The Propagation-Separation Approach

Theoretical Study
and Application to Magnetic Resonance Imaging

DISSERTATION

zur Erlangung des akademischen Grades

doctor rerum naturalium (Dr. rer. nat.)
im Fach Mathematik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät II
Humboldt-Universität zu Berlin

von
Dipl.-Math. Saskia Myriam Alice Becker

Präsident der Humboldt-Universität zu Berlin
Prof. Dr. Jan-Hendrik Olbertz

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät II
Prof. Dr. Elmar Kulke

Gutachter:
1. Prof. Dr. Markus Reiß, Humboldt-Universitaät zu Berlin
2. Prof. Dr. Stephan Huckemann, Universität Göttingen
3. PD. Dr. Peter Mathé, WIAS Berlin/ Freie Universität Berlin

Eingereicht am: 05. Dezember 2013
Tag der Verteidigung: 08. Mai 2014

Abstract

In statistics, nonparametric estimation is often based on local parametric modeling. For pointwise estimation of the target function, the corresponding parametric neighborhoods can be described by weights that depend on design points or (additionally) on observations. As it turned out, the comparison of noisy observations at single points suffers from a lack of robustness. The Propagation-Separation Approach by Polzehl and Spokoiny [2006] overcomes this problem by using a multiscale approach with iteratively updated weights. This allows simultaneous estimation of the parametric neighborhoods and the associated parameters. The method has been successfully applied to a large variety of statistical problems. Nevertheless, only few properties are known. Here, we present a theoretical study and numerical results, which provide a better understanding of this versatile procedure.

For this purpose, we introduce and analyse a novel strategy for the choice of the crucial parameter of the algorithm, namely the adaptation bandwidth. In particular, we study its variability with respect to the unknown target function. This justifies a choice of the adaptation bandwidth by simulations, but independent of the data at hand. For piecewise constant and piecewise bounded functions, this choice enables theoretical proofs of the main heuristic properties of the algorithm, which are propagation under homogeneity and separation of distinct regions. Additionally, we consider the case of a misspecified model. Here, we introduce a specific step function, and we establish a pointwise error bound between this function and the corresponding estimates of the Propagation-Separation Approach.

Finally, we develop a method for the denoising of diffusion-weighted magnetic resonance data, which is based on the Propagation-Separation Approach. Our new procedure, called (ms)POAS, relies on a specific description of the data, which enables simultaneous smoothing in the measured positions and with respect to the directions of the applied diffusion-weighting magnetic field gradients. We define and justify two distance functions on the combined measurement space $\mathbb{R}^3 \times \mathbb{S}^2$, where we follow a differential geometric approach. We demonstrate the capability of (ms)POAS on simulated and experimental data.

Zusammenfassung

Lokal parametrische Modelle werden in der mathematischen Statistik häufig im Kontext der nichtparametrischen Schätzung verwendet. Bei einer punktweisen Schätzung der Zielfunktion können die zugehörigen parametrischen Umgebungen mithilfe von Gewichten beschrieben werden, die entweder von den Designpunkten oder (zusätzlich) von den Beobachtungen abhängen. Der Vergleich von verrauschten Beobachtungen in einzelnen Punkten leidet allerdings unter einem Mangel an Robustheit. Der Propagations-Separations-Ansatz von Polzehl und Spokoiny [2006] verwendet daher einen Multiskalen-Ansatz mit iterativ aktualisierten Gewichten. Das ermöglicht die gleichzeitige Schätzung der parametrischen Umgebungen und ihrer Parameter. Obwohl der Algorithmus auf eine Vielzahl statistischer Probleme erfolgreich angewendet wurde, sind nur wenige Eigenschaften bekannt. Deshalb präsentieren wir eine theoretische Studie und numerische Resultate, die ein besseres Verständnis des Verfahrens ermöglichen.

Zu diesem Zweck definieren und untersuchen wir eine neue Strategie für die Wahl des entscheidenden Parameters des Verfahrens, der Adaptationsbandweite. Insbesondere untersuchen wir ihre Variabilität in Abhängigkeit von der unbekannten Zielfunktion. Unsere Resultate rechtfertigen eine Wahl der Adaptationsbandweite anhand von simulierten Daten und damit unabhängig von den jeweils vorliegenden Beobachtungen. Die neue Parameterwahl liefert für stückweise konstante und stückweise beschränkte Funktionen theoretische Beweise der Haupteigenschaften des Algorithmus, ungehinderte Glättung unter Homogenität und Separation unterschiedlicher Regionen. Für den Fall eines falsch spezifizierten Modells führen wir eine spezielle Stufenfunktion ein und weisen eine punktweise Fehlerschranke im Vergleich zum Schätzer des Propagations-Separations-Algorithmus nach.

Des Weiteren entwickeln wir auf der Grundlage des Propagations-Separations-Ansatzes eine neue Methode zur Entrauschung von diffusionsgewichteten Magnetresonanzdaten. Unser neues Verfahren (ms)POAS basiert auf einer speziellen Beschreibung der Daten, die eine zeitgleiche Glättung bezüglich der gemessenen Positionen und der Richtungen der verwendeten Diffusiongradienten ermöglicht. Für den kombinierten Messraum $\mathbb{R}^3 \times \mathbb{S}^2$ schlagen wir zwei Distanzfunktionen vor, deren Eignung wir mithilfe eines differentialgeometrischen Ansatzes nachweisen. Schließlich demonstrieren wir das große Potential von (ms)POAS auf simulierten und experimentellen Daten.

Acknowledgment

There are many people who supported me during my PhD studies. First of all, I would like to thank Alexander Dittrich, my family, and my friends for their patience and encouragement. I am very grateful to my supervisor Prof. Dr. Markus Reiß for giving me the opportunity to carry out this research and for his mathematical guidance during the last years. I would like to express my sincere gratitude to my co-advisor Priv.-Doz. Dr. Peter Mathé for his constant interest in this project, the inspired discussions, and many useful advices. My colleagues at the WIAS Berlin contributed to a stimulating environment. In particular, I thank Dr. Karsten Tabelow and Dr. Jörg Polzehl for introducing me to the interesting world of MRI, our fruitful cooperation, and their technical help. I thank my co-authors Dr. Henning U. Voss (Weill Cornell Medical College, New York, USA), Dr. Alfred Anwander and Dr. Robin M. Heidemann (Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany), Dr. Nikolaus Weiskopf and Dr. Siawoosh Mohammadi (Wellcome Trust Centre for Neuroimaging, London, UK) for the permission to use their experimental diffusion-weighted MRI data. Ariane Beier and Dr. Rüdiger Thul proofread parts of this manuscript, and I would like to express my deepest gratitude for their comments. Finally, I would like to acknowledge financial support from the Stiftung der Deutschen Wirtschaft (SDW).

Contents

Chapter 1

Introduction

In statistics, regression analysis is one of the most commonly used approaches for parametric and nonparametric estimation, see, for instance, Simonoff [1996]. This technique aims to construct a function from a given data set $\{(X_i, Y_i)\}_i$ in order to describe the relationship between the explanatory variables X_i and the associated response variables Y_i. For example, in the Gaussian regression model, we observe the random variables $Y_i = f(X_i) + \epsilon_i$, $i \in \{1, ..., n\}$, where f is the target function and $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ denotes an independent and identically distributed Gaussian error term with variance $\sigma^2 > 0$. In this case, the value of the target function $f(X_i)$ equals the expectation of the random variable Y_i. Regression analysis is applicable to a large variety of scientific research and real-world questions.

In the literature, a large number of regression techniques have been proposed. Naturally, they have different strengths in their respective domains of application. As typical examples of parametric regression models, we mention the linear and the polynomial regression, which are usually solved by means of the ordinary least squares estimation. Here, the regression function is assumed to depend on a finite number of unknown parameters, which are estimated from the data. This leads to a global description which may be inappropriate for a satisfying data-fitting. In contrast, nonparametric regression relaxes this restrictive assumption by imposing as few assumptions as possible on the regression function. Usually, the corresponding statistical models are infinite-dimensional, or the dimensionality of the parameter space grows with the sample size.

For instance, we can introduce more flexibility by allowing the derivatives of the regression function to have discontinuities. The regression function can then be estimated by fitting piecewise polynomials, called *splines*. Alternatively, we can expand the regression function in an orthogonal series and compute an approximation, using an appropriate subset of its basis functions. This leads to the well-known *wavelet* approach. The classic approach is based on localization, where we want to concentrate on. The monograph by Fan and Gijbels [1996] gives a comprehensive overview of local polynomial estimation for regression and some other statistical models. We also note the monograph by Wand and Jones [1995]. For recent reviews of local modeling in image processing, see Buades et al. [2005a] and Katkovnik et al. [2010]. Due to the large number of publications concerning local modeling, we refer the reader to the references in these studies. They treat the main techniques as well as their historical development.

Local modeling

Instead of increasing the number of describing parameters, the *local modeling approach* assumes the parametric model to be satisfied locally. For each explanatory variable X_i, we apply the polynomial regression technique on a certain fraction of the data around X_i. Then, the value of the estimated regression function at X_i is used as the pointwise estimator at X_i. Globally, this leads to a nonparametric description of the target function. Obviously, the choice of the corresponding neighborhood of X_i is crucial for the estimation quality when using the local modeling approach.

A local model can be described by weights. For each explanatory variable X_i, we fix some weighting scheme W_i which assigns a weight $w(X_i, X_j)$ to every explanatory variable X_j. This determines the impact of the corresponding response variable Y_j on the pointwise estimator at X_i. For instance, in order to estimate the expectation of the random variable Y_i under the above introduced Gaussian regression model, we may apply a weighted mean $\hat{f}(X_i) := \sum_j w(X_i, X_j) Y_j$ with weights that satisfy $\sum_j w(X_i, X_j) = 1$. The simplest example of a weighting scheme $W_i = \{w(X_i, X_j)\}_j$ is localization by a window, where the weighting function equals the indicator function of a certain subset (window) of the design around X_i. Applications of this approach can be found in [Müller, 1992; Qiu, 1998; Spokoiny, 1998], among many others. Alternatively, the weights can be defined by a kernel function, where the corresponding bandwidth determines the amount of smoothing. Popular examples are the Nadaraya-Watson and the Gasser-Müller estimators [Fan and Gijbels, 1996, §2.2]. The bandwidth can be chosen either by the data analyst or by a data-driven procedure. Additionally, we distinguish between global choices of the bandwidth and local choices, which allow more flexible modeling. Some interesting proposals for a structure-adaptive bandwidth choice in the context of image processing are compared in the review by Katkovnik et al. [2010]. These are mostly linked to Lepski's method, which was introduced in [Lepskiĭ, 1990], but see also Lepski et al. [1997], Lepski and Spokoiny [1997], Mathé and Pereverzev [2006], and Spokoiny [1998].

The above weights depend on the explanatory variables $\{X_i\}_i$ only. An alternative approach for local modeling is based on weighting schemes which depend (additionally) on the response variables $\{Y_i\}_i$. This helps to avoid blurring at discontinuities. An overview of such algorithms can be found in [Katkovnik et al., 2010]. The authors distinguish between local and non-local procedures. They call a method local if it assigns small weights to spatially distant design points, which restricts the size of the estimation support. On the contrary, a non-local method allows large weights to distant points, resulting in a (possibly) disconnected estimation support. In a nutshell, the distinction between local and non-local methods relies on the question whether the explanatory or the response variables dominate the weights. The Propagation-Separation Approach by Polzehl and Spokoiny [2006] has an outstanding position in this classification. By using iteratively updated weights at increasing scales, it relates the local with the non-local approach.

Tibshirani and Hastie [1987] and later on Fan et al. [1998] and Loader [1999] proposed an extension of local modeling to the likelihood approach. We recall that the local exponential family model provides, under regularity conditions, an explicit representation of the maximum likelihood estimator. This considerably simplifies the theoretical analysis and the computations, while still including many of the usual probability distributions. Following Polzehl and Spokoiny [2006], we will restrict a large part of our study to the local exponential family model, and we will use, for our pointwise estimator, a local maximum likelihood estimator with specifically defined adaptive weights.

Previous results and objective of this thesis

In this thesis, we will study the Propagation-Separation Approach by Polzehl and Spokoiny [2006]. Due to its structural adaptivity, this method for nonparametric estimation avoids blurring at discontinuities. Moreover, it can be applied on any design space which is endowed with a metric, independent of the dimension or geometry of this space. Therefore, it is applicable to a large variety of problems. The method has been successfully applied in the context of image denoising [Li et al., 2011, 2012; Polzehl and Spokoiny, 2000, 2008; Tabelow et al., 2008], time series analysis [Divine et al., 2008], density estimation, and classification [Polzehl and Spokoiny, 2006], for example. Despite the practical use of this method, only few properties are known. The aim of this thesis is to provide a better understanding of the Propagation-Separation Approach, the involved parameters, its theoretical properties, and its behavior in practice. The corresponding results were published in [Becker, 2013] and [Becker and Mathé, 2013].

We know from the original study by Polzehl and Spokoiny [2006] that the Propagation-Separation Approach possesses the following theoretical properties.

- Propagation: In a homogeneous setting, the algorithm provides similar results as the corresponding non-adaptive estimator, which is optimal in this situation.
- Separation: The method avoids blurring at structural borders by separating distinct homogeneous regions.
- Stability: The procedure provides a certain stability of estimates.

We will revisit the above properties for the following two reasons.

First, the proof of the propagation property by Polzehl and Spokoiny [2006] is based on a doubtable assumption, called (S0). This assumption requires the statistical independence of the adaptive weights from the observations, which is problematic as discussed in [Polzehl and Spokoiny, 2006, Rem. 5.1]. Theoretically, the standard splitting technique could be used to ensure the required statistical independence. However, in practice, such a split is questionable due to the iterative approach of the algorithm. Here, we will present theoretical results, which do not require Assumption (S0).

Second, Polzehl and Spokoiny [2006] included an additional *memory step* into the algorithm. The authors argued that in some situations the adaptivity of the procedure may not suffice in order to avoid an increase of the estimation error after some iterations. Therefore, they included the memory step which is constructed to ensure a stability of estimates up to some constant. However, the corresponding constant can be large, which leads to a reduced stability. Additionally, the use of the memory step turned out to be questionable in practice. More precisely, to our best knowledge, no situation has been reported to date where the memory step considerably improved the results of the Propagation-Separation Approach. Hence, in many applications of the algorithm, the memory step was omitted, see for example Becker et al. [2012], Divine et al. [2008], Li et al. [2011, 2012], and Tabelow et al. [2008], still yielding the desired behavior.

We aim to justify the simplified Propagation-Separation Algorithm where the memory step is omitted. Our theoretical and numerical results will answer the question whether the memory step is needed and if, where. A general justification of the Propagation-Separation Approach itself and a comparison with other estimation methods is however beyond the scope of this study. Instead, we refer the reader to the previous articles by Polzehl and Spokoiny [2000, 2004, 2006, 2008].

In addition, we will develop a new method for the denoising of diffusion-weighted magnetic resonance data, which is based on the simplified Propagation-Separation Approach. Our proposed algorithm uses a specific perspective on the measured data which allows simultaneous smoothing in the voxel space $\mathbb{R}^3$ and on the sphere $\mathbb{S}^2$, where the diffusion-weighting magnetic field gradients are uniformly sampled. Additionally, we will improve adaptation by a coupling between measurements with different magnetic field strengths and diffusion time, called *q-shells*. As it turns out, this approach leads to a very promising and efficient method. The crucial steps are the search for and the justification of an appropriate distance function on the combined measurement space $\mathbb{R}^3 \times \mathbb{S}^2$. We refer the reader to the recent articles [Becker et al., 2012] and [Becker et al., 2013].

Research highlights

In practice, the original and the simplified Propagation-Separation Algorithm show a very similar behavior with obvious heuristic properties. Unfortunately, the theoretical verification of these properties is complicated due to the iterative approach. Polzehl and Spokoiny [2006] tackled this problem with the help of the questionable Assumption (S0) and the additionally included memory step. Here, we will follow a different approach which uses neither the memory step nor Assumption (S0). Instead, we will take advantage of what is known as the *propagation condition*. As proposed by Polzehl and Spokoiny [2006], such a condition can be used for the choice of the crucial parameter of the algorithm, namely the *adaptation bandwidth*. We will present a new formulation which enables our subsequent study and which improves the interpretability of the resulting bandwidth choice.

The new propagation condition allows the verification of the propagation property and a certain stability of estimates for (piecewise) constant parameter functions. The corresponding results are stated in Chapter 4. In particular, we will analyze, for the first time, the interaction of propagation and separation during iteration. All previous studies have concentrated on the respective properties on their own. Our results provide an insight in the reciprocal effects of the different components which influence the estimation quality.

Then, we will extend the presented results from piecewise constant to piecewise bounded parameter functions with sharp discontinuities, supposing the adaptation bandwidth to be in accordance with an inhomogeneous extension of the propagation condition. Moreover, we will study consequences of a misspecified structural assumption. We will introduce an *associated step function*, and we will explore its heuristic properties by numerical simulations in Chapter 5. In all examples, the Propagation-Separation Approach converges to the associated step function. This result could be interpreted as an intrinsic stopping criterion which provides a certain stability of estimates for arbitrary parameter functions. Unfortunately, we still lack a definite proof for reasons that we will discuss in Chapter 5 (§ 5.4.1). As a consequence, a stopping criterion with respect to the maximal number of iterations is needed to still ensure a general stability property. Particularly, in the case of a piecewise smooth function, where the formation of a step function may worsen the smoothing result, an appropriate stopping is advantageous.

Finally, we will generalize the setting of our study, first by relaxing the assumed local exponential family model, and second by considering a local likelihood model without additional restrictions. We will discuss the theoretical and numerical results for the simplified algorithm in Chapter 5 (Section 5.4). This includes a detailed comparison of the presented results with the original study by Polzehl and Spokoiny [2006], and some proposals for future research.

Our theoretical results and the corresponding numerical simulations in Chapter 5 justify the simplified Propagation-Separation Algorithm. The novel approach via the propagation condition provides a better understanding of the procedure by evaluating its behavior during iteration, the impact of the involved components, and the effects of different probability distributions. Nevertheless, we pay a price for the omittance of the memory step and the avoidance of Assumption (S0), which follows from the propagation condition. For our analysis, an appropriate choice of the adaptation bandwidth is crucial. In the above mentioned results, we will assume the adaptation bandwidth to be in accordance with our new propagation condition. However, the original as well as the new propagation condition rely on an artificial data set which satisfies a parametric model with some fixed parameter value. Hence, we need a justification that the adaptation bandwidth is in accordance with the propagation condition for the unknown parameter values which we aim to estimate.

For this purpose, we will evaluate in Chapter 3 (§ 3.1.2) the variability of the propagation condition with respect to the fixed parameter. This analysis is based on a sufficient criterion for the invariance of the composition of two functions with respect to some parameter. Then, our main result of this chapter will be stated in Theorem 3.8, where the non-adaptive estimator will be considered. An extension to the adaptive estimator is hampered by the iterative approach of the algorithm, which leads to an unknown probability distribution of the adaptive weights. Therefore, we will illustrate by simulations the close relation of the adaptive and the non-adaptive estimator under a satisfied propagation condition.

Together, Theorem 3.8 and the numerical simulations suggest the desired invariance of the propagation condition for several probability distributions, such as the Gaussian and exponential distributions and, as a consequence, the log-normal, Rayleigh, Weibull, and Pareto distributions. For probability distributions where the propagation condition varies with the fixed parameter, we recommend to choose the adaptation bandwidth with respect to some least favorable parameter. If this parameter is chosen appropriately, then the propagation condition remains valid for the unknown target parameters. However, the resulting adaptation bandwidth may be much larger than needed, leading to a loss of adaptation. To our knowledge, this is the first theoretical analysis of the propagation condition, forming an essential step in our analysis. In summary, we will replace the obviously violated Assumption (S0) by the more realistic assumption of an appropriate choice of the adaptation bandwidth.

