn. Design and analysis of computer experiments. *Statistical Science*, 4(4):409–423, 1989.

[123] T. J. Santner, B. J. Williams, and W. Notz. *The Design and Analysis of Computer Experiments*. Springer, 2003.

[124] M. J. Sasena. *Flexibility and Efficiency Enhancements for Constrained Global Design Optimization with Kriging Approximations*. PhD thesis, University of Michigan, 2002.

[125] M. J. Sasena, P. Papalambros, and P. Goovaerts. Exploration of metamodeling sampling criteria for constrained global optimization. *Engineering Optimization*, 34:263–278, 2002.

[126] O. Schabenberger and C. Gotway. *Statistical Methods for Spatial Data Analysis*. CRC Press, 2005.

[127] D. Schwamborn, T. Gerhold, and R. Heinrich. The DLR TAU-code: Recent applications in research and industry. In P. Wesseling, E. Oñate, and J. Périaux, editors, *European conference on computational fluid dynamics, ECCOMAS CFD*, 2006.

[128] H. Seim, D. Kainmueller, M. Heller, H. Lamecker, S. Zachow, and H.-C. Hege. Automatic segmentation of the pelvic bones from CT data based on a statistical shape model. In *Proceedings of the First Eurographics conference on Visual Computing for Biomedicine*, pages 93–100, 2008.

[129] S. Shan and G. Wang. Survey of modeling and optimization strategies to solve high-dimensional design problems with computationally-expensive black-box functions. *Structural and Multidisciplinary Optimization*, 41:219–241, 2010.

[130] M. C. Shewry and H. P. Wynn. Maximum entropy sampling. *Journal of Applied Statistics*, 14(2):165–170, 1987.

[131] T. Simpson, D. Lin, and W. Chen. Sampling strategies for computer experiments: design and analysis. *International Journal of Reliability and Safety (IJRS)*, 2(3):209–240, 2001.

[132] T. Simpson, J. Poplinski, P. N. Koch, and J. Allen. Metamodels for computer-based engineering design: Survey and recommendations. *Engineering with Computers*, 17:129–150, 2001.

[133] L. Sirovich. Turbulence and the dynamics of coherent structures. Part I: Coherent structures. *Quarterly of Applied Mathematics*, 45(3):561–571, 1987.

[134] M. L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, 1999.

[135] R. Szeliski. Image alignment and stitching: A tutorial. *Foundations and Trends in Computer Graphics and Vision*, 2(1):1–104, 2006.

[136] B. Tang. Orthogonal array-based latin hypercubes. *Journal of the American Statistical Association*, 88(424):1392–1397, 1993.

[137] C. Tang, K. Gee, and S. Lawrence. Generation of aerodynamic data using a design of experiment and data fusion approach. In *43rd AIAA Aerospace Sciences meeting, Reno, Nevada*, 2005.

[138] S. Thompson. *Sampling*. Wiley, 1992.

[139] D. Toal, N. Bressloff, and A. Keane. Kriging hyperparameter tuning strategies. *AIAA Journal*, 46(5):1240–1252, 2008.

[140] D. J. J. Toal, A. I. J. Forrester, N. W. Bressloff, A. J. Keane, and C. Holden. An adjoint for likelihood maximization. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*, 465(2111):3267–3287, 2009.

[141] A. Vendl and H. Faßbender. Proper orthogonal decomposition for steady aerodynamic applications. *PAMM*, 10(1):635–636, 2010.

[142] J. M. Ver Hoef, N. Cressie, and R. P. Barry. Flexible spatial models for Kriging and Cokriging using moving averages and the fast fourier transform (FFT). *Journal of Computational & Graphical Statistics*, 13(2):265–282, 2004.

[143] F. A. C. Viana and R. T. Haftka. Cross validation can estimate how well prediction variance correlates with error. *AIAA Journal*, 47(9):2266–2270, 2009.

[144] F. A. C. Viana, G. Venter, and V. Balabanov. An algorithm for fast optimal latin hypercube design of experiments. *International Journal for Numerical Methods in Engineering*, 82(2):135–156, 2010.

[145] S. Volkwein. Optimal control of a phase-field model using proper orthogonal decomposition. *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik*, 81(2):83–97, 2001.

[146] H. Wackernagel. *Multivariate Geostatistics: An Introduction with Applications*. Springer Verlag, 2003.

[147] H. Wendland. *Scattered Data Approximation*. Cambridge University Press, 2005.

[148] Y. Xiong, W. Chen, D. Apley, and X. Ding. A non-stationary covariance-based kriging method for metamodelling in engineering design. *International Journal for Numerical Methods in Engineering*, 71(6):733–756, 2007.

[149] A. Yaglom. *An Introduction to the Theory of Stationary Random Functions*. Dover Publications, 2004.

[150] W. Yamazaki, M. Rumpfkeil, and D. Mavriplis. Design optimization utilizing gradient/hessian enhanced surrogate model. *AIAA Paper*, 2010-4363, 2010.

[151] L. Zhao. *Reliability-based Design Optimization Using Surrogate Model with Assessment of Confidence Level*. PhD thesis, University of Iowa, 2011.

[152] R. Zimmermann. Asymptotic behavior of the likelihood function of covariance matrices of spatial gaussian processes. *Journal of Applied Mathematics*, Article ID 494070, 2010.

[153] R. Zimmermann and S. Görtz. Non-linear reduced order models for steady aerodynamics. *Procedia Computer Science*, 1(1):165–174, 2010.

[154] B. Zitová and J. Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21(11):977–1000, 2003.