In our last chapter, we will demonstrate the practical value of the Propagation-Separation Approach. Here, we will use the simplified algorithm for the denoising of diffusion-weighted magnetic resonance images. Basically, the Propagation-Separation Approach is applicable on any data set where the design and the observation space are endowed with possibly asymmetric distance functions. As it turns out, the smoothing results can be considerably improved by taking the specific properties of the data into account. Therefore, we aim to benefit from the whole information provided by the measurement process of magnetic resonance imaging, in position, orientation, and with respect to different magnetic field strengths and diffusion times. We will introduce a specific description of the measured data, and we will extend the adaptive weights, where we will reconstruct possibly missing data by spherical interpolation.

Our new method is theoretically justified via a natural embedding of diffusion-weighted data into the theory of *orientation scores* as proposed by Duits and Franken [2011] and Franken [2008]. It is well-known that the use of invertible transformations in image processing simplifies processing of a certain feature of interest in the transformed domain. For example, the Fourier transform concentrates on global frequencies, the Gabor transform relates to local frequencies, and wavelet transforms manipulate features at different scales. Similarly, the invertible orientation score

transform considers the local orientation as the feature of interest. Here, we will replace the design space $\mathbb{R}^3 \times \mathbb{S}^2$ by a Lie group, using a lifting of real valued functions on $\mathbb{R}^3 \times \mathbb{S}^2$ to real valued functions on the special Euclidean motion group $\mathrm{SE}(3)$ with a certain invariance property which ensures invertibility. Then, our main result will be stated in Theorem 6.38.

In order to make the procedure applicable in practice, we seek for a distance function on the design space. There are two natural approaches. First, the above embedding provides the Carnot-Carathéodory distance as a natural metric on $\mathrm{SE}(3)$, and we will establish its appropriateness for diffusion-weighted data. However, this metric does not have an explicit representation, and the proposed approximation violates a certain invariance. Therefore, we will introduce a second distance which is given as the sum of the Euclidean metric on $\mathbb{R}^3$ and the great circle distance on the sphere $\mathbb{S}^2$. This distance satisfies all desired properties.

The promising behavior of our smoothing method will be illustrated on simulated and experimental data. Finally, we will discuss its advantages and limitations (Section 6.6). In particular, we will review alternative perspectives on the data, give a brief overview of competing smoothing methods, and propose several topics for future research.

Notes to the reader

This thesis is written in such a manner that most chapters can be considered on their own. Possibly needed previous results are explicitly mentioned. Nevertheless, we recommend at least the reading of Section 2.2, where we introduce the original and the simplified Propagation-Separation Approach. The underlying statistical model and some basic concepts can be found in Section 2.1. Usually, we postpone longer proofs to the end of the respective chapter, in order to improve readability.

In Chapter 3, we will present our new propagation condition, its inhomogeneous extension, and the corresponding results concerning the invariance with respect to the unknown parameters and the application in practice. Here, we will use some auxiliary results by Polzehl and Spokoiny [2006] which are stated in Chapter 2. Then, in Chapter 4, we will deduce our theoretical study of the simplified Propagation-Separation Approach, assuming that the adaptation bandwidth has been chosen in accordance with the propagation condition. For the results on (piecewise) constant functions, it suffices to know the homogeneous propagation condition in Definition 3.2. Similarly, the results on (piecewise) bounded functions rely on the inhomogeneous propagation condition in Definition 3.15. Throughout this chapter, we will assume the local exponential family model (§ 2.1.2). Additionally, we will again use the auxiliary results in Chapter 2. We will illustrate several aspects of our theoretical study by numerical simulations, see Chapter 5. Although Chapter 4 helps for a better understanding of the examples and the subsequent discussion, the main points only require the knowledge of the original and the simplified algorithm.

Finally, Chapter 6 treats the application of the simplified Propagation-Separation Approach in the context of diffusion-weighted magnetic resonance imaging. This chapter can be considered on its own. Nevertheless, Chapters 2, 3, and 5 may help for a better understanding of the algorithm and the corresponding parameter choices. Most of the theoretical results in Chapter 4 do not apply since the assumed local exponential family model is not satisfied in the case of diffusion-weighted magnetic resonance data. We refer the reader to Section 4.4 and § 5.3.5 for more details about the Propagation-Separation Approach in this case.

Notation

We summarize some quantities that we will frequently use. For reasons of clarity and in order to avoid confusion, we concentrate on the most important ones.

General symbols

B, C	$B'(\theta) = \theta C'(\theta)$, see $p(y,\theta)$	10
$\mathcal{E}\tilde{\theta}_i^{(k)}$	$\mathcal{E}\tilde{\theta}_i^{(k)} = \sum_{j=1}^n \tilde{w}_{ij}^{(k)} \theta_j / \tilde{N}_i^{(k)}$	40
$\{h^{(k)}\}_{k=0}^{k^*}$	Increasing sequence of location bandwidths with $h^{(0)} > 0$	15
I	Fisher information, $I(\theta) = C'(\theta)$	11
$\mathcal{KL}$	Kullback-Leibler divergence, it holds under Ass. A1 that $\mathcal{KL}(\theta,\theta') = \theta\,[C(\theta) - C(\theta')] - [B(\theta) - B(\theta')]$	11, 95
$K_{\mathrm{loc}}, K_{\mathrm{ad}}, K_{\mathrm{me}}$	Location, adaptation, and memory kernels	15
k	Iteration step, $k \in \{0, ..., k^*\}$	16
$\overline{N}_i^{(k)}$	Sum of the non-adaptive weights, $\overline{N}_i^{(k)} = \sum_j \overline{w}_{ij}^{(k)}$	15
$\tilde{N}_i^{(k)}$	Sum of the adaptive weights, $\tilde{N}_i^{(k)} = \sum_j \tilde{w}_{ij}^{(k)}$	18
$\hat{N}_i^{(k)}$	Relaxed sum of weights, $\hat{N}_i^{(k)} = \eta_i^{(k)} \tilde{N}_i^{(k)} + (1 - \eta_i^{(k)}) \hat{N}_i^{(k-1)}$	16
n	Sample size, $n \in \mathbb{N}$	9
$\mathcal{P} = \{\mathbb{P}_\theta\}_{\theta \in \Theta}$	Parametric family of probability distributions	9
$\mathbb{P}$	Dominating σ-finite measure	9
$p(y,\theta)$	$p(y,\theta) = d\mathbb{P}_\theta / d\mathbb{P} = p(y) \exp\left[T(y)C(\theta) - B(\theta)\right]$	10
$p_{\varkappa}, \breve{p}_{\varkappa}, \breve{p}_{\varkappa,0}$	Probabilities of unfavorable realizations	22, 25
$s_{ij}^{(k)}$	Statistical penalty, $s_{ij}^{(k)} = \tilde{N}_i^{(k-1)} \mathcal{KL}(\tilde{\theta}_i^{(k-1)}, \tilde{\theta}_j^{(k-1)})$	18
T	$\mathbb{E}_\theta\left[T(Y)\right] = \theta$, see $p(y,\theta)$	10
$U_i^{(k)}$	Neighborhood under consideration, $U_i^{(k)} = \{X_j : \overline{w}_{ij}^{(k)} > 0\}$	40
$\overline{w}_{ij}^{(k)}$	Non-adaptive weights, $\overline{w}_{ij}^{(k)} = K_{\mathrm{loc}}\left(\delta(X_i, X_j)/h^{(k)}\right)$	15
$\tilde{w}_{ij}^{(k)}$	Adaptive weights, $\tilde{w}_{ij}^{(k)} = \overline{w}_{ij}^{(k)} \cdot K_{\mathrm{ad}}\left(s_{ij}^{(k)}/\lambda\right)$	18
$\mathcal{X}$	Metric space with metric δ	9
$\{X_i\}_{i=1}^n$	Deterministic design, $\{X_i\}_{i=1}^n \subseteq \mathcal{X}$	9
$\mathcal{Y}$	Measurable observation space, $\mathcal{Y} \subseteq \mathbb{R}$	9
$\{Y_i\}_{i=1}^n$	Statistically independent observations, $Y_i \sim \mathbb{P}_{\theta(X_i)}$	9
$\mathfrak{Z}_\lambda$	Function for the homogeneous propagation condition	33
$\hat{\mathfrak{Z}}_\lambda$	Function for the inhomogeneous propagation condition	41

Specific symbols in Chapter 6

b	b-value, $b \geq 0$	81
B	Set of applied b-values $b > 0$	91
B_0	$B \cup \{0\}$	91
$\mathfrak{B}$	Number of applied b-values $b > 0$, $\mathfrak{B} = \lvert B \rvert \in \mathbb{N}$	91
G_b	Set of applied gradient directions at b-value b, $G_b \subseteq \mathbb{S}^2 \cup \{\vec{0}\}$	86
L	Number of MR receiver coils, $L \in \mathbb{N}$	§ 6.1.4
L'	Effectively utilized MR receiver coils, $L' \leq L$	§ 6.1.4
S_0	Non-diffusion weighted image	86, 91
S_b	Diffusion weighted image at b-value $b > 0$	86, 91
$\overline{S}_b^{(k)}$	Non-adaptive estimator of (ms)POAS	97
$\tilde{S}_b^{(k)}$	Adaptive estimator of (ms)POAS	97
V	Voxel space	86

Greek symbols

δ	Metric on the design space $\mathcal{X}$	9
$\delta_\kappa, \Delta_\kappa$	Discrepancies on $\mathbb{R}^3 \times \mathbb{S}^2$ (only Chapter 6)	96, 120, 122
ϵ	Propagation level	33, 41
$\eta_i^{(k)}$	Relaxation weight, see $\hat{\theta}_i^{(k)}$	16
Θ	Convex parameter set, $\Theta \subseteq \mathbb{R}$	9
$\Theta_\varkappa$	Compact and convex subset $\Theta_\varkappa \subseteq \Theta$, see $\varkappa$	11
Θ^*	Restriction of the range of $\theta(.)$, $\{\theta_i\}_{i=1}^n \in (\Theta^*)^n$, $\Theta^* \subseteq \Theta$	22
$\theta(.)$	Parameter function that we aim to estimate, $\theta_i = \theta(X_i)$	9
$\theta_{b,m}$	Non-centrality parameter, $S_b(m) \sim \chi_{2L'}(\theta_{b,m})$, $\mathbb{E}\left[S_b(\vec{v},\vec{g})\right]/\sigma = \mu(\theta_{b,m})$ with μ as in Equation (6.7)	91
$\overline{\theta}_i, \overline{\theta}_i^{(k)}$	Non-adaptive estimator, $\overline{\theta}_i^{(k)} = \sum_{j=1}^n \overline{w}_{ij}^{(k)} T(Y_j)/\overline{N}_i^{(k)}$	15
$\tilde{\theta}_i^{(k)}$	Adaptive estimator, $\tilde{\theta}_i^{(k)} = \sum_{j=1}^n \tilde{w}_{ij}^{(k)} T(Y_j)/\tilde{N}_i^{(k)}$	18
$\hat{\theta}_i^{(k)}$	Aggregated estimator, $\hat{\theta}_i^{(k)} = \eta_i^{(k)}\tilde{\theta}_i^{(k)} + (1-\eta_i^{(k)})\hat{\theta}_i^{(k-1)}$	16
$\varkappa$	$I(\theta_1)/I(\theta_2) \leq \varkappa^2$ for all $\theta_1, \theta_2 \in \Theta_\varkappa$, under Ass. A1	11
κ	Balancing parameter between spatial and spherical distances	96
κ_0	Specific choice: $\kappa(h^{(k)}) := \kappa_0/h^{(k)}$, see κ	§ 6.2.2
λ	Adaptation bandwidth, see $\tilde{w}_{ij}^{(k)}$	15, Chapter 3
τ	Memory bandwidth	18
φ_0, φ	Variability bounds	22, 40
$\Omega_\varkappa$	Set of the favorable outcomes under homogeneity	25
$\breve{\Omega}_\varkappa$	Set of the favorable outcomes under inhomogeneity	25
$(\Omega, \mathcal{F}, \mathbb{P}_\theta)$	Probability space	9

Chapter 2

The Propagation-Separation Approach

In this chapter, we will recall some basic concepts, the original and the simplified algorithm of the Propagation-Separation Approach, and the corresponding parameter choices. Additionally, we will compare the Propagation-Separation Approach with the previously introduced Adaptive Weights Smoothing [Polzehl and Spokoiny, 2000], with Lepski's method [Lepskiĭ, 1990; Lepski et al., 1997; Spokoiny, 1998], and with the recent approaches for non-local smoothing as reviewed by Katkovnik et al. [2010]. Finally, we will provide some auxiliary results.

2.1 Basic concepts

We will introduce the statistical setting of our study and several illustrating examples. Then, we will consider a local exponential family model. In particular, we will recall some basic results concerning the Fisher information and the Kullback-Leibler divergence of an exponential family. Finally, we will concentrate on local likelihood estimation.

2.1.1 Statistical model

We assume a local parametric model, more precisely the local likelihood model. This general setting enables a unified approach to a broad class of nonparametric estimation problems as we emphasize in the subsequent example.

Notation 2.1 (Setting). *Let $\mathcal{P} := \{\mathbb{P}_\theta\}_{\theta\in\Theta}$ denote a parametric family of probability distributions with a convex parameter set $\Theta \subseteq \mathbb{R}$, where $(\Omega, \mathcal{F}, \mathbb{P}_\theta)$ forms, for every $\theta \in \Theta$, a probability space with dominating σ-finite measure $\mathbb{P}$. We consider a metric space $\mathcal{X}$ with metric δ, and a measurable observation space $(\mathcal{Y}, \mathcal{B})$, where $\mathcal{Y} \subseteq \mathbb{R}$ and $\mathcal{B}$ denotes the Borel algebra. On the deterministic design $\{X_i\}_{i=1}^n \subseteq \mathcal{X}$ with $n \in \mathbb{N}$, we observe the statistically independent random variables $\{Y_i\}_{i=1}^n$, where $Y_i \sim \mathbb{P}_{\theta(X_i)} \in \mathcal{P}$ and $Y_i(\omega) \in \mathcal{Y}$, $\omega \in \Omega$, for every $i \in \{1, ..., n\}$. Then, we aim to estimate, the unknown parameter function $\theta : \mathcal{X} \to \Theta \subseteq \mathbb{R}$ on the design $\{X_i\}_{i=1}^n$, that is $\{\theta_i\}_{i=1}^n$ with $\theta_i := \theta(X_i)$.*

For the sake of simplicity, we assume the design to be known and the observation space as well as the parameter set to be one-dimensional, that is $\mathcal{Y}, \Theta \subseteq \mathbb{R}$. Basically, the Propagation-

Separation Approach can be applied on any measurable vector space $\mathcal{Y} \subseteq M$ with $Y_i \sim \mathbb{P}_{\theta(X_i)}$ for every $i \in \{1,...,n\}$ and $\theta : \mathcal{X} \to \Theta \subseteq M$, where M is endowed with a possibly asymmetric distance function. The extension of this setting to a random design would require conditional probabilities. We recall some examples from Polzehl and Spokoiny [2006] in order to illustrate the introduced setting. The last example was added in preparation of Chapter 6.

Example 2.2.

1. Gaussian regression: For every $i \in \{1,...,n\}$, we assume that $Y_i = \theta(X_i) + \epsilon_i$, where $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma^2)$ denotes an independent and identically distributed Gaussian error term with variance $\sigma^2 > 0$. The Gaussian regression model appears in many applications, for instance in image processing.
2. Inhomogeneous exponential model: For every $i \in \{1,...,n\}$, let $Y_i \sim \mathrm{Exp}\left(\theta(X_i)\right)$ follow an exponential distribution with parameter $\theta(X_i)$. Possible applications of this model are the reliability or survival analysis and the tail-index estimation theory.
3. Binary response model: For every $i \in \{1,...,n\}$, we assume that $Y_i \sim \mathrm{Bernoulli}\left(\theta(X_i)\right)$ follows a Bernoulli distribution with parameter $\theta(X_i)$. This model is often used for classification, digital imaging, and various econometric applications.
4. Inhomogeneous Poisson model: Let $\mathcal{Y} \subseteq \mathbb{N}$, and assume for every $i \in \{1,...,n\}$ that $Y_i \sim \mathrm{Poiss}\left(\theta(X_i)\right)$ follows a Poisson distribution with parameter $\theta(X_i)$. The binning procedure, see Fan and Gijbels [1996] for more details, provides this model as approximation of the density model. Additionally, it is used in the queuing theory.
5. In the context of diffusion-weighted magnetic resonance images, the setting in Notation 2.1 appears for every $b \in B_0$ with $\mathcal{X} := V \times G_b \subseteq \mathbb{R}^3 \times \mathbb{S}^2$, $\mathcal{Y} := \mathbb{R}$, and non-central chi-distributed observations $S_b(m)/\sigma \sim \chi_{2L'}(\theta_{b,m})$, $m \in \mathcal{X}$. This will be motivated in Section 6.1 and summarized in the introduction of Section 6.2, where the corresponding notation is clarified. In § 6.2.1, a vector-based description of the observations will be constructed. This leads to a multidimensional observation space $\mathcal{Y} := \mathbb{R}^{\mathfrak{B}+1}$, $\mathfrak{B} \in \mathbb{N}$.

2.1.2 The local exponential family model

Following Polzehl and Spokoiny [2006], we assume the parametric family of probability distributions in Notation 2.1 to be an exponential family with standard regularity conditions. We use the common notation

$$C^2(\Theta,\mathbb{R}) := \{f : \Theta \to \mathbb{R} : \text{ the first and second derivative of } f \text{ exist and are continuous}\}.$$

Assumption A1 (Local exponential family model)**.** *The parametric family $\mathcal{P} = \{\mathbb{P}_\theta\}_{\theta\in\Theta}$ in Notation 2.1 is an exponential family. More precisely, there are two functions $C, B \in C^2(\Theta,\mathbb{R})$, a non-negative function $p : \mathcal{Y} \to [0,\infty)$, and a sufficient statistic $T : \mathcal{Y} \to \mathbb{R}$ such that*

$$p(y,\theta) := d\mathbb{P}_\theta/d\mathbb{P}(y) = p(y)\exp\left[T(y)C(\theta) - B(\theta)\right], \qquad \theta \in \Theta,$$

where C is strictly monotonic increasing. The parameter θ satisfies $B'(\theta) = \theta\, C'(\theta)$,

$$\int p(y,\theta)\mathbb{P}(dy) = 1, \quad \textit{and} \quad \mathbb{E}_\theta\left[T(Y)\right] = \int T(y)p(y,\theta)\mathbb{P}(dy) = \theta. \tag{2.1}$$

Remark 2.3.

- In [Polzehl and Spokoiny, 2006, Ass. (A1)], the authors presumed the sufficient statistic T to equal the identity map. In this study, we explicitly allow any sufficient statistic in order

to clarify where this transformation T comes into play. As it turns out in Lemma 2.5 (3), this extension of Assumption A1 leaves the Kullback-Leibler divergence unchanged due to Equation (2.1).
- The required unbiasedness $\mathbb{E}_\theta[T(Y)] = \theta$ of the parameter θ in Equation (2.1) can be achieved via reparametrization with $\theta := t(\vartheta)$, where $t(\vartheta) := \mathbb{E}_\vartheta[T(Y)]$. This will be discussed in § 4.4.1.
- A list of parametric families which are in accordance with Assumption A1 are summarized in Table 2.1.

In our subsequent analysis, the notions of the Kullback-Leibler divergence $\mathcal{KL}(.,.)$ and the Fisher information $I(.)$ will be important.

Notation 2.4. *For $\theta, \theta' \in \Theta$, we set*

$$\mathcal{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) := \int \log\left(\frac{d\mathbb{P}_\theta}{d\mathbb{P}_{\theta'}}(y)\right)\mathbb{P}_\theta(dy) \quad \text{and} \quad I(\theta) := -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log p(Y,\theta)\right].$$

Moreover, we introduce the abbreviatory notation $\mathcal{KL}(\theta,\theta') := \mathcal{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'})$ for the Kullback-Leibler divergence between the probability distributions $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$ with parameters $\theta, \theta' \in \Theta$.

Lemma 2.5. *Under Assumption A1 it holds the following.*

1. *The Fisher information satisfies $I(\theta) = C'(\theta)$ for all $\theta \in \Theta$.*
2. *For every compact and convex subset $\Theta' \subseteq \Theta$, there is a constant $\varkappa \geq 1$ such that*

$$\frac{I(\theta_1)}{I(\theta_2)} \leq \varkappa^2 \qquad \text{for all } \theta_1, \theta_2 \in \Theta'. \tag{2.2}$$

3. *The Kullback-Leibler divergence is convex with respect to the first argument. Additionally, it has an explicit representation, and it satisfies a quadratic expression*

$$\begin{aligned} \mathcal{KL}(\theta,\theta') &= \theta\left[C(\theta) - C(\theta')\right] - \left[B(\theta) - B(\theta')\right] && (2.3)\\ &= r(\theta_*,\theta_0)\left[\theta - \theta'\right]^2/2, && (2.4) \end{aligned}$$

where $\theta, \theta' \in \Theta$, $r(\theta_*,\theta_0) := [I(\theta_*)]^2/I(\theta_0)$, and θ_*, θ_0 lie between θ and θ'.

Equation (2.2) allows the following notations.

Notation 2.6. *For every compact and convex subset $\Theta' \subseteq \Theta$, we set*

$$\varkappa := \max\{I(\theta_1)/I(\theta_2) : \theta_1, \theta_2 \in \Theta'\} \geq 1 \quad \text{and} \quad \Theta' := \Theta_\varkappa.$$

Vice versa, for every constant $\varkappa \geq 1$, we use the notation $\Theta' := \Theta_\varkappa$ for any compact and convex set $\Theta' \subseteq \Theta$ which satisfies Equation (2.2)*.*

The set $\Theta_\varkappa$ should be sufficiently large such that $\theta(X_i) \in \Theta_\varkappa$ for all $i \in \{1,...,n\}$. We specify its precise choice where necessary. Lemma 2.5 and Table 2.1 provide under Assumption A1 explicit expressions for the Kullback-Leibler divergence and the Fisher information.

Example 2.7. We consider the same probability distributions as in Example 2.2 (1) and (2). For $\mathcal{P} := \{\mathcal{N}(\theta,\sigma^2)\}_{\theta\in\mathbb{R}}$, it holds

$$I(\theta) = 1/\sigma^2 \quad \text{and} \quad \mathcal{KL}(\theta,\theta') = (\theta-\theta')^2/(2\sigma^2), \quad \theta,\theta' \in \Theta.$$

For $\mathcal{P} := \{\mathrm{Exp}(1/\theta)\}_{\theta\in(0,\infty)}$, it holds

$$I(\theta) = 1/\theta^2 \quad \text{and} \quad \mathcal{KL}(\theta,\theta') = \theta/\theta' - 1 - \ln(\theta/\theta'), \quad \theta,\theta' \in \Theta.$$

$\mathcal{P}, \mathcal{Y}$	Θ	$p(y)$	$T(y)$	$C_t(\vartheta)$	$B_t(\vartheta)$	$\mathbb{E}_\vartheta[T(Y)]$
$\mathcal{N}(\vartheta, \sigma^2)$ $y \in \mathbb{R}$	$\mathbb{R}$	$\frac{e^{-y^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}}$	y	$\frac{\vartheta}{\sigma^2}$	$\frac{\vartheta^2}{2\sigma^2}$	ϑ
$\mathcal{N}(0, \vartheta)$ $y \in \mathbb{R}$	$(0,\infty)$	$\frac{1}{\sqrt{2\pi}}$	y^2	$-\frac{1}{2\vartheta}$	$\frac{\ln\vartheta}{2}$	ϑ
$\log\mathcal{N}(\vartheta, \sigma^2)$ $y \in (0,\infty)$	$(0,\infty)$	$\frac{e^{-(\ln y)^2/(2\sigma^2)}}{y\sqrt{2\pi\sigma^2}}$	$\ln y$	$\frac{\vartheta}{\sigma^2}$	$\frac{\vartheta^2}{2\sigma^2}$	ϑ
$\Gamma(p, \vartheta)$ $y \in (0,\infty)$	$(0,\infty)$	$\frac{y^{p-1}}{\Gamma(p)}$	y	$-\frac{1}{\vartheta}$	$p\ln\vartheta$	$p\vartheta$
$\mathrm{Exp}\left(\frac{1}{\vartheta}\right)$ $y \in [0,\infty)$	$(0,\infty)$	1	y	$-\frac{1}{\vartheta}$	$\ln\vartheta$	ϑ
$\mathrm{Erlang}\left(n, \frac{1}{\vartheta}\right)$ $y \in [0,\infty)$	$(0,\infty)$	$\frac{y^{n-1}}{(n-1)!}$	y	$-\frac{1}{\vartheta}$	$n\ln\vartheta$	$n\vartheta$
$\mathrm{Rayleigh}(\vartheta)$ $y \in [0,\infty)$	$(0,\infty)$	y	y^2	$-\frac{1}{2\vartheta^2}$	$2\ln\vartheta$	$2\vartheta^2$
$\mathrm{Weibull}(\vartheta, k)$ $y \in [0,\infty)$	$(0,\infty)$	ky^{k-1}	y^k	$-\frac{1}{\vartheta^k}$	$k\ln\vartheta$	ϑ^k
$kY/\vartheta \sim \chi^2(k)$ $y \in [0,\infty)$	$(0,\infty)$	$\frac{k^{k/2}y^{k/2-1}}{2^{k/2}\Gamma(k/2)}$	y	$-\frac{k}{2\vartheta}$	$\frac{k\ln\vartheta}{2}$	ϑ
$\mathrm{Pareto}\left(x_m, \frac{1}{\vartheta}\right)$ $y \in [x_m,\infty)$	$(0,1)$	$\frac{1}{y}$	$\ln\left(\frac{y}{x_m}\right)$	$-\frac{1}{\vartheta}$	$\ln(\vartheta)$	ϑ
$\mathrm{Poiss}(\vartheta)$ $y := k \in \mathbb{N}$	$(0,\infty)$	$1/k!$	k	$\ln\vartheta$	ϑ	ϑ
$\mathrm{Bin}(n, \vartheta)$ $y := k \in \{0,1,...,n\}$	$(0,1)$	$\binom{n}{k}$	k	$\ln\left(\frac{\vartheta}{1-\vartheta}\right)$	$-n\ln(1-\vartheta)$	$n\vartheta$
$\mathrm{NegativeBin}(r, \vartheta)$ $y := k \in \mathbb{N}$	$(0,1)$	$\binom{k+r-1}{k}$	k	$\ln\vartheta$	$-r\ln(1-\vartheta)$	$\frac{r\vartheta}{1-\vartheta}$
$\mathrm{Bernoulli}(\vartheta)$ $y := k \in \{0,1\}$	$(0,1)$	1	k	$\ln\left(\frac{\vartheta}{1-\vartheta}\right)$	$-\ln(1-\vartheta)$	ϑ

Table 2.1: One-parametric exponential families which satisfy Assumption A1, possibly after reparametrization with $\theta := t(\vartheta)$, where $t(\vartheta) := \mathbb{E}_\vartheta[T(Y)]$.

Finally, we recall a technical lemma by Polzehl and Spokoiny [2006, Lem. 5.2]. In Section 2.4, we will discuss some details concerning the applicability of this lemma.

Lemma 2.8. *Suppose Assumption A1, and let $\Theta_\varkappa \subseteq \Theta$ and $\varkappa \geq 1$ be as in Notation 2.6. For any sequence $\theta_0, \theta_1, ..., \theta_m \in \Theta_\varkappa$, it holds*

$$\mathcal{KL}^{1/2}(\theta_0, \theta_m) \leq \varkappa \sum_{l=1}^{m} \mathcal{KL}^{1/2}(\theta_{l-1}, \theta_l).$$

2.1.3 Local likelihood estimation

Let us consider the local likelihood model in Notation 2.1. Recall that the standard maximum likelihood estimator (MLE) is constructed to select the set of parameter values which maximize the log-likelihood function L. More precisely, the MLE is given as

$$\operatorname*{argsup}_{\theta \in \Theta} L(\theta) \quad \text{with} \quad L(\theta) := \sum_{j=1}^{n} \log p(Y_j, \theta),$$

where $p := d\mathbb{P}_\theta / d\mathbb{P}$ denotes the probability density with respect to the dominating measure $\mathbb{P}$. In order to enable more flexible modeling, we turn to the locally weighted MLE.

Notation 2.9. *The weighted maximum likelihood estimator is given as*

$$\theta_i^{(\mathrm{MLE})} := \operatorname*{argsup}_{\theta \in \Theta} L(\overline{W}_i, \theta) \quad \textit{with} \quad L(\overline{W}_i, \theta) := \sum_{j=1}^{n} \overline{w}_{ij} \log p(Y_j, \theta), \tag{2.5}$$

where the weighting scheme $\overline{W}_i$ equals the set $\{\overline{w}_{ij}\}_{j=1}^n$.

Under the very general setting in Notation 2.1, this estimator does not need to have an explicit or unique solution. Therefore, we presume the local exponential family model in Assumption A1 (page 10) and state the following lemma.

Lemma 2.10. *Under Assumption A1 it holds the following.*

1. *The weighted MLE in Equation* (2.5) *has a unique solution which equals the corresponding weighted mean of the transformed observations. More precisely, for $\overline{W}_i := \{\overline{w}_{ij}\}_{j=1}^n$, it holds*
$$\theta_i^{(\mathrm{MLE})} = \operatorname*{argsup}_{\theta \in \Theta} \sum_{j=1}^{n} \overline{w}_{ij} \log p(Y_j, \theta) = \sum_{j=1}^{n} \overline{w}_{ij} T(Y_j) / \overline{N}_i =: \overline{\theta}_i, \tag{2.6}$$
where $\overline{N}_i$ denotes the sum of weights $\sum_j \overline{w}_{ij}$.
2. *If the transformed observations have a homogeneous variance $\sigma^2 = \mathrm{Var}[T(Y_i)]$ for every $i \in \{1, ..., n\}$, then the variance reduction by the weighted MLE is given by the factor*
$$\mathrm{Var}\left(\theta_i^{(\mathrm{MLE})}\right) / \sigma^2 = \sum_{j=1}^{n} \overline{w}_{ij}^2 / (\overline{N}_i)^2.$$
Additionally, we get the upper bound $\mathrm{Var}(\theta_i^{(\mathrm{MLE})}) \leq \sigma^2 / \overline{N}_i$ if $\overline{W}_i \in [0,1]^n$.

Remark 2.11. The weighted mean in Equation (2.6) is a refinement of the simple running local average. It coincides with the Nadaraya-Watson estimator if $\overline{w}_{ij} := K(\|X_i - X_j\|/h)/h$ for all $i, j \in \{1, ..., n\}$, where $h > 0$ denotes a bandwidth and K is usually a symmetric probability density [Fan and Gijbels, 1996, §2.2].

For additive noise models, the approximation error is usually measured by some function which depends on the residual $\overline{\theta}_i - \theta$. For the local likelihood model, this approach is less natural since the target parameter does not need to be additive. For instance, in the case of the exponential distribution the target parameter θ is a scale parameter. Then, it is more natural to consider the ratio $\overline{\theta}_i/\theta$. The fitted log-likelihood

$$L(\overline{W}_i, \theta, \theta') := L(\overline{W}_i, \theta) - L(\overline{W}_i, \theta'), \qquad \theta, \theta' \in \Theta,$$

provides a natural measure for the approximation distance under a local likelihood model. Under Assumption A1 the fitted log-likelihood is closely related with the Kullback-Leibler divergence.

Lemma 2.12. *Under Assumption A1, it holds* $L(\overline{W}_i, \overline{\theta}_i, \theta) = \overline{N}_i \mathcal{KL}(\overline{\theta}_i, \theta)$ *for every* $\theta \in \Theta$*, where* $\overline{\theta}_i$ *is as in Equation* (2.6)*.*

This shows with Example 2.7 that the fitted log-likelihood depends for the (additive) Gaussian model on the difference $\overline{\theta}_i - \theta$, while considering for the exponential distribution the ratio $\overline{\theta}_i/\theta$. We recall an exponential bound for the excess probability $\mathbb{P}(\overline{N}_i \mathcal{KL}(\overline{\theta}_i, \theta) > z)$ of the Kullback-Leibler divergence between the weighted mean $\overline{\theta}_i$ and the true parameter θ. This result will be crucial in our study. It was established by Polzehl and Spokoiny [2006, Thm. 2.1].

Theorem 2.13. *Let Assumption A1 be satisfied, and presume a parametric model,* $\theta(.) \equiv \theta$*. Then, for each* $i \in \{1, ..., n\}$ *and every weighting scheme* $\overline{W}_i := \{\overline{w}_{ij}\}_{j=1}^n \in [0,1]^n$*, we get*

$$\mathbb{P}\left(\overline{N}_i \mathcal{KL}(\overline{\theta}_i, \theta) > z\right) \leq 2e^{-z} \quad \text{for all } z > 0$$

with $\overline{N}_i$ *and* $\overline{\theta}_i$ *as in Equation* (2.6)*.*

2.2 Methodology of the Propagation-Separation Approach

The Propagation-Separation Approach provides pointwise estimates of the unknown parameter function $\theta(.)$ in Notation 2.1. The method is constructed to yield similar results as non-adaptive smoothing within homogeneity regions (propagation), while avoiding smoothing across discontinuities (separation). Therefore, it is especially powerful in the case of large homogeneous regions and sharp discontinuities. Originally, it relies on a local constant model. However, it can be extended to a local polynomial model in an analogous manner as presented in [Polzehl and Spokoiny, 2004] for an additive random noise model. Hence, the procedure is applicable to a broad class of nonparametric models. In our study, we concentrate on the local constant model for the sake of simplicity. Important application can be found in image processing, where this is often reasonable.

First, we will recall the original algorithm as introduced by Polzehl and Spokoiny [2006]. Then, we will detail the corresponding parameter choices. Furthermore, we will introduce a simplified version of the method, where the additionally included memory step of the original procedure is omitted. We will close with two illustrating examples in order to provide some intuition for the heuristic behavior of the simplified algorithm.

2.2.1 The original algorithm

The Propagation-Separation Approach presumes a local parametric model. Then, it estimates the parametric neighborhoods and the associated parameters simultaneously during an iterative

procedure. The algorithm is based on the weighted MLE in Equation (2.5), where the non-adaptive weights are replaced by structure-adaptive weights which penalize large differences between the corresponding unknown parameter values. For this purpose, the algorithm uses previously aggregated information to improve the pointwise estimates during iteration. We emphasize that the Propagation-Separation Approach does not use adaptive parameters. It is adaptive in the sense that the returned estimator function is based on structure-adaptive weights which describe the homogeneity regions of the unknown parameter function $\theta(.)$.

In each iteration step, the pointwise estimator is defined as the weighted mean in Equation (2.6), which equals the weighted MLE under Assumption A1 (page 10). For each design point X_i with $i \in \{1, ..., n\}$, the local weights are chosen adaptively as a product of two kernel functions. The *location kernel* acts on the design space $\mathcal{X}$, where it determines the vicinity under consideration, using the metric δ in Notation 2.1. The *adaptation kernel* compares the pointwise parameter estimates of the previous iteration step in terms of the Kullback-Leibler divergence, providing iteratively updated local weights. Here, a *statistical penalty*, which is based on the Kullback-Leibler divergence, is used as a test statistic for homogeneity. Basically, it checks whether the previous estimate at X_j belongs to the confidence interval of the previous estimate at X_i [Polzehl and Spokoiny, 2006, Eq. (3.1)]. An additional motivation follows from the relationship between the fitted log-likelihood and the Kullback-Leibler divergence, which we established in Lemma 2.12. Polzehl and Spokoiny [2004] introduced the fitted log-likelihood as an asymmetric modification of the classic two populations likelihood-ratio test statistic, assuming a Gaussian regression model. They interpreted the statistical penalty as the difference between the weighted MLE at X_i, using the weighting scheme $\overline{W}_i$ and the 'plug-in' MLE at X_j, which is evaluated with respect to the weighting scheme $\overline{W}_i$ although it was calculated with the weighting scheme $\overline{W}_j$.

For both kernels, a bandwidth controls how much information is taken into account. Usually, the adaptation bandwidth is chosen as a fixed constant, while the location bandwidth increases along the number of iterations. Starting at a small neighborhood, in each iteration step, the considered region is extended. According to Lemma 2.12, the statistical penalty becomes more restrictive during iteration by the factor $\hat{N}_i^{(k-1)}$ which equals the relaxed sum of the adaptive weights. This approximately compensates the already achieved variance reduction, see Lemma 2.10 (2) for the case of non-adaptive weights and a homogeneous variance. The described procedure enables an advancing variance reduction during iteration, while avoiding blurring at structural borders.

Finally, an additional *memory step* ensures a certain stability of estimates. In each iteration step, the *memory penalty* compares, for every design point, the new estimate with the previous one. In the case of a significant difference the new estimate is relaxed, replacing it by a value between the two estimates. The memory step provides a smooth transition of the pointwise estimates during iteration. For a detailed study about spatial aggregation of local likelihood estimates, we refer the reader to Belomestny and Spokoiny [2007].

We turn to a formal description of the algorithm, and we start by introducing some notation.

Notation 2.14. *Suppose Assumption A1. We fix three non-increasing kernel functions*

$$K_{\mathrm{loc}}, K_{\mathrm{ad}}, K_{\mathrm{me}} : [0, \infty) \to [0, 1]$$

with support $[0, 1)$*, satisfying* $K_.(0) = 1$*. These kernels will be used for location, for adaptation, and for the memory step, respectively. Moreover, let* $\lambda > 0$ *denote the bandwidth of the adaptation kernel, and let* $\{h^{(k)}\}_{k=0}^{k^*}$ *be an increasing sequence of pre-specified location bandwidths with* $h^{(0)} > 0$*. For the memory step, we choose the minimal memory effect* $\eta_0 \in [0, 1)$

and the memory bandwidth $\tau > 0$. Then, we call the weighted mean $\overline{\theta}_i^{(k)}$ in Equation (2.6) *with $\overline{w}_{ij}^{(k)} := K_{\mathrm{loc}}\left(\delta(X_i, X_j)/h^{(k)}\right)$ the non-adaptive estimator of θ_i. Additionally, we recall the notion $\overline{N}_i^{(k)} = \sum_j \overline{w}_{ij}^{(k)}$.*

The effect of different choices for the kernel functions is negligible. This follows from experience and, for the location kernel K_{loc}, from the theoretical results in [Scott, 1992, §6.2.3]. One possible choice is given in Equation (2.10) (page 19). The other quantities will be specified in § 2.2.2. Now we present the algorithm of the Propagation-Separation Approach. More details can be found in [Polzehl and Spokoiny, 2006, §3].

Algorithm 1 (Propagation-Separation Approach with memory step)**.**

1. *Input parameters: Sequence of bandwidths $\{h^{(k)}\}_{k=0}^{k^*}$, adaptation bandwidth λ, the memory bandwidth τ, and the minimal memory effect η_0.*
2. *Initialization: $\hat{\theta}_i^{(0)} := \overline{\theta}_i^{(0)}$ and $\hat{N}_i^{(0)} := \overline{N}_i^{(0)}$ for all $i \in \{1, ..., n\}$, $k := 1$.*
3. *Iteration: Calculate, for every $i, j = 1, ..., n$,*
   the non-adaptive weights $\overline{w}_{ij}^{(k)} := K_{\mathrm{loc}}\left(\delta(X_i, X_j)/h^{(k)}\right)$,
   the statistical penalty $s_{ij}^{(k)} := \hat{N}_i^{(k-1)} \mathcal{KL}(\hat{\theta}_i^{(k-1)}, \hat{\theta}_j^{(k-1)})$,
   the adaptive weights $\tilde{w}_{ij}^{(k)} := \overline{w}_{ij}^{(k)} \cdot K_{\mathrm{ad}}\left(s_{ij}^{(k)}/\lambda\right)$,
   the sum of the adaptive weights $\tilde{N}_i^{(k)} := \sum_j \tilde{w}_{ij}^{(k)}$,
   and the adaptive estimator

$$\tilde{\theta}_i^{(k)} := \sum_{j=1}^{n} \tilde{w}_{ij}^{(k)} T(Y_j) / \tilde{N}_i^{(k)}.$$

4. *Memory step: Calculate, for every $i, j = 1, ..., n$,*
   the sum of the non-adaptive weights $\overline{N}_i^{(k)} := \sum_j \overline{w}_{ij}^{(k)}$,
   the memory penalty $m_i^{(k)} := \overline{N}_i^{(k)} \mathcal{KL}(\tilde{\theta}_i^{(k)}, \hat{\theta}_i^{(k-1)})$,
   the relaxation weight $\eta_i^{(k)} := (1 - \eta_0) K_{\mathrm{me}}\left(m_i^{(k)}/\tau\right)$,
   the relaxed estimator $\hat{\theta}_i^{(k)} := \eta_i^{(k)} \tilde{\theta}_i^{(k)} + (1 - \eta_i^{(k)}) \hat{\theta}_i^{(k-1)}$, and the relaxed sum of the adaptive weights $\hat{N}_i^{(k)} := \eta_i^{(k)} \tilde{N}_i^{(k)} + (1 - \eta_i^{(k)}) \hat{N}_i^{(k-1)}$.
5. *Stopping: Stop if $k = k^*$, and return $\hat{\theta}_i^{(k^*)}$ for all $i \in \{1, ..., n\}$, otherwise increase k by 1.*

We emphasize that the the data-driven statistical penalty $s_{ij}^{(k)}$ makes the adaptive weights $\tilde{w}_{ij}^{(k)}$, their sum $\tilde{N}_i^{(k)}$, and the relaxed sum $\hat{N}_i^{(k)}$ random. In contrast, we notice that the input parameters, the non-adaptive weights $\overline{w}_{ij}^{(k)}$, and their sum $\overline{N}_i^{(k)}$ are deterministic.

Corollary 2.15. *If $s_{ij}^{(k)} = 0$ and $\eta_i^{(k)} = 1$ for all $i, j \in \{1, ..., n\}$, then it holds $K_{\mathrm{ad}}\left(s_{ij}^{(k)}/\lambda\right) = 1$ for all i, j, and the non-adaptive estimator coincides with the adaptive and the relaxed one, $\tilde{\theta}_i^{(k)} = \hat{\theta}_i^{(k)} = \overline{\theta}_i^{(k)}$.*

2.2.2 Choice of the input parameters

Next we render the choices of the input parameters more precisely. As usual, the basic strategies follow from mathematical arguments, while the precise choices within a certain range may be arbitrary or follow from experience.

First, we consider the increasing sequence of location bandwidths. Recall that the algorithm is initialized by the non-adaptive estimator. A choice of the initial bandwidth $h^{(0)}$ such that $\overline{w}_{ij}^{(0)} = 0$ for all $i, j \in \{1, ..., n\}$ with $i \neq j$ avoids smoothing among distinct homogeneity regions. A theoretical drawback of this choice will be discussed in Remark 2.18. For the subsequent bandwidths $\{h^{(k)}\}_{k=1}^{k^*}$, there have been two proposals. Polzehl and Spokoiny [2006, §3.4] recommended to set $h^{(k)} := a^k h^{(0)}$ with $a \approx 1.25^{1/d}$, where d denotes the dimension of the design space $\mathcal{X}$. For every $X_i \in \mathcal{X}$, this choice ensures, up to boundary effects, an exponential growth of the mean number of design points X_j with non-zero weights $\overline{w}_{ij}^{(k)} \neq 0$. Alternatively, one could ensure a constant variance reduction of the non-adaptive estimator in Notation 2.14, see Becker et al. [2012, 2013]. Motivated by Lemma 2.10 (2), we consider the quantities

$$q_i(h) := \frac{\sum_{j=1}^n K_{\mathrm{loc}} \left(\delta(X_i, X_j)/h\right)^2}{\left[\sum_{j=1}^n K_{\mathrm{loc}} \left(\delta(X_i, X_j)/h\right)\right]^2}, \qquad h > 0,$$

that satisfy $q_i(h) \in [n^{-1}, 1]$ for all $h > 0$, and $q_i(h) \to n^{-1}$ if $h \to \infty$. Starting with some fixed initial bandwidth $h^{(0)} > 0$, the subsequent values $h^{(k)}$, $k \geq 1$, are determined by numerically solving the equation

$$q_i(h^{(k-1)}) = 1.25 \cdot q_i(h^{(k)}), \qquad h^{(k)} > 0.$$

If the location kernel K_{loc} is non-increasing, continuous, and has the support $[0, 1)$ with $K_{\mathrm{loc}}(0) = 1$, then the existence of a unique solution $h^{(k)}$ is ensured for all iteration steps $k \in \{1, ..., k_0 - 1\}$, where

$$k_0 := \min\left\{k \in \{1, ..., k^*\} : 1.25^{-1} \cdot q_i(h^{(k-1)}) < n^{-1}\right\}.$$

For all later iteration steps $k \in \{k_0, ..., k^*\}$, we apply the formal choice $h^{(k)} := \infty$, where $\overline{w}_{ij}^{(k)} = 1$ for all $i, j \in \{1, ..., n\}$. The first bandwidth choice has the drawback to depend on the design dimension d via the factor $a \approx 1.25^{1/d}$. In both cases, the specific value 1.25 could be replaced by any constant $a > 1$. For instance in [Li et al., 2011, 2012], the authors used the first bandwidth choice with $a = 1.1^{1/d}$.

The maximal location bandwidth $h^{(k^*)}$ is determined by the maximal number of iterations k^*. Polzehl and Spokoiny [2006] allowed an arbitrarily large choice which is only bounded by the available computation time. Theoretically, this was motivated by the established stability of estimates up to some constant [Polzehl and Spokoiny, 2006, Thm. 5.7], which results from the memory step. Additionally, the numerical examples in Chapter 5 indicate that the adaptivity itself provides an intrinsic stopping criterion. For further details concerning the choice of k^*, we refer the reader to Section 5.4.3.

The amount of adaptivity is determined by the adaptation bandwidth λ, which can be specified by the propagation condition independent of the observations at hand, see Chapter 3 and Polzehl and Spokoiny [2006, §3.4 & 3.5]. For $\lambda \to \infty$, the algorithm results in non-adaptive estimates as defined in Notation 2.14 (over-smoothing), while small values lead to adaptation to noise (under-smoothing), where the estimation function just interpolates the observations.

The memory bandwidth scales the memory penalty $m_i^{(k)}$ in a similar manner as the adaptation bandwidth the statistical penalty. It can be chosen by the propagation condition after λ has been fixed. Polzehl and Spokoiny [2006, §3.4] recommended the choice $\tau = \max\{\tau_1 - \tau_2 \log h^{(k)}, \tau_0\}$ for some $\tau_0, \tau_1, \tau_2 \geq 0$, leading to a decreasing memory bandwidth during iteration. For the minimal memory effect η_0, they propose a default value of 0.25, without giving further details. In their theoretical study, they set $\eta_0 = 0$, mentioning that the extension to $\eta_0 < 1/2$ is straightforward.

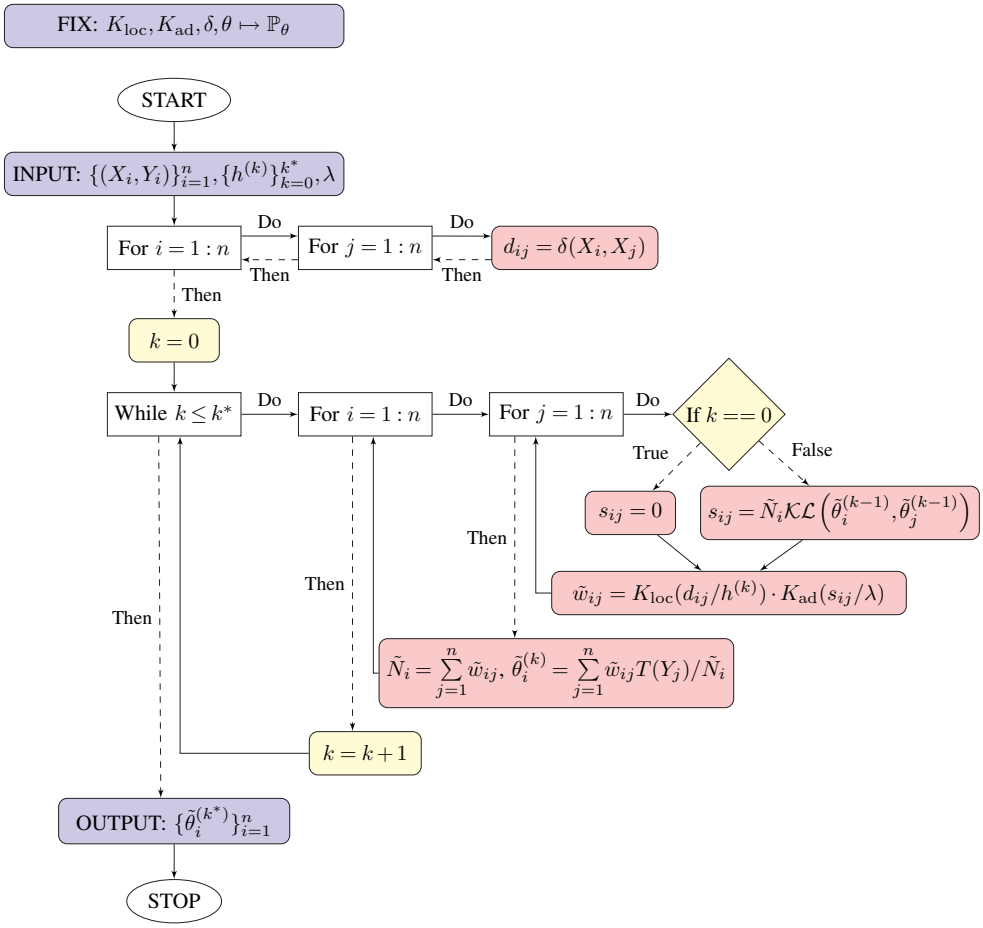

Figure 2.1: Flowchart of the simplified Propagation-Separation Approach in Algorithm 2.

2.2.3 A simplified algorithm

In many applications of the Propagation-Separation Approach, see for example Becker et al. [2012, 2013], Divine et al. [2008], Li et al. [2011, 2012], and Tabelow et al. [2008], the memory step was omitted without considerable change of the general behavior. This raises the question whether the memory step is really needed. Therefore, in this study, we will attempt to provide, for the simplified Propagation-Separation Algorithm, similar results as stated by Polzehl and Spokoiny [2006]. Additionally, we will examine the impact of the memory step within a numerical study. The simplified version of the original Propagation-Separation Approach in Algorithm 1 results from the formal choice $\eta_i^{(k)} = 1$ for all $i \in \{1, ..., n\}$ and $k \in \{0, ..., k^*\}$. The resulting procedure is introduced in Algorithm 2 and illustrated in Figure 2.1.

Algorithm 2 (Propagation-Separation Approach without memory step)**.**

1. *Input parameters: Sequence of bandwidths* $\{h^{(k)}\}_{k=0}^{k^*}$ *and adaptation bandwidth* λ.
2. *Initialization: Set* $k := 0$*, and apply, for all* $i, j = 1, ..., n$*, Equation* (2.7) *with* $s_{ij}^{(0)} := 0$*. Then, increase* k *by* 1.
3. *Iteration: Calculate, for every* $i, j = 1, ..., n$*,*
   the statistical penalty $s_{ij}^{(k)} := \tilde{N}_i^{(k-1)} \mathcal{KL}(\tilde{\theta}_i^{(k-1)}, \tilde{\theta}_j^{(k-1)})$*,*
   the adaptive weights $\tilde{w}_{ij}^{(k)} := K_{\text{loc}}\left(\delta(X_i, X_j)/h^{(k)}\right) \cdot K_{\text{ad}}\left(s_{ij}^{(k)}/\lambda\right)$*,*

the sum of weights $\tilde{N}_i^{(k)} := \sum_j \tilde{w}_{ij}^{(k)}$,
and the adaptive estimator

$$\tilde{\theta}_i^{(k)} := \sum_{j=1}^{n} \tilde{w}_{ij}^{(k)} T(Y_j) / \tilde{N}_i^{(k)}. \tag{2.7}$$

4. *Stopping: Stop if* $k = k^*$*, and return* $\tilde{\theta}_i^{(k^*)}$ *for all* $i \in \{1, ..., n\}$*, otherwise increase* k *by* 1.

2.2.4 An illustrative example

In order to provide some intuition, we illustrate the general behavior of the simplified Propagation-Separation Approach in Algorithm 2. Here, we used the **R**-package **aws** by Polzehl [2012], where the memory step is omitted by default.

On $\mathcal{X} := \{1, ..., 1000\}$, we considered a piecewise constant function,

$$\theta(x) := \begin{cases} 5 & \text{if } x \in \{1, ..., 200\}, \\ 7 & \text{if } x \in \{201, ..., 400\}, \\ 1 & \text{if } x \in \{401, ..., 550\}, \\ 1.5 & \text{if } x \in \{551, ..., 700\}, \\ 2 & \text{if } x \in \{701, ..., 850\}, \\ 1.5 & \text{if } x \in \{851, ..., 1000\}, \end{cases} \tag{2.8}$$

and a piecewise polynomial one,

$$\theta(x) := \begin{cases} 7 + x/250 & \text{if } x \in \{1, ..., 250\}, \\ 11 + ((x - 450)/100)^2/2 & \text{if } x \in \{251, ..., 750\}, \\ 6 - (x - 750)/200 & \text{if } x \in \{751, ..., 1000\}. \end{cases} \tag{2.9}$$

The statistically independent observations were sampled from a Gaussian distribution with locally varying expectation, $Y_i \sim \mathcal{N}(\theta(X_i), 1)$. Then, the plots were provided by the function `aws`, using the default parameter choices and the following kernel functions,

$$K_{\text{loc}}(x) := (1 - x^2)_+ \quad \text{and} \quad K_{\text{ad}}(x) := \min\{1, 2 - 2x\}_+. \tag{2.10}$$

In the first row of Figure 2.2, we show the results for the piecewise constant function (2.8) with increasing location bandwidths `hmax` $= 10, 90, 2000$. The second row of Figure 2.2 is based on the piecewise smooth function (2.9), setting `hmax` $= 10, 50, 2000$. For both examples, the intermediate bandwidth minimizes the mean absolute error.

We summarize the following heuristic observations.

- Homogeneous regions with sufficiently sharp discontinuities are separated by the algorithm, leading to a consistent estimator, see $x \in \{1, ..., 400\}$ in the first row of Figure 2.2. If the discontinuities or the sample size of the corresponding homogeneous regions are too small, separation fails. Then, different homogeneous regions are treated as one, yielding a bounded estimation bias. This is illustrated for $x \in \{401, ..., 1000\}$. The variance of the separation points where the algorithm creates a discontinuity for a sufficiently large location bandwidth decreases with increasing contrast between the true homogeneous regions.

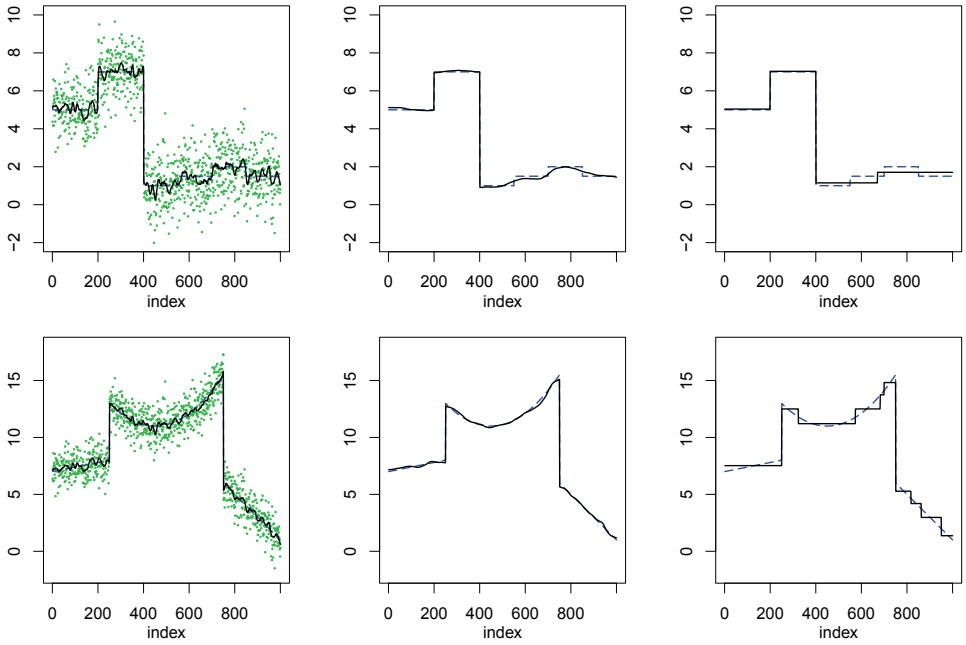

Figure 2.2: Results of Algorithm 2 (black solid line) on Gaussian observations (circles in the first column) for the parameter function $\theta(.)$ (blue dashed line). Above: Results of the piecewise constant parameter function (2.8) for increasing location bandwidths (from left to right) $\texttt{hmax} = 10, 90, 2000$. Below: Results of the piecewise smooth function (2.9) for $\texttt{hmax} = 10, 50, 2000$.

- In the second row of Figure 2.2, we consider the case of model misspecification, that is a parameter function $\theta(.)$ that is not piecewise constant. Here, the algorithm forces the final estimator into a step function. The step size mainly depends on the smoothness of the parameter function $\theta(.)$ and on the size of the adaptation bandwidth λ. However, the estimation bias can be reduced by an accurate stopping.

Thus, the heuristic properties are clear. However, the iterative approach complicates a theoretical verification considerably. As a consequence, some heuristic properties do not allow a definite proof. Nevertheless, the following theoretical study will provide a deeper insight into the behavior of the algorithm and the respective impact of the involved components, such as the adaptation bandwidth or the smoothness properties of the unknown parameter function. Together with the numerical study in Chapter 5 this will justify the omittance of the memory step, confirming the above heuristic properties.

2.3 Related work

Polzehl and Spokoiny [2006] introduced the Propagation-Separation Approach as an extension of their *Adaptive Weights Smoothing (AWS)* procedure [Polzehl and Spokoiny, 2000]. Essentially, the algorithm of AWS works very similar to Algorithm 1 (page 16). The pointwise estimates in both methods are defined by weighted means with iteratively updated weights and a subsequent relaxation by a memory step. In AWS, the localization and memory kernels are based on the uniform kernel, which leads to windows instead of monotonically decreasing weighting schemes. However, the main difference is the restriction of AWS to an additive noise model. In accordance to this model the statistical penalty of AWS compares the estimates of the previous

iteration steps by the standardized absolute value of their numerical difference, taking the already achieved variance reduction into account. For instance, in the case of homoscedastic Gaussian distributed observations, the resulting term (approximately) conforms with the statistical penalty in Algorithm 1. The extension to the local likelihood model has two main benefits. First, it enabled the justification of theoretical properties, and second it provided a considerable increase of the application area.

The basic idea of the Propagation-Separation Approach is related to Lepski's method [Lepskiĭ, 1990]. In [Lepski et al., 1997; Lepski and Spokoiny, 1997], this was adapted to kernel based estimation of functions with inhomogeneous smoothness. It aims to stop smoothing at structural borders in order to avoid the increasing estimation bias due to smoothing across discontinuities. For this purpose, the method chooses the largest location bandwidth such that the associated pointwise estimator does not considerably differ from the corresponding estimates with smaller bandwidths. This is ensured by requiring a non-empty intersection of all corresponding confidence intervals. Basically, the adaptive weights of the Propagation-Separation Approach attempt to provide an intrinsic stopping criterion which resembles the model selection by Lepski's method. Here, the statistical penalty is used as a test statistic, that tests whether the estimator $\hat{\theta}_j^{(k-1)}$ lies in the confidence interval of the estimator $\hat{\theta}_i^{(k-1)}$ from the last iteration step. Hence, the adaptive weights depend, for the respective iteration step, on the estimation results of all design points within the considered neighborhood, while Lepski's method proceeds independently at each point, considering the estimates of all previous steps.

Particularly in image denoising, there are several methods that use weighted means or weighted likelihood estimators with signal-dependent weights. In the review by Katkovnik et al. [2010], the authors give a brief overview. There are several approaches that can be considered as special cases of the Propagation-Separation Approach with a single iteration step, such as the Yaroslavsky filter [Yaroslavsky, 1985], the SUSAN algorithm [Smith and Brady, 1997], or the bilateral filter [Tomasi and Manduchi, 1998]. All of them suffer from a lack of robustness when comparing noisy observations in single points, see also Buades et al. [2005a,b]. In contrast, the Propagation-Separation Approach benefits from the iteratively updated weights.

The non-local means algorithm [Buades et al., 2005a,b; Katkovnik et al., 2010] improves robustness by comparing not only the signals in single points, but in a whole neighborhood. This requires a sufficient self-redundancy of the considered image in order to ensure a sufficient number of similar neighborhoods. Additionally, in order to improve adaptation and accelerate computation a non-local means algorithm by blocks was proposed. Here, overlapping neighborhoods provide several estimators for a single point, which are averaged to achieve the final estimator. This aggregation can yield a drastically improved estimate compared to the underlying single-window estimates. Katkovnik et al. [2010] reported several proposals for further improvement of the original or the block-wise non-local means algorithm. Recently, Deledalle et al. [2009] presented an iterative method for weighted maximum likelihood estimation, combining the patch-based adaptation of the non-local means algorithm with the general setting and the iterative procedure of the Propagation-Separation Approach. Despite this close relation to the Propagation-Separation Approach, comparison was only reported with respect to other smoothing methods.

2.4 Some auxiliary results

In our study, we will use some results by Polzehl and Spokoiny [2006], which do not base on the memory step or the problematic Assumption (S0). Figure 2.3 illustrates the interdependence of the theoretical results in [Polzehl and Spokoiny, 2006] and the application of Assumption (S0). Here, we distinguish between general results and results with respect to the original Propagation-Separation Approach. In the following, we will only use Lemma 5.2, Theorem 2.1, and Theorem 2.2 from Polzehl and Spokoiny [2006], see Figure 2.3. The former is a direct consequence of Assumption A1 (page 10). It is given in Lemma 2.8 (page 13). The two theorems consider the non-adaptive estimator in Equation (2.6) (page 13). The first one is given in Theorem 2.13 (page 14) and the second one will be stated in Theorem 2.24, at the end of this section. It extends Theorem 2.13 to inhomogeneous parameter functions. Both theorems, 2.13 and 2.24, follow from a more general result, Theorem 2.19, that was stated in [Polzehl and Spokoiny, 2006, Thm. 6.1]. The corresponding proofs are given in Section 2.5.

First, we give some details concerning the applicability of the technical Lemma 2.8 and the related Equation (2.2) (page 11). The addressed challenges appear in [Polzehl and Spokoiny, 2006] as well, although this is not explicitly mentioned there. For several results, we will apply Equation (2.2) and Lemma 2.8 not only with respect to the true parameters $\{\theta_i\}_i$, but as well with respect to the transformed observations $\{T(Y_i)\}_i$ or the associated estimates $\{\tilde{\theta}_i^{(k)}\}_i$, $k \in \{0,...,k^*\}$. There, we will need that $T(Y_i) \in \Theta_\varkappa$ or at least $\tilde{\theta}_i^{(k)} \in \Theta_\varkappa$ for all $i \in \{1,....,n\}$, the respective iteration step $k \in \{0,...,k^*\}$, and some constant $\varkappa \geq 1$, where $\varkappa$ and $\Theta_\varkappa$ are as in Notation 2.6. However, if $\varkappa > 1$ and $\mathbb{P}_\theta$ has unbounded support, this cannot be satisfied with probability one. Therefore, we will restrict our analysis to the favorable realizations $\{T(Y_i) \in \Theta_\varkappa \text{ for all } i\}$, and we will quantify the probability of its complementary set. For every $\varkappa$, we will use the most convenient choice of the set $\Theta_\varkappa$. We will restrict the range of $\theta(.)$ by the subset $\Theta^* \subseteq \Theta$, which may influence the respective choice of $\Theta_\varkappa$ and, as a consequence, the corresponding value $p_\varkappa$ which we introduce now.

Notation 2.16. *We fix a subset $\Theta^* \subseteq \Theta$ and a constant $\varphi_0 \geq 0$. Then, we recall Notation 2.6, and we choose $\varkappa \geq 1$ sufficiently large such that $\Theta^* \subseteq \Theta_\varkappa$. The function $\mathfrak{p}_\varkappa : (\Theta^*)^n \to [0,1]$ maps to the probability of the event $T(Y_i) \notin \Theta_\varkappa$ for some i, where $\Theta_\varkappa$ is chosen such that $\mathfrak{p}_\varkappa$ is minimal. More precisely, we set*

$$\mathfrak{p}_\varkappa(\{\theta_i\}_{i=1}^n) := \inf\{\mathbb{P}(\exists i \in \{1,...,n\} : T(Y_i) \notin \Theta_\varkappa) : Y_i \sim \mathbb{P}_{\theta_i}, \{\theta_i\}_{i=1}^n \in (\Theta_\varkappa)^n\}.$$

Furthermore, we consider the worst choice of $\{\theta_i\}_{i=1}^n \in (\Theta^*)^n$ with bounded Kullback-Leibler divergence via

$$p_\varkappa \quad := \quad \sup\{\mathfrak{p}_\varkappa(\{\theta_i\}_{i=1}^n) : \{\theta_i\}_{i=1}^n \in (\Theta^*)^n \text{ and } \max_{i,j} \mathcal{KL}(\theta_i, \theta_j) \leq \varphi_0^2\}.$$

The probability $p_\varkappa$ is the smaller the larger we choose $\varkappa \geq 1$. The following example illustrates the trade-off between $\varkappa$ and $p_\varkappa$. In practice, the consequences are attenuated since the effective values of $\varkappa$ and $p_\varkappa$ may be much smaller than the global ones.

Example 2.17.

- For Gaussian and log-normal distributed observations with $\mathcal{P} = \{\mathcal{N}(\theta, \sigma^2)\}_{\theta \in \Theta}$ and $\mathcal{P} = \{\log\mathcal{N}(\theta, \sigma^2)\}_{\theta \in \Theta}$, respectively, it holds $I(\theta) = 1/\sigma^2$, leading to $\varkappa = 1$. In this case, Equation (2.2) and Lemma 2.8 hold for every subset $\Theta' \subseteq \Theta$ without the restriction to compact sets, and we get $p_\varkappa = 0$. This is the optimal scenario.

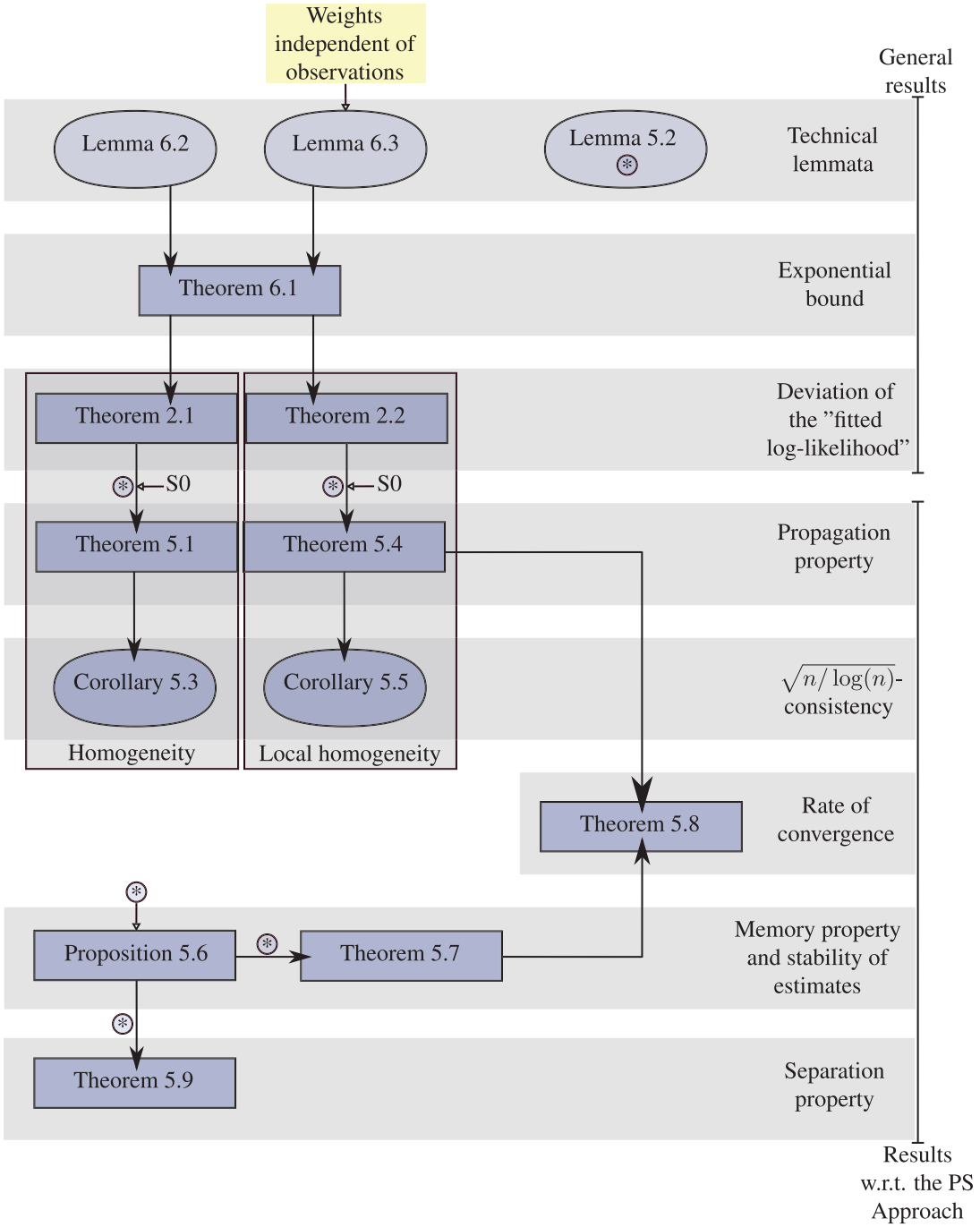

Figure 2.3: Structure of the original study by Polzehl and Spokoiny [2006]. The repeated application of Lemma 5.2 (see the first line) is emphasized by the star $(*)$. Proposition 5.6 establishes the memory property. Hence, Theorems 5.7, 5.8, and 5.9 rely on the memory step, which we omit in the simplified algorithm (page 18). The propagation property and, as a consequence, the rate of convergence require the problematic Assumption (S0).

- For the Gamma, Erlang, scaled chi-squared, exponential, Rayleigh, Weibull, and Pareto distributions, see Table 2.1 (page 12), it holds after reparametrization $I(\theta) = 1/\theta^2$. This leads to large values of $\varkappa$ and $p_\varkappa$. More precisely, every $\varkappa \geq 1$ implies that $\Theta_\varkappa = [a, \varkappa a]$ with an appropriate choice $a > 0$ which minimizes $p_\varkappa$. For $\mathcal{P} = \{\mathrm{Exp}(1/\theta)\}_{\theta\in\Theta}$ and $Y \sim \mathbb{P}_\theta$, it follows
$$\mathbb{P}(Y \in \Theta_\varkappa) = e^{-a/\theta} - e^{-\varkappa a/\theta}.$$
For $\varphi_0 = 0$ and $\Theta^* = \{\theta\}$, this depends on the explicit choices of θ and $\Theta_\varkappa$ via the quotient a/θ only. Hence, we get by maximization of $\mathbb{P}(Y \in \Theta_\varkappa)$ with respect to a/θ for every θ the most favorable $a > 0$ and consequently the most favorable choice of $\Theta_\varkappa$. Then, for each value of $\varkappa$, the probability $p_\varkappa$ equals $1 - \mathbb{P}(Y \in \Theta_\varkappa)^n$, where $\mathbb{P}(Y \in \Theta_\varkappa)$ is given in Table 2.2. This probability $p_\varkappa$ increases exponentially with n.

$\varkappa$	5	8	20	50	100
$\mathbb{P}(Y \in \Theta_\varkappa)$	0.535	0.65	0.811	0.905	0.945
a/θ	0.402	0.297	0.158	0.08	0.047

Table 2.2: Maximization of $\mathbb{P}(Y \in \Theta_\varkappa)$ with respect to a/θ for varying values of $\varkappa$.

Remark 2.18. Alternatively, we could modify the algorithm, replacing the adaptive estimator (2.7) in Algorithm 2 (page 18) by

$$\tilde{\theta}_i^{(k)} := \operatorname*{argmin}_{\theta' \in \Theta_\varkappa} |\theta' - \sum_{j=1}^n \tilde{w}_{ij}^{(k)} T(Y_j)/\tilde{N}_i^{(k)}|.$$

This projects the adaptive estimator into the set $\Theta_\varkappa$. Here, it might be advantageous to decrease the probability of $\overline{\theta}_i^{(0)} \notin \Theta_\varkappa$ by choosing the initial bandwidth $h^{(0)}$ such that the neighborhood $U_i^{(0)} := \{X_j \in \mathcal{X} : \overline{w}_{ij}^{(0)} > 0\}$ contains more design points than X_i, $U_i^{(0)} \neq \{X_i\}$, for each $i \in \{1, ..., n\}$. Else, the projection may change the impact of the statistical penalty in later iteration steps, leading to slightly shifted estimators. On the other hand, initialization with $U_i^{(0)} = \{X_i\}$, $i \in \{1, ..., n\}$, avoids smoothing among distinct homogeneous regions before adaptation starts. The projection complicates the theoretical analysis since the effect of the projection would need to be quantified. Therefore, we prefer the first approach, that is the restriction to the favorable realizations.

Next we recall an exponential bound by Polzehl and Spokoiny [2006, Thm. 6.1], from which the later on applied Theorems 2.13 and 2.24 (pages 14 and 25) follow as special cases.

Theorem 2.19. *Suppose Assumption A1, and reparametrize $v := C(\theta)$ and $D(v) := B(\theta)$. Furthermore, let $\overline{W}_i := \{\overline{w}_{ij}\}_{j=1}^n \in [0,1]^n$ be a weighting scheme, and consider the corresponding MLE $\overline{\theta}_i$ in Equation* (2.6) *and its expectation $\breve{\theta}_i := \mathbb{E}\overline{\theta}_i = \sum_j \overline{w}_{ij}\theta_j/\overline{N}_i$. We set $q(u|v) := \mathcal{KL}(v, v+u)$ and define, for a given constant $z \geq 0$ and $\breve{v}_i = C(\breve{\theta}_i)$, the set*

$$\mathcal{U}\left(\overline{W}_i, z\right) := \left\{u \in \mathbb{R} : \int_0^u xD''(\breve{v}_i + x)dx = z/\overline{N}_i\right\},$$

where it holds $\int_0^u xD''(\breve{v}_i + x)dx = \mathcal{KL}(\breve{v}_i + u, \breve{v}_i)$. Finally, we assume the existence of some constant $\alpha \geq 0$ such that

$$q(\mu u \overline{w}_{ij}|v_j) \leq (1+\alpha)\mu^2 \overline{w}_{ij} q(u|\breve{v}_i), \qquad j = 1, ..., n, \tag{2.11}$$

for $\mu := (1+\alpha)^{-1} \in (0,1]$ and all $u \in \mathcal{U}\left(\overline{W}_i, z\right)$. Then, we get

$$\mathbb{P}\left(\overline{N}_i \mathcal{KL}\left(\overline{\theta}_i, \mathbb{E}\overline{\theta}_i\right) > z\right) \leq 2e^{-z/(1+\alpha)}.$$

Remark 2.20. Polzehl and Spokoiny [2006] assumed the sufficient statistic T in Assumption A1 to be the identity map. Fortunately, the recalled results depend on the probability distribution and consequently on T via the Kullback-Leibler divergence only. This ensures with Lemma 2.5 (3) that the choice of T does not have any effect, and the original results remain valid.

Finally, we extend Theorem 2.13 to parameter functions with bounded variability. In the corresponding proof, Polzehl and Spokoiny [2006] used Equation (2.2) (page 11). Although not stated in [Polzehl and Spokoiny, 2006, Thm. 2.2], this requires a restriction to the favorable realizations. In order to quantify the probability of the complementary set, we proceed in an analogous manner as in Notation 2.16. We consider a different set of realizations, whose definition will be motivated in the proof of Theorem 2.24 (page 30).

Notation 2.21. *Recall Notation 2.6. We fix a subset $\Theta^* \subseteq \Theta$ and a constant $\varphi_0 \geq 0$. Let $\varkappa \geq 1$ be sufficiently large such that $\Theta^* \subseteq \Theta_\varkappa$. Then, for every $i \in \{1,...,n\}$, we consider the MLE $\overline{\theta}_i$ in Equation* (2.6) *with weighting scheme $\overline{W}_i := \{\overline{w}_{ij}\}_{j=1}^n \in [0,1]^n$. The function $\breve{\mathfrak{p}}_\varkappa : (\Theta^*)^n \to [0,1]$ is given by*

$$\begin{aligned}\breve{\mathfrak{p}}_\varkappa(\{\theta_i\}_{i=1}^n) \;&:=\; \inf\{\mathbb{P}\left(\exists i,j \in \{1,...,n\} : C^{-1}[C(\theta_j) + C(\overline{\theta}_i) - C(\mathbb{E}\overline{\theta}_i)] \notin \Theta_\varkappa\right) : \\ &\qquad Y_i \sim \mathbb{P}_{\theta_i}, \{\theta_i\}_{i=1}^n \in (\Theta_\varkappa)^n\}.\end{aligned}$$

We consider the worst choice of $\{\theta_i\}_{i=1}^n \in (\Theta^*)^n$ with bounded Kullback-Leibler divergence via

$$\breve{p}_\varkappa \;:=\; \sup\left\{\breve{\mathfrak{p}}_\varkappa(\{\theta_i\}_{i=1}^n) : \{\theta_i\}_{i=1}^n \in (\Theta^*)^n \text{ and } \max_{i,j} \mathcal{KL}(\theta_i,\theta_j) \leq \varphi_0^2\right\}.$$

For every $i \in \{1,...,n\}$, let the weighting scheme $\overline{W}_i := \{\overline{w}_{ij}\}_{j=1}^n \in [0,1]^n$ be given as $\overline{w}_{ii} = 1$ and $\overline{w}_{ij} = 0$ for all $i \neq j$. Then, it holds $\overline{\theta}_i = T(Y_i)$ for every i, and we set $\breve{p}_{\varkappa,0} := \breve{p}_\varkappa$ in order to distinguish the specific weighting scheme.

Example 2.22. For Gaussian and log-normal distributed observations, it holds $\breve{p}_\varkappa = 0$ since $\varkappa = 1$ for every set $\Theta_\varkappa \subseteq \Theta$. For the Gamma and its related distributions, the probability $\breve{p}_\varkappa$ may be large, and it increases with decreasing values of $\varkappa$ as well as with increasing sample sizes.

The probabilities $\breve{p}_{\varkappa,0}$ in Notation 2.21 and $p_\varkappa$ in Notation 2.16 are closely related.

Lemma 2.23. *Suppose Assumption A1 and the setting of Notation 2.21. Then, it holds*

$$\begin{aligned}\max_i \overline{\theta}_i \;&\leq\; \max_{i,j} C^{-1}\left[C(\theta_j) + C(\overline{\theta}_i) - C(\mathbb{E}\overline{\theta}_i)\right] \\ \text{and} \qquad \min_i \overline{\theta}_i \;&\geq\; \min_{i,j} C^{-1}\left[C(\theta_j) + C(\overline{\theta}_i) - C(\mathbb{E}\overline{\theta}_i)\right].\end{aligned}$$

Furthermore, we have

$$\breve{\Omega}_\varkappa := \bigcap_{i,j=1}^n \left\{C^{-1}\left[C(\theta_j) + C(T(Y_i)) - C(\theta_i)\right] \in \Theta_\varkappa\right\} \subseteq \bigcap_{i=1}^n \{T(Y_i) \in \Theta_\varkappa\} =: \Omega_\varkappa,$$

and, as a consequence, we get $\breve{p}_{\varkappa,0} \geq p_\varkappa$, where $p_\varkappa$ is as in Notation 2.16.

Theorem 2.24. *Suppose Assumption A1, and fix a subset $\Theta^* \subseteq \Theta$ and a constant $\varphi_0 \geq 0$ such that $\{\theta_i\}_{i=1}^n \in (\Theta^*)^n$ and $\max_{i,j} \mathcal{KL}(\theta_i,\theta_j) \leq \varphi_0^2$. Moreover, recall Notation 2.6, and let $\varkappa \geq 1$ be sufficiently large such that $\Theta^* \subseteq \Theta_\varkappa$. Finally, let $\overline{W}_i := \{\overline{w}_{ij}\}_{j=1}^n \in [0,1]^n$ denote a weighting scheme, and recall the corresponding quantities $\overline{\theta}_i$ and $\overline{N}_i$ in Equation* (2.6)*. Then, for each $i \in \{1,...,n\}$ and every $z > 0$, it holds*

$$\mathbb{P}\left(\overline{N}_i \mathcal{KL}(\overline{\theta}_i, \mathbb{E}\overline{\theta}_i) > z\right) \leq 2e^{-z/\varkappa^2} + \breve{p}_\varkappa,$$

where $\breve{p}_\varkappa$ is as in Notation 2.21.

2.5 Proofs

Proof of Lemma 2.5.

1. For every $\theta \in \Theta$, it follows from Assumption A1 that

$$\log p(Y,\theta) = \log\left(p(Y)\right) + T(Y)C(\theta) - B(\theta).$$

Due to $B'(\theta) = \theta C'(\theta)$, this leads to

$$\begin{aligned} \tfrac{\partial}{\partial\theta}\log p(Y,\theta) &= T(Y)C'(\theta) - B'(\theta) = C'(\theta)[T(Y)-\theta] \\ \text{and} \quad \tfrac{\partial^2}{\partial\theta^2}\log p(Y,\theta) &= C''(\theta)[T(Y)-\theta] - C'(\theta). \end{aligned}$$

Hence, Notation 2.4 and Equation (2.1) yield

$$I(\theta) = -C''(\theta)\mathbb{E}_\theta\left[T(Y)-\theta\right] + C'(\theta) = C'(\theta).$$

2. The compactness of Θ' and $C \in C^2(\Theta,\mathbb{R})$ ensure by the extreme value theorem that the derivative C' attains its minimum and maximum. Then, for every $\theta_1,\theta_2 \in \Theta'$, we get

$$\frac{I(\theta_1)}{I(\theta_2)} = \frac{C'(\theta_1)}{C'(\theta_2)} \le \frac{\max\{C'(\theta) : \theta \in \Theta'\}}{\min\{C'(\theta) : \theta \in \Theta'\}} =: \varkappa^2 \in [1,\infty)$$

since $C' > 0$.

3. For all $\theta,\theta' \in \Theta$, we observe that

$$\begin{aligned} \mathcal{KL}\left(\theta,\theta'\right) &= \mathbb{E}_\theta\left[\log p(Y,\theta) - \log p(Y,\theta')\right] \\ &= \theta C(\theta) - B(\theta) - \theta C(\theta') + B(\theta'), \end{aligned}$$

which leads to Equation (2.3). Next we consider the reparametrization $v := C(\theta)$ and $D(v) := B(\theta)$, satisfying $D'(v) = C^{-1}(v) = \theta$ and $D''(v) = 1/I(C^{-1}(v)) = 1/I(\theta)$. Then, the quadratic expression follows from Taylor's Theorem, where the remainder is given in Lagrange form,

$$\begin{aligned} \mathcal{KL}\left(\theta,\theta'\right) &= C^{-1}(v)(v-v') + D(v') - D(v) \qquad (2.12) \\ &= C^{-1}(v)(v-v') + D'(v)(v'-v) + D''(v_0)(v'-v)^2/2 \\ &= \frac{1}{2I(\theta_0)}\left[C(\theta') - C(\theta)\right]^2 \qquad (2.13) \\ &= \frac{1}{2I(\theta_0)}\left[C'(\theta_*)(\theta'-\theta)\right]^2 \\ &= \frac{\left[I(\theta_*)\right]^2}{2I(\theta_0)}(\theta'-\theta)^2 \end{aligned}$$

with θ_* between θ and θ' and v_0 between v and v', leading by the monotonicity of C to θ_0 between θ and θ'. Finally, the convexity of the Kullback-Leibler divergence with respect to the first argument is satisfied since the second derivative is non-negative

$$\frac{\partial^2}{\partial\theta^2}\mathcal{KL}\left(\theta,\theta'\right) = \frac{\partial}{\partial\theta}\left[C(\theta) - C(\theta')\right] = C'(\theta) > 0.$$

□

Proof of Lemma 2.8. The proof follows the same ideas as [Polzehl and Spokoiny, 2006, Lem. 5.2]. We use the reparametrization $v := C(\theta)$ and $D(v) := B(\theta)$ as introduced in the proof of Lemma 2.5 (3). This leads for all $\theta_1, \theta_2, \theta_3$ with v^*, v_1^*, v_3^* appropriate to

$$\begin{aligned}
\mathcal{KL}^{1/2}(\theta_1, \theta_3) &\overset{\text{Eq. (2.13)}}{=} \sqrt{\tfrac{1}{2}D''(v^*)}\,|v_1 - v_3| \\
&\leq \sqrt{\tfrac{1}{2}D''(v^*)}\,|v_1 - v_2| + \sqrt{\tfrac{1}{2}D''(v^*)}\,|v_2 - v_3| \\
&\overset{\text{Eq. (2.2)}}{\leq} \varkappa\sqrt{\tfrac{1}{2}D''(v_1^*)}\,|v_1 - v_2| + \varkappa\sqrt{\tfrac{1}{2}D''(v_3^*)}\,|v_2 - v_3| \\
&\overset{\text{Eq. (2.13)}}{=} \varkappa\left[\mathcal{KL}^{1/2}(\theta_1, \theta_2) + \mathcal{KL}^{1/2}(\theta_2, \theta_3)\right]
\end{aligned}$$

and analogously for any sequence $\theta_0, \theta_1, ..., \theta_m \in \Theta_\varkappa$. □

Proof of Lemma 2.10. We use the notations $\overline{N}_i = \sum_{j=1}^n \overline{w}_{ij}$ and $S_i := \sum_{j=1}^n \overline{w}_{ij} T(Y_j)$.

1. Assumption A1 yields, for every $\theta \in \Theta$, that

$$L(\overline{W}_i, \theta) = \sum_{j=1}^n \overline{w}_{ij} \left[\log\left(p(Y_j)\right) + T(Y_j)C(\theta) - B(\theta)\right],$$

   and we have

$$\frac{\partial}{\partial\theta} L(\overline{W}_i, \theta) = C'(\theta)\left[S_i - \theta\overline{N}_i\right],$$

   where we used that $B'(\theta) = \theta C'(\theta)$. Since $C'(\theta) > 0$ for all $\theta \in \Theta$, the only extremum of $L(\overline{W}_i, \theta)$ is at the point $\overline{\theta}_i = S_i/\overline{N}_i$. Additionally, we observe that the monotonicity of the function C ensures that the second derivative

$$\frac{\partial^2}{\partial\theta^2} L(\overline{W}_i, \theta) = C''(\theta)\left[S_i - \theta\overline{N}_i\right] - C'(\theta)\overline{N}_i$$

   is negative at $\theta = S_i/\overline{N}_i$. Hence, the log-likelihood function has a unique global maximum at $\theta = \overline{\theta}_i$, and the weighted mean $\overline{\theta}_i$ equals the maximum likelihood estimator θ_i^{MLE}.
2. Recall that the observations $\{Y_i\}_{i=1}^n$ are statistically independent. Therefore, the variance of the weighted MLE equals

$$\begin{aligned}
\operatorname{Var}\left(\theta_i^{\text{MLE}}\right) &= \mathbb{E}_\theta\left[\left(\sum_{j=1}^n \overline{w}_{ij}[T(Y_j) - \theta_j]/\overline{N}_i\right)^2\right] \\
&= \sum_{j=1}^n \overline{w}_{ij}^2 \mathbb{E}_\theta\left[(T(Y_j) - \theta_j)^2\right]/\overline{N}_i^2 \quad = \quad \sigma^2 \sum_{j=1}^n \overline{w}_{ij}^2/\overline{N}_i^2.
\end{aligned}$$

   Finally, $\{w_{ij}\}_j \in [0,1]^n$ implies that $\sum_j \overline{w}_{ij}^2/\overline{N}_i^2 \leq \sum_j \overline{w}_{ij}/\overline{N}_i^2 = 1/\overline{N}_i$.

□

Proof of Lemma 2.12. Let S_i and $\overline{N}_i$ be as in the proof of Lemma 2.10, and recall that $S_i = \overline{N}_i\overline{\theta}_i$ as established in Lemma 2.10 (1). Then, for every $\theta \in \Theta$, we get

$$\begin{aligned}
L(\overline{W}_i, \overline{\theta}_i, \theta) &= \sum_{j=1}^n \overline{w}_{ij}\left[T(Y_j)C(\overline{\theta}_i) - B(\overline{\theta}_i) - T(Y_j)C(\theta) + B(\theta)\right] \\
&= S_i\left[C(\overline{\theta}_i) - C(\theta)\right] - \overline{N}_i\left[B(\overline{\theta}_i) - B(\theta)\right] \quad = \quad \overline{N}_i \mathcal{KL}\left(\overline{\theta}_i, \theta\right),
\end{aligned}$$

where we used Equation (2.3) in Lemma 2.5 (3). □

The proofs of Theorems 2.13 and 2.24 are based on Theorem 2.19. Therefore, we prove the latter first. For this purpose, we recall the exponential Chebyshev inequality, see for example Durrett [2010, Thm 1.6.4, page 28].

Lemma 2.25 (Exponential Chebyshev inequality)**.** *Let Z denote a random variable with finite exponential expected value. Then, it holds $\mathbb{P}(Z > z) \leq e^{-\mu z}\,\mathbb{E}[e^{\mu Z}]$ for all $z > 0$ and every $\mu > 0$.*

Proof of Theorem 2.19. Recall that the function C is strictly monotonic increasing. Therefore, the parameter v is uniquely defined, and it holds by definition of the weighted MLE and the fitted log-likelihood that

$$\overline{v}_i = \operatorname*{argsup}_{v \in \mathrm{range}(C)} L\left(\overline{W}_i, v\right) \quad \text{and} \quad L\left(\overline{W}_i, \overline{v}_i, v'\right) = \sup_{v \in \mathrm{range}(C)} L\left(\overline{W}_i, v\right) - L\left(\overline{W}_i, v'\right)$$

for $v' \in \mathrm{range}(C)$. We use the reparametrization $v := C(\theta)$ and $D(v) := B(\theta)$, see the proof of Lemma 2.5 (3). For all $v \in \mathrm{range}(C)$, Lemma 2.12 shows with $D'(\overline{v}_i) = \overline{\theta}_i = S_i/\overline{N}_i$, where $S_i = \sum_j \overline{w}_{ij} T(Y_j)$, that

$$L\left(\overline{W}_i, \overline{v}_i, v\right) \overset{\text{Lem. 2.12}}{=} \overline{N}_i \mathcal{KL}\left(\overline{v}_i, v\right) \overset{\text{Eq. (2.12)}}{=} S_i(\overline{v}_i - v) - \overline{N}_i\left(D(\overline{v}_i) - D(v)\right).$$

Then, we consider the function $f(u) := \left[z + \overline{N}_i\left(D(\breve{v}_i + u) - D(\breve{v}_i)\right)\right]/u$, which attains its minimum at some point $u \in \mathcal{U}(\overline{W}_i, z)$, since

$$\begin{aligned} 0 = f'(u) &\iff 0 = \left[\overline{N}_i u D'(\breve{v}_i + u) - z - \overline{N}_i\left(D(\breve{v}_i + u) - D(\breve{v}_i)\right)\right]/u^2 \\ &\iff z/\overline{N}_i = u D'(\breve{v}_i + u) - \left(D(\breve{v}_i + u) - D(\breve{v}_i)\right) \end{aligned}$$

and

$$\begin{aligned} \mathcal{KL}\left(\breve{v}_i + u, \breve{v}_i\right) &= u D'(\breve{v}_i + u) - \left(D(\breve{v}_i + u) - D(\breve{v}_i)\right) \\ &= \int_0^u x D''(\breve{v}_i + x) + D'(\breve{v}_i + x)dx - \int_0^u D'(\breve{v}_i + x)dx \\ &= \int_0^u x D''(\breve{v}_i + x)dx. \end{aligned}$$

The same holds for the function $f_*(u) := -f(u)$. Hence, for all $z \geq 0$ and some appropriate values $u_1, u_2 \in \mathcal{U}(\overline{W}_i, z)$, it follows in the same manner as in the proof of [Polzehl and Spokoiny, 2006, Lem. 6.2] that

$$\begin{aligned} &\left\{L\left(\overline{W}_i, \overline{v}_i, \breve{v}_i\right) > z\right\} \\ = &\left\{\sup_{v \in \mathrm{range}(C)} \left[S_i(v - \breve{v}_i) - \overline{N}_i\left(D(v) - D(\breve{v}_i)\right)\right] > z\right\} \\ \subseteq &\left\{S_i > \inf_{v > \breve{v}_i} \frac{z + \overline{N}_i\left(D(v) - D(\breve{v}_i)\right)}{v - \breve{v}_i}\right\} \cup \left\{-S_i > \inf_{v < \breve{v}_i} \frac{z + \overline{N}_i\left(D(v) - D(\breve{v}_i)\right)}{\breve{v}_i - v}\right\} \\ \subseteq &\left\{S_i > \frac{z + \overline{N}_i\left(D(\breve{v}_i + u_1) - D(\breve{v}_i)\right)}{u_1}\right\} \cup \left\{-S_i > \frac{z + \overline{N}_i\left(D(\breve{v}_i + u_2) - D(\breve{v}_i)\right)}{-u_2}\right\} \\ \subseteq &\left\{L\left(\overline{W}_i, \breve{v}_i + u_1, \breve{v}_i\right) > z\right\} \cup \left\{L\left(\overline{W}_i, \breve{v}_i + u_2, \breve{v}_i\right) > z\right\}, \end{aligned}$$

For every $u \in \mathcal{U}(\overline{W}_i, z)$, we get from $D'(v) = \theta$ that

$$\begin{aligned} L\left(\overline{W}_i, \breve{v}_i + u, \breve{v}_i\right) &= u\left[S_i - \overline{N}_i \breve{\theta}_i\right] - \overline{N}_i\left[D(\breve{v}_i + u) - D(\breve{v}_i) - u D'(\breve{v}_i)\right] \\ &= u\left[S_i - \overline{N}_i \breve{\theta}_i\right] - \overline{N}_i q(u|\breve{v}_i). \end{aligned}$$

Then, for $\mu > 0$, Lemma 2.25 implies with the statistical independence of the observations that

$$\log \mathbb{P}\left(L(\overline{W}_i, \breve{v}_i + u, \breve{v}_i) > z\right) \quad \leq \quad -\mu z - \mu \overline{N}_i q(u|\breve{v}_i) - \mu u \overline{N}_i \breve{\theta}_i + \sum_{j=1}^{n} \log \mathbb{E}\left[e^{\mu u \overline{w}_{ij} T(Y_j)}\right].$$

For every $v, v^* \in \mathrm{range}(C)$, it holds $(d\mathbb{P}_v / d\mathbb{P}_{v^*})(y) = e^{T(y)[v-v^*]-[D(v)-D(v^*)]}$, yielding, for every j with $a_j := \mu u \overline{w}_{ij}$, that

$$\begin{aligned}
\mathbb{E}\left[e^{a_j T(Y_j)}\right] &= \int e^{a_j T(y)} \frac{d\mathbb{P}_{v_j}}{d\mathbb{P}_{v^*}}(y) d\mathbb{P}_{v^*}(y) \\
&= e^{-[D(v_j)-D(a_j+v_j)]} \int e^{T(y)[(a_j+v_j)-v^*]-[D(a_j+v_j)-D(v^*)]} d\mathbb{P}_{v^*}(y) \\
&= e^{-[D(v_j)-D(a_j+v_j)]} \int \frac{d\mathbb{P}_{(a_j+v_j)}}{d\mathbb{P}_{v^*}}(y) d\mathbb{P}_{v^*}(y) \\
&= e^{q(a_j|v_j)+a_j D'(v_j)}.
\end{aligned}$$

Finally, $\mu = (1+\alpha)^{-1}$ and Equation (2.11) (page 24) lead to

$$\begin{aligned}
& \log \mathbb{P}\left(L(\overline{W}_i, \breve{v}_i + u, \breve{v}_i) > z\right) \\
\leq \quad & -\mu z - \mu \overline{N}_i q(u|\breve{v}_i) - \mu u \overline{N}_i \breve{\theta}_i + \sum_{j=1}^{n} \left[q(\mu u \overline{w}_{ij}|v_j) + \mu u \overline{w}_{ij} D'(v_j)\right] \\
\leq \quad & -\mu z - \mu \overline{N}_i q(u|\breve{v}_i) - \mu u \overline{N}_i \breve{\theta}_i + (1+\alpha)\mu^2 \overline{N}_i q(u|\breve{v}_i) + \mu u \sum_{j=1}^{n} \overline{w}_{ij} \theta_j \\
= \quad & -z/(1+\alpha),
\end{aligned}$$

which terminates the proof. □

Proof of Theorem 2.13. Let $q(u|v) = \mathcal{KL}(v, v+u)$ be as in Theorem 2.19 and hence $q(0|v) = 0$. We justify the assertions of Theorem 2.19. Since $\theta(.) \equiv \theta$, it holds $v_j = \breve{v}_i$ for all $i, j \in \{1, ..., n\}$. Recall that the Kullback-Leibler divergence is convex with respect to the first argument. For every weighting scheme $\overline{W}_i \in [0,1]^n$ and all $u \in \mathcal{U}(\overline{W}_i, z)$, this yields

$$q(\overline{w}_{ij} u|v_j) \leq \overline{w}_{ij} q(u|\breve{v}_i), \qquad j = 1, ..., n.$$

Hence, in the setting of Theorem 2.13, Equation (2.11) (page 24) is satisfied with $\alpha = 0$ and $\mu = 1$, and the assertion follows from Theorem 2.19 with $\theta = \mathbb{E}\overline{\theta}$. □

Proof of Lemma 2.23. Recall that the function C is strictly monotonic increasing. Additionally, the estimator $\overline{\theta}_i$ is a weighted mean, and we get

$$\max_j C(\theta_j) \geq C(\mathbb{E}\overline{\theta}_i) \quad \text{and} \quad \min_j C(\theta_j) \leq C(\mathbb{E}\overline{\theta}_i)$$

for every $i \in \{1, ..., n\}$. Therefore, it holds

$$\begin{aligned}
& \max_{i,j} C^{-1}\left[C(\theta_j) + C(\overline{\theta}_i) - C(\mathbb{E}\theta_i)\right] \\
= \quad & C^{-1}\left[\max_j C(\theta_j) + \max_i \{C(\overline{\theta}_i) - C(\mathbb{E}\overline{\theta}_i)\}\right] \geq C^{-1}\left[\max_i C(\overline{\theta}_i)\right] = \max_i \overline{\theta}_i
\end{aligned}$$

and analogously

$$\min_{i,j} C^{-1}\left[C(\theta_j)+C(\overline{\theta}_i)-C(\mathbb{E}\theta_i)\right] \leq \min_i \overline{\theta}_i.$$

The other assertions follow as direct consequences by the definition of $\breve{p}_{\varkappa,0}$ and the corresponding weighting scheme. □

Proof of Theorem 2.24. Let $q(u|v) = \mathcal{KL}(v, v+u)$ be as in Theorem 2.19. Equation (2.12) (page 26) provides with the Taylor expansion that

$$q(u|v) = D(v+u) - D(v) - uD'(v) = u^2 D''(v+cu)/2,$$

where the remainder is in Lagrange form, and $c \in [0,1]$ is chosen appropriately. We set $\alpha := \varkappa^2 - 1$, and recall that $\overline{w}_{ij}^2 \leq \overline{w}_{ij}$ since $\overline{w}_{ij} \in [0,1]$. For $c_1, c_2 \in [0,1]$ appropriate and all $i,j \in \{1,...,n\}$, this yields Equation (2.11) (page 24) via

$$\begin{aligned} q(\mu u \overline{w}_{ij}|v_j) &= (\mu u \overline{w}_{ij})^2 D''(v_j + c_1 \mu u \overline{w}_{ij})/2 \\ &\overset{\text{Eq. (2.2)}}{\leq} \mu^2 \overline{w}_{ij}^2 \varkappa^2 u^2 D''(\breve{v}_i + c_2 u)/2 \leq (1+\alpha)\mu^2 \overline{w}_{ij} q(u|\breve{v}_i) \end{aligned}$$

if $C^{-1}(v_j + c_1 \overline{w}_{ij}\mu u), C^{-1}(\breve{v}_i + c_2 u) \in \Theta_\varkappa$. The function C is strictly monotonic increasing, and the expectation satisfies $\breve{v}_i \in [\min_j v_j, \max_j v_j]$. It holds by assumption that $C^{-1}(v_i) \in \Theta_\varkappa$ for all $i \in \{1,...,n\}$ and, as a consequence, $C^{-1}(\breve{v}_i) \in \Theta_\varkappa$. Therefore, it suffices to ensure that $C^{-1}(v_j + u) \in \Theta_\varkappa$ for all $j \in \{1,...,n\}$ and $u \in \mathcal{U}(\overline{W}_i, z)$. The assertion of Theorem 2.19 remains valid if Equation (2.11) is only satisfied for $u \in \mathcal{U}(\overline{W}_i, z)$ with $u := \overline{v}_i - \breve{v}_i$. Hence, we restrict our analysis to the favorable realizations, where $C^{-1}(v_j + \overline{v}_i - \breve{v}_i) \in \Theta_\varkappa$ for all $i,j \in \{1,...,n\}$ and some most favorable subset $\Theta_\varkappa \subseteq \Theta$. The probability of the complementary set of realizations is bounded by the probability $\breve{p}_\varkappa$ in Notation 2.21, and we get by Theorem 2.19 that

$$\mathbb{P}\left(\overline{N}_i \mathcal{KL}(\overline{\theta}_i, \mathbb{E}\overline{\theta}_i) > z\right) \leq 2e^{-z/\varkappa^2} + \breve{p}_\varkappa,$$

which leads to the assertion. □

Chapter 3

The propagation condition

We dwell on the choice of the adaptation bandwidth λ. This specifies the amount of adaptation and is therefore crucial for the behavior of the Propagation-Separation Approach, see Algorithms 1 and 2 (pages 16 and 18). For $\lambda \to \infty$, the algorithm behaves as non-adaptive smoothing. In contrast, for a small adaptation bandwidth, that is if $\lambda \to 0$, the algorithm adapts to noise, and the pointwise adaptive estimator equals the respective observation, $\tilde{\theta}_i^{(k)} = T(Y_i)$. Hence, we aim to find a bandwidth λ which allows as much smoothing as possible within homogeneous regions, while still providing structural adaptation.

Polzehl and Spokoiny [2006, §3.5] suggested a parameter choice strategy for the adaptation bandwidth, called *propagation condition*. The basic idea is that the impact of the statistical penalty in the adaptive weights should be negligible under homogeneity, yielding almost unhindered smoothing within homogeneous regions. More precisely, the authors proposed to adjust λ by Monte-Carlo simulations in accordance with the following criterion, where an artificial data set is considered.

> "(...) the parameter λ can be selected as the minimal value of λ that, in the case of a homogeneous (parametric) model $\theta(x) \equiv \theta$, provides a prescribed probability to obtain the global model at the end of the iteration process."

Here, we formally introduce a new criterion which allows the verification of propagation and stability under (piecewise) homogeneity for the simplified procedure as given in Algorithm 2 (page 18). Additionally, it provides a better interpretability than earlier formulations, see for example Polzehl et al. [2010]. Spokoiny and Vial [2009] formulated in the context of model selection a *propagation condition* which is based on confidence intervals. Here, we consider quantiles.

The chapter is divided in three sections. First, we will present our novel propagation condition for the choice of the adaptation bandwidth, considering an artificial homogeneous setting. This choice will be justified by a detailed evaluation of its variability with respect to the fixed parameter $\theta \in \Theta$. The corresponding results in § 3.1.2 will be crucial for our subsequent study of the simplified Propagation-Separation Approach in Section 4.2. Additionally, we will give some hints concerning the application of the propagation condition in practice. Then, we will extend the homogeneous propagation condition to bounded parameter functions. This requires an increased adaptation bandwidth as we will specify in Claim 3.22. Although the iterative procedure of the algorithm impedes a definite proof, this claim suggests the existence of a bandwidth λ which is in accordance with the inhomogeneous propagation condition

in Definition 3.15. The justification of Claim 3.22 will be based on some technical results, namely Lemmas 3.19, 3.20, and Proposition 3.21. These results will deliver an insight in the homogeneous propagation condition and the impact of the Kullback-Leibler divergence, where it is based on. All proofs and the justification of Claim 3.22 are collected in the last part of the chapter.

Throughout this chapter we will consider the local exponential family model in Assumption A1 (page 10). Additionally, we will concentrate on the simplified Propagation-Separation Approach in Algorithm 2 (page 18), where the memory step is omitted. Our novel formulation of the propagation condition is motivated by two results by Polzehl and Spokoiny [2006], which we recalled in Chapter 2, namely Theorems 2.13 and 2.24.

3.1 Choice of the adaptation bandwidth

Under homogeneity, where $\theta(.) \equiv \theta$, Theorem 2.13 shows that the non-adaptive estimator in Notation 2.14 satisfies $\mathbb{P}(\overline{N}_i^{(k)} \mathcal{KL}(\overline{\theta}_i^{(k)}, \theta) > z) \leq 2e^{-z}$ for all $i \in \{1, ..., n\}$ and every iteration step $k \in \{0, ..., k^*\}$. This describes the probability of the Kullback-Leibler divergence between the non-adaptive estimator $\overline{\theta}_i^{(k)}$ and the true parameter $\theta \in \Theta$ to exceed the upper bound $z/\overline{N}_i^{(k)}$. Hence, in probability, the divergence $\mathcal{KL}(\overline{\theta}_i^{(k)}, \theta)$ decreases at least with rate $\overline{N}_i^{(k)}$. The following propagation condition is constructed to ensure a similar behavior for the adaptive estimator. Recall the relation of the adaptive and the non-adaptive estimators, that we established in Corollary 2.15.

3.1.1 Homogeneous propagation condition

For all $k \in \{0, ..., k^*\}$, $p \in (0,1)$, $\theta \in \Theta$, and $i \in \{1, ..., n\}$, Theorem 2.13 implies that

$$Z(k, p; \theta, i) := \inf \left\{ z > 0 : \mathbb{P}\left(\overline{N}_i^{(k)} \mathcal{KL}(\overline{\theta}_i^{(k)}, \theta) > z \right) \leq p \right\} \leq \ln(2/p)$$

since $p = 2e^{-z}$ if and only if $z = \ln(2/p)$. However, this is a rough bound, and $Z(k, p; \theta, i)$ can be smaller than $\ln(2/p)$. In fact, heuristic observations suggest that the function $Z(., p; \theta, i)$ is non-increasing with respect to the first argument, that is during iteration. We introduce the adaptive analog of the non-adaptive function Z.

Notation 3.1. *For every $\lambda > 0$, we consider the function*

$$\mathfrak{Z}_\lambda : \{0, ..., k^*\} \times (0,1) \times \Theta \times \{1, ..., n\} \to [0, \infty)$$

defined by

$$\mathfrak{Z}_\lambda(k, p; \theta, i) := \inf \left\{ z > 0 : \mathbb{P}\left(\overline{N}_i^{(k)} \mathcal{KL}(\tilde{\theta}_i^{(k)}(\lambda), \theta) > z \right) \leq p \right\},$$

where $\tilde{\theta}_i^{(k)}(\lambda)$ denotes the adaptive estimator at the position $X_i \in \mathcal{X}$, resulting from Algorithm 2 (page 18) with the adaptation bandwidth λ and observations $Y_j \overset{iid}{\sim} \mathbb{P}_\theta$ for all $j \in \{1, ..., n\}$ with $\theta(.) \equiv \theta$.

Note that we replaced the non-adaptive estimator $\overline{\theta}_i^{(k)}$ by the adaptive one, but not the sum of the non-adaptive weights $\overline{N}_i^{(k)}$. This is crucial for our analysis as it requires the adaptive estimator to provide a similar reduction of the Kullback-Leibler divergence as the non-adaptive one.

Definition 3.2 (Homogeneous Propagation condition)**.** *We say that the adaptation bandwidth λ is chosen in accordance with the (homogeneous) propagation condition at level $\epsilon > 0$ for $\theta \in \Theta$ if the function $\mathfrak{Z}_\lambda(.,p;\theta,i)$ in Notation 3.1 is non-increasing for all $p \in (\epsilon,1)$ and every $i \in \{1,...,n\}$.*

Remark 3.3.

- In § 3.1.2, we will evaluate the variability of the propagation condition with respect to the fixed parameter $\theta \in \Theta$. There, we will show some examples of the propagation condition, assuming a Gaussian, exponential, and Poisson distribution, see Figures 3.1, 3.2, and 3.3.
- The probability $\mathbb{P}(\overline{N}_i^{(k)} \mathcal{KL}(\tilde{\theta}_i^{(k)}(\lambda),\theta) > z)$ cannot be calculated exactly. In § 3.1.3, we will introduce an appropriate approximation which can be used in practice.
- The propagation level ϵ can be interpreted as the expected number of observations out of one hundred to which the Propagation-Separation Approach adapts under homogeneity. In practice, this number should be chosen such that the algorithm provides a certain estimation quality that is appropriate for the respective application.

In general, it is advantageous to allow as much adaptation as possible, without violating the propagation condition. Hence, the optimal choice of λ is given by the infimum over the values which are in accordance with the propagation condition. In order to ensure that $\lambda > 0$, we introduce an additional constant $\lambda_{\min} > 0$.

Notation 3.4. *Let $\lambda_{\min} > 0$ be fixed, and consider the set*

$$\Lambda(\epsilon;\theta) := \{\lambda > 0 : \mathfrak{Z}_\lambda(.,p;\theta,i) \text{ is non-increasing for } p \in (\epsilon,1) \text{ and all } i\}.$$

Then, we introduce

$$\lambda_{\text{opt}}(\epsilon,\theta;\lambda_{\min}) := \max\{\lambda_{\min}, \inf\{\lambda \in \Lambda(\epsilon;\theta)\}\}.$$

3.1.2 Justification of the propagation condition

As in previous versions, the propagation condition in Definition 3.2 is formulated with respect to some fixed parameter $\theta \in \Theta$. In practice, the parameter function $\theta(.)$ is unknown. Hence, we need to ensure that the propagation condition is satisfied for the unknown values θ_i with $i \in \{1,...,n\}$. At best, the choice of λ by the propagation condition is invariant with respect to the underlying parameter θ. The study below points out that this is the case for the Gaussian and the exponential distribution and, as a consequence, for the log-normal, Rayleigh, Weibull, and Pareto distributions. Else, we recommend to identify some *least favorable* parameter θ^* which yields a sufficiently large choice of the adaptation bandwidth λ such that the propagation condition remains valid for all values θ_i, $i \in \{1,...,n\}$.

We start our analysis by introducing a general criterion for the invariance of the composition of two functions with respect to some parameter θ.

Proposition 3.5. *Let $f : \Omega^f \to \mathbb{R}$ and $g : \Omega^g \to \mathbb{R}$ be continuously differentiable functions with open domains $\Omega^f, \Omega^g \subseteq \mathbb{R}^2$. We denote $\Omega_\theta^f := \{y : (y,\theta) \in \Omega^f\}$ and $f_\theta : \Omega_\theta^f \to \mathbb{R}$ with $f_\theta(y) := f(y,\theta)$ and analogously Ω^g and g_θ. Then, we suppose $\Omega_\theta^g \subseteq \Omega_\theta^f$ and $\left|\frac{\partial g_\theta}{\partial y}\right| > 0$, such that the composition $f_\theta \circ g_\theta^{-1} : g_\theta(\Omega_\theta^g) \to \mathbb{R}$ is well-defined. The function*

$$h(z,\theta) := f_\theta\left(g_\theta^{-1}(z)\right), \qquad z \in g_\theta(\Omega_\theta^g),$$

is invariant with respect to θ if a variable $\zeta(y,\theta)$ and functions $\tilde{f}$ and $\tilde{g}$ exist such that

$$\tilde{f}(\zeta) = f_\theta(y) \quad \textit{and} \quad \tilde{g}(\zeta) = g_\theta(y). \tag{3.1}$$

Now we are well prepared to evaluate the variability of the propagation condition with respect to the parameter θ. Recall that the estimator of the Propagation-Separation Approach is defined as a linear combination of the terms $T(Y_j)$, where the adaptive and the non-adaptive estimator only differ in the definition of the weights. Thus, we approach the problem in three steps. We start from the special case where the estimator is restricted to a single point $T(Y_j)$. Then, we consider the *non-adaptive* estimator, describing its probability density as the convolution of the respective densities corresponding to the weighted observations. Here, we take advantage of the statistical independence of the involved random variables $\overline{w}_{ij}^{(k)} T(Y_j)/\overline{N}_i^{(k)}$. In the case of the *adaptive* estimator, we cannot follow the same approach. This would require the specification of the probability distributions of the random variables $\tilde{w}_{ij}^{(k)} T(Y_j)/\tilde{N}_i^{(k)}$, where the distributions of the adaptive weights are unknown. Furthermore, these variables are not statistically independent. In order to compensate the resulting lack of a theoretical proof, we illustrate by simulations that the adaptive estimator shows almost the same behavior as the non-adaptive estimator if the propagation condition is satisfied. This suggests that the probability distribution of $\mathcal{KL}(\tilde{\theta}_i^{(k)}, \theta)$ is invariant with respect to θ if the same holds for the non-adaptive estimator. The single observation case is treated first.

Lemma 3.6. *Suppose Assumption A1, and let $\mathcal{P} = \{\mathbb{P}_\theta\}_{\theta\in\Theta}$ with $\Theta \subseteq \mathbb{R}$ be a parametric family of continuous probability distributions. Presume that $Y \sim \mathbb{P}_\theta$ with $T(Y) \in \Theta$ almost surely and that the density $f_\theta^Y = d\mathbb{P}_\theta/d\mathbb{P}$ of Y is continuously differentiable. We consider the random variable $Z := g_\theta(Y) := \mathcal{KL}(T(Y),\theta)$, and we require that g_θ is continuously differentiable and $|\frac{\partial g_\theta}{\partial y}| > 0$ almost surely. The density f_θ^Z of Z is invariant with respect to the parameter θ if a variable $\zeta(y,\theta)$ and functions $\tilde{f}$ and $\tilde{g}$ exist such that*

$$\tilde{f}(\zeta) = f_\theta^Y(y) \cdot \left| \frac{\partial g_\theta}{\partial y}(y) \right|^{-1} \quad \textit{and} \quad \tilde{g}(\zeta) = g_\theta(y).$$

This lemma yields the desired results for Gaussian and Gamma distributed observations.

Example 3.7. We use the same notations as in Lemma 3.6. In the following cases, the density of Z is invariant with respect to the parameter θ.

- $\mathcal{P} = \{\mathcal{N}(\theta,\sigma^2)\}_{\theta\in\Theta}$ with $\sigma > 0$ fixed: We know from Example 2.7 that

  $$g_\theta(y) = \mathcal{KL}(y,\theta) = \frac{(y-\theta)^2}{2\sigma^2} \quad \text{and we get} \quad \frac{\partial g_\theta}{\partial y}(y) = \frac{y-\theta}{\sigma^2}.$$

  Hence, g_θ is strictly monotonic on the open sets $V_1 = (-\infty,\theta)$ and $V_2 = (\theta,\infty)$. Additionally, it holds $\mathbb{P}_\theta(Y = \theta) = 0$. Since $f_\theta^Y(y) = \exp\left(-\frac{(y-\theta)^2}{2\sigma^2}\right)/\sqrt{2\pi\sigma^2}$, we get the invariance with respect to θ from Lemma 3.6 by setting

  $$\zeta(y,\theta) := y - \theta, \qquad \tilde{f}(\zeta) := \frac{\sigma\, e^{-\frac{\zeta^2}{2\sigma^2}}}{\zeta\sqrt{2\pi}}, \quad \text{and} \quad \tilde{g}(\zeta) := \frac{\zeta^2}{2\sigma^2}.$$

- $\mathcal{P} = \{\Gamma(p,\theta)\}_{\theta\in\Theta}$ with $p > 0$ fixed: Equation (2.3) (page 11) and Table 2.1 (page 12) yield for the Kullback-Leibler divergence between $\mathbb{P}_y, \mathbb{P}_\theta \in \mathcal{P}$ the explicit formula

  $$g_\theta(y) = \mathcal{KL}(y,\theta) = p\,[y/\theta - 1 - \ln(y/\theta)] \quad \text{and hence} \quad \frac{\partial g_\theta}{\partial y}(y) = p\left(\frac{1}{\theta} - \frac{1}{y}\right).$$

The function g_θ is monotonic on the open sets $V_1 = (0,\theta)$ and $V_2 = (\theta,\infty)$, and it holds $\mathbb{P}_\theta(Y = \theta) = 0$. Lemma 3.6 leads with $f_\theta^Y(y) = \frac{y^{p-1}e^{-y/\theta}}{\theta^p\Gamma(p)}$ to

$$\zeta(y,\theta) := \frac{y}{\theta}, \qquad \tilde{f}(\zeta) := \frac{\zeta^p e^{-\zeta}}{p(\zeta-1)\Gamma(p)}, \quad \text{and} \quad \tilde{g}(\zeta) := p\,[\zeta - 1 - \ln\zeta]\,.$$

This extends to non-adaptive linear combinations as follows. Lemma 3.6 can be applied with respect to the non-adaptive estimator in Notation 2.14, setting $Y := \overline{\theta}_i^{(k)}$ and considering the composition of the density $f_\theta^{\overline{\theta}_i^{(k)}}$ and the Kullback-Leibler divergence described by the function g_θ. While the latter depends on the assumed parametric family $\mathcal{P}$ only, the density $f_\theta^{\overline{\theta}_i^{(k)}}$ is determined via convolution of the probability densities of $\overline{w}_{ij}^{(k)}T(Y_j)/\overline{N}_i^{(k)}$, where $Y_j \sim \mathbb{P}_\theta$. Hence, it depends on the observations via the sufficient statistic T. This will simplify the proof (page 44), where it suffices to show the assertion for the Gaussian and the exponential distribution.

Theorem 3.8. *Suppose Assumption A1. We consider the random variable*

$$Z := g_\theta(\overline{\theta}_i^{(k)}) := \mathcal{KL}\left(\overline{\theta}_i^{(k)},\theta\right),$$

where $\overline{\theta}_i^{(k)}$ denotes the non-adaptive estimator in Notation 2.14 for observations $Y_j \overset{\text{iid}}{\sim} \mathbb{P}_\theta$ with $j \in \{1,...,n\}$ and $\theta \in \Theta$. The density of Z is invariant with respect to the parameter θ if

- $\mathcal{P} = \{\mathcal{N}(\theta,\sigma^2)\}_{\theta\in\Theta}$ *with $\sigma > 0$ fixed,*
- $\mathcal{P} = \{\log\mathcal{N}(\theta,\sigma^2)\}_{\theta\in\Theta}$ *with $\sigma > 0$ fixed,*
- $\mathcal{P} = \{\mathrm{Exp}(1/\theta)\}_{\theta\in\Theta}$,
- $\mathcal{P} = \{\mathrm{Rayleigh}(\theta)\}_{\theta\in\Theta}$,
- $\mathcal{P} = \{\mathrm{Weibull}(\theta,k)\}_{\theta\in\Theta}$ *with $k > 0$,*
- $\mathcal{P} = \{\mathrm{Pareto}(x_m,\theta)\}_{\theta\in\Theta}$ *with $x_m \geq 1$.*

The convolution of the densities of exponential distributions was studied, for instance, by Akkouchi [2008]. We know from Example 3.7 that the density of the random variable $\mathcal{KL}(T(Y),\theta)$ is invariant with respect to the parameter θ if the observations follow a Gamma distribution. However, the probability distribution of the corresponding non-adaptive estimator has a sophisticated form [Mathai, 1982; Moschopoulos, 1985], where the corresponding summands could not been proven to be invariant with respect to θ. Though, we get via Example 3.7 the following corollary.

Corollary 3.9. *Let the location kernel only attain values in $\{0,1\}$, ensuring that $\overline{w}_{ij}^{(k)} \in \{0,1\}$ for all $i,j \in \{1,...,n\}$. Then,*

$$Y_j \overset{\text{iid}}{\sim} \Gamma(p,\theta) \quad \textit{implies that} \quad \overline{\theta}_i^{(k)} \sim \Gamma(\overline{N}_i^{(k)}p, \theta/\overline{N}_i^{(k)}),$$

and, as a consequence, we get the invariance of the distribution of $\mathcal{KL}(\overline{\theta}_i^{(k)},\theta)$ with respect to θ. The same holds for the Erlang and the scaled chi-squared distribution, where

$$\mathrm{Erlang}(k,\theta) = \Gamma(k,\theta), \qquad k \in \mathbb{N},$$

and

$$Y \sim \Gamma(k/2, 2\theta/k) \textit{ if } kY/\theta \sim \chi^2(k) = \Gamma(k/2,2), \qquad k \in \mathbb{N}.$$

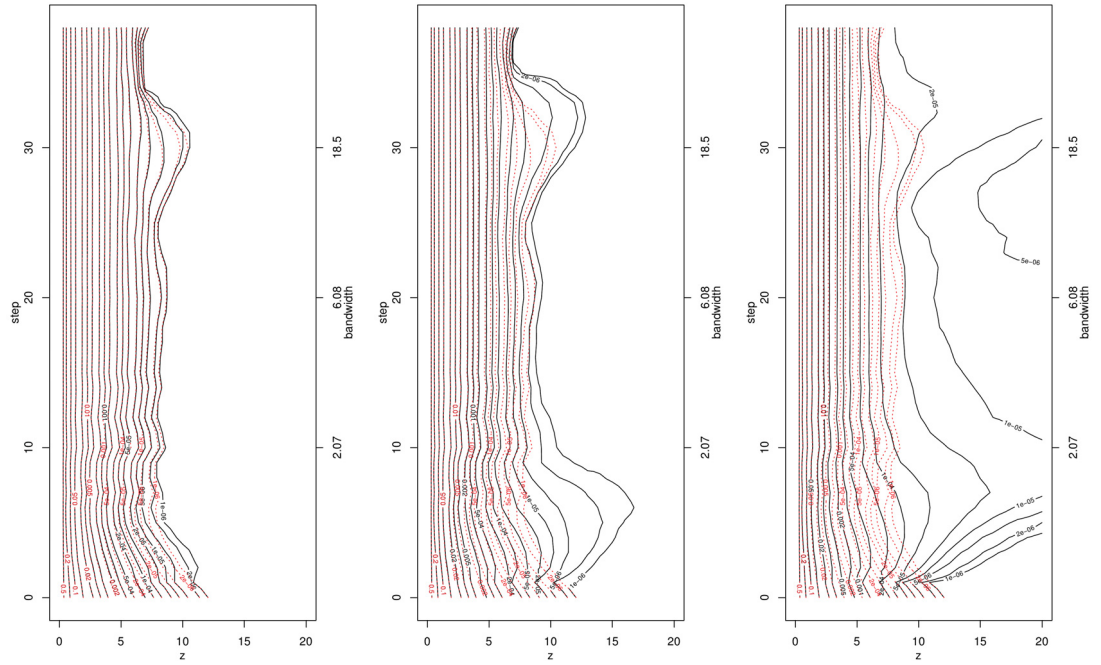

Figure 3.1: Plots of the propagation condition for the Gaussian distribution with (from left to right) $\lambda = 22.4, 13.6, 9.72$. The isolines of the probability p, for values between 10^{-6} and 0.5, are plotted with respect to the location bandwidth $h^{(k)}$ described by the iteration step k and the corresponding value $z = \mathfrak{Z}_\lambda(k, p; \theta = 1, i)$ for some $i \in \{1, ..., n\}$. The black solid lines represent the isolines of the adaptive estimator, the red dotted lines correspond to the non-adaptive estimator.

The new propagation condition is included in the **R**-package **aws** by Polzehl [2012]. In Figures 3.1 and 3.2, we show some examples to illustrate the close relation of the adaptive and the non-adaptive estimator under a satisfied propagation condition. The plots were realized using the function `awstestprop` on a two-dimensional design with 5000×5000 points and the kernels in Equation (2.10) (page 19). The maximal location bandwidth $h^{(k^*)}$ was set to 50 requiring 38 iteration steps. Running the simulation with different parameters θ yield exactly the same plots. In Figure 3.1, we show the results for the Gaussian distribution with three different values of λ. In Figure 3.2, we consider the same setting with respect to the exponential distribution. Together, Theorem 3.8 and the numerical simulations suggest the invariance of the propagation condition with respect to the parameter θ.

Finally, we discuss how to proceed if the function $\mathfrak{Z}_\lambda$ in Notation 3.1 varies with the parameter θ. We aim to ensure that our choice of the adaptation bandwidth λ is in accordance with the propagation condition for all θ_i, $i \in \{1, ..., n\}$. Certainly, we do not know the exact parameters $\{\theta_i\}_i$. Instead, we could analyze the monotonicity of the optimal choice $\lambda_{\text{opt}}(\epsilon, \theta, \lambda_{\min})$, see Notation 3.4, for a fixed constant $\epsilon > 0$ and varying parameters $\theta \in \Theta$. For the sake of simplicity, we prefer to observe, for a fixed adaptation bandwidth λ and varying parameters θ, for which probabilities p the propagation condition is satisfied. This can be done by the function `awstestprop` in the **R**-package **aws**. Thus, for every θ, we get the corresponding value $\epsilon_\lambda(\theta)$. Then, $\epsilon_\lambda(\theta) \geq \epsilon_\lambda(\theta')$ indicates that the parameter θ requires a larger adaptation bandwidth than the parameter θ'. Taking the range of our observations into account, we tempt to identify a small number of parameters $\theta^* \in \Theta$ such that every λ that satisfies the propagation condition for these parameters $\theta^* \in \Theta$ remains valid with high probability for the unknown parameters θ_i, $i \in \{1, ..., n\}$. This approach can be nicely illustrated, considering some examples.

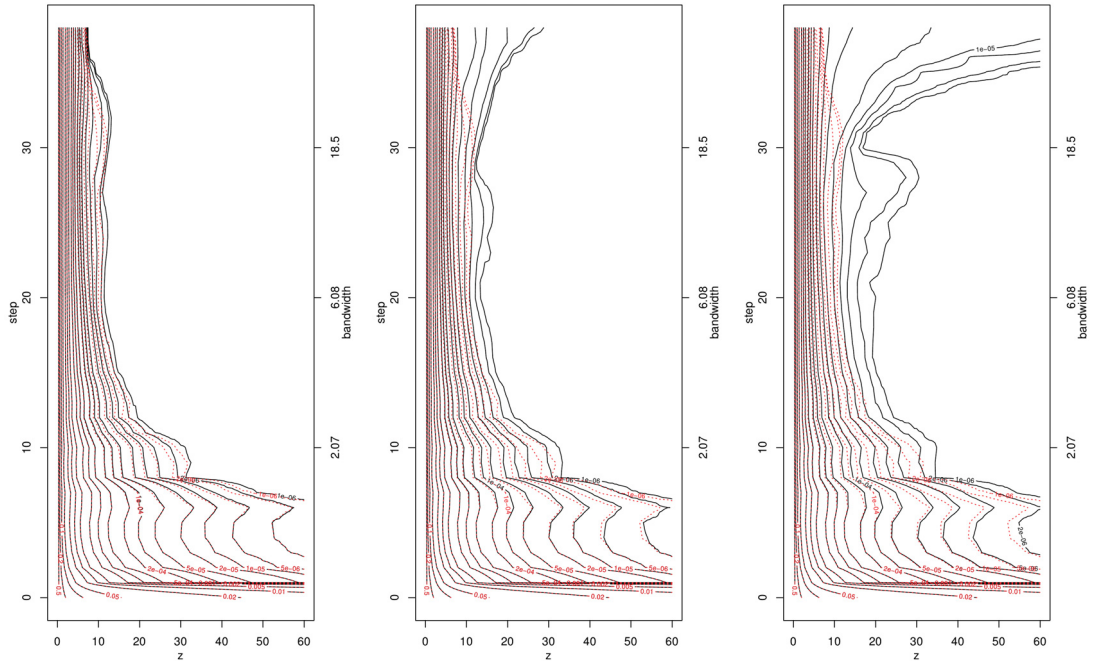

Figure 3.2: Plots of the propagation condition for the exponential distribution with (from left to right) $\lambda = 13.2, 10.2, 8.78$.

Example 3.10.

- For observations which follow a Poisson distribution, it turned out that different parameters θ yield comparable propagation levels $\epsilon_\lambda(\theta)$, even though the resulting isolines differ clearly. This is illustrated in Figure 3.3, where we again consider the kernel functions (2.10), a regular design with 5000×5000 points, and $h^{(k^*)} = 50$.
- in the case of Bernoulli distributed observations, the propagation condition should be ensured for $\theta^* := 0.5$.
- In § 6.5.3, we will show another example, applying the propagation condition to non-central chi-distributed observations on the design space $\mathcal{X} = \mathbb{R}^3 \times \mathbb{S}^2$.

By slightly shifting the estimator, the implemented algorithm avoids that the Kullback-Leibler divergence between two Poisson or Bernoulli distributions becomes infinity.

3.1.3 Practical application

The propagation condition is based on the function $\mathfrak{Z}_\lambda$ in Notation 3.1. This depends on the probability $\mathbb{P}(\overline{N}_i^{(k)} \mathcal{KL}(\tilde{\theta}_i^{(k)}(\lambda), \theta) > z)$ which cannot be calculated exactly. Therefore, in practice, we need an appropriate approximation. Recall that the propagation condition depends on the function $\mathfrak{Z}_\lambda$ via its behavior during iteration only. We know from Theorem 2.13 that the behavior of the non-adaptive term $\overline{N}_i^{(k)} \mathcal{KL}(\overline{\theta}_i^{(k)}, \theta)$ during iteration does not depend on the position X_i within the design $\mathcal{X}$. Since the observations are independent and identically distributed, we may assume that this property extends to the adaptive estimator and consequently to the function $\mathfrak{Z}_\lambda(., p; \theta, i)$. Then, we can estimate the above probability by the relative frequency of design points $X_i \in \mathcal{X}$ with $\overline{N}_i^{(k)} \mathcal{KL}(\tilde{\theta}_i^{(k)}(\lambda), \theta) > z$ as established in Definition 3.11. In Section 4.2, we will justify the presented estimator providing exponential bounds for the estimation bias and

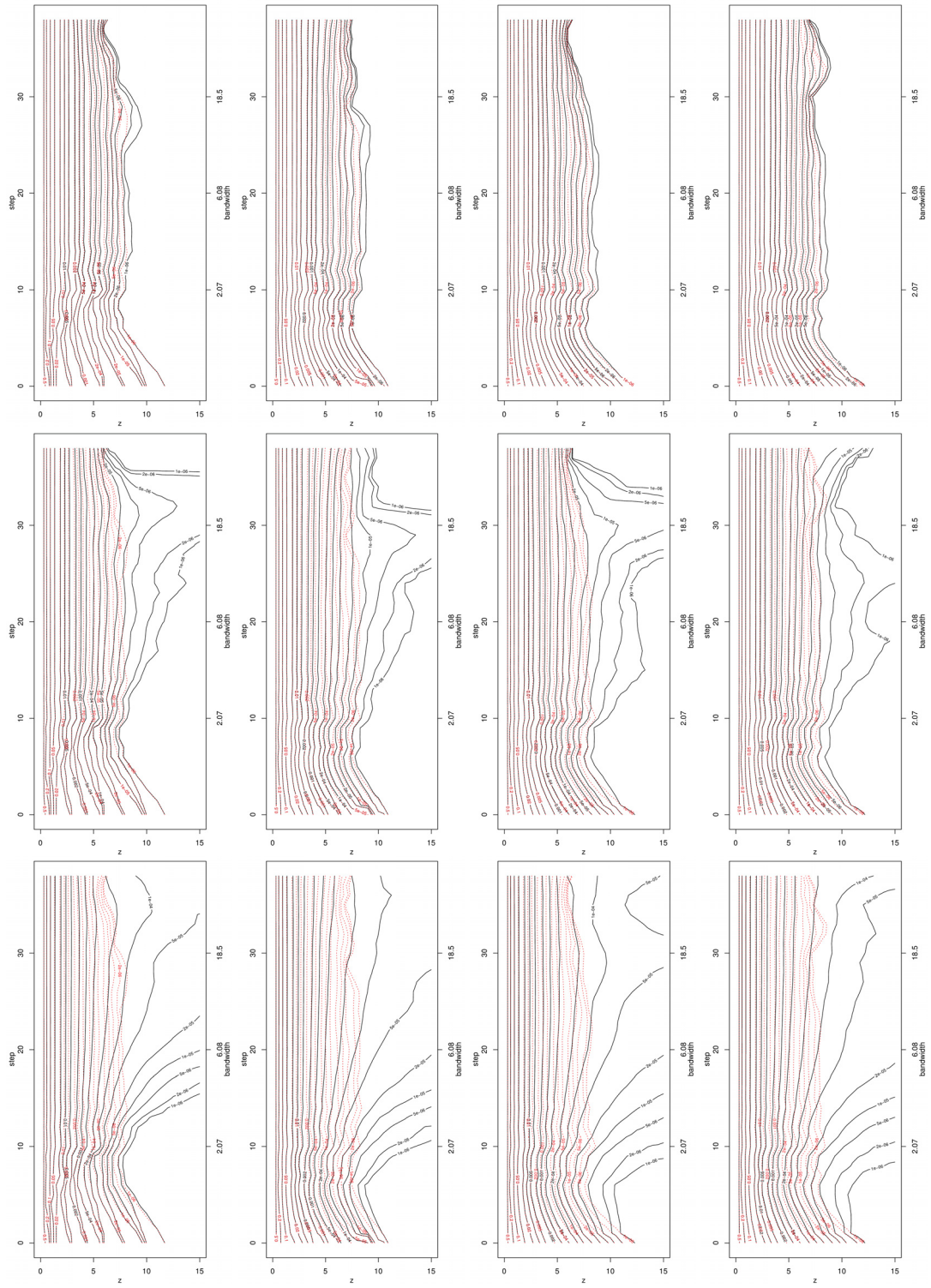

Figure 3.3: Plots of the propagation condition for the Poisson distribution with (from left to right) $\theta = 1, 10, 100, 1000$ and (from top to bottom) $\lambda = 13.2, 9.88, 7.69$, yielding $\epsilon_{13.2}(\theta) \leq 10^{-6}$, $\epsilon_{9.88}(\theta) \approx 5 \cdot 10^{-5}$, and $\epsilon_{7.69}(\theta) \approx 5 \cdot 10^{-4}$.

the corresponding variance, see Lemma 4.5. In order to avoid boundary effects in the resulting estimate, we restrict the approximation to the interior of the design space, that is to all design points $X_i \in \mathcal{X}$ where the final neighborhood $U_i^{(k^*)} := \{X_j \in \mathcal{X} : \overline{w}_{ij}^{(k^*)} > 0\}$ is not restricted by the boundaries of the considered set $\{X_i\}_{i=1}^n \subseteq \mathcal{X}$. This subset of $\{X_i\}_{i=1}^n$ is denoted by $\mathcal{X}^0$. Without loss of generality we assume that $\mathcal{X}^0 = \{X_i\}_{i=1}^{n_0}$ for some $n_0 < n$.

Definition 3.11. *We consider the same setting as in Notation 3.1, and we set*

$$M_\lambda^{(k)}(z) := \{X_i \in \mathcal{X}^0 : \overline{N}_i^{(k)} \mathcal{KL}(\tilde{\theta}_i^{(k)}(\lambda), \theta) > z\}.$$

Then we define the following estimator

$$\hat{p}_\lambda^{(k)}(z) := n_0^{-1} \sum_{i=1}^{n_0} \mathbf{1}_{M_\lambda^{(k)}(z)}(X_i),$$

where $\mathbf{1}$ *denotes the indicator function.*

Remark 3.12. Simulations are carried out using an artificial data set which ensures a sufficiently large number of effectively independent regions for estimating the propagation level on the basis of a single realization. In practice, choices of k^* with $\overline{N}_i^{(k^*)} \ll n$ are often sufficient or even recommendable, see Section 5.4. Here, we approximate the probability $\mathbb{P}(\overline{N}_i^{(k)} \mathcal{KL}(\tilde{\theta}_i^{(k)}(\lambda), \theta) > z)$ by the corresponding relative frequency. This estimate can be calculated for $\epsilon \geq 1/n$ only. Additionally, it becomes unstable if ϵ is close to $1/n$. in the case of a regular design, the sample can be extended in a natural way, allowing arbitrary sample sizes and, as a consequence, any $\epsilon > 0$. Otherwise, that is for random or irregular designs, one should evaluate carefully under which conditions the propagation condition generalizes from the artificial data set to the data set at hand.

3.2 Extension for locally varying parameter functions

The homogeneous propagation condition in Definition 3.2 bounds the probability of adaptation to noise, supposing a constant parameter function. In the next chapter, this will be used to verify propagation and a certain stability of estimates for (piecewise) constant parameter functions. In order to extend these properties to (piecewise) bounded parameter functions, we will formulate an inhomogeneous propagation condition. Like before under homogeneity, we will consider an artificial data set, but this time we will allow local variability. Then, we aim to ensure a similar behavior of the algorithm as for non-adaptive estimation for every locally varying function which satisfies a pre-specified variability bound. Recall that we consider the local exponential family model in Assumption A1 (page 10) and the simplified procedure in Algorithm 2 (page 18).

3.2.1 An inhomogeneous propagation condition

Our inhomogeneous propagation condition is motivated by Theorem 2.24. This can be considered as the inhomogeneous analog of Theorem 2.13, where the homogeneous propagation condition was based on. For the non-adaptive estimator, Theorem 2.24 establishes the exponential bound

$$\mathbb{P}(\overline{N}_i^{(k)} \mathcal{KL}(\overline{\theta}_i^{(k)}, \mathbb{E}\overline{\theta}_i^{(k)}) > z) \leq 2e^{-z/\varkappa^2} + \breve{p}_\varkappa$$

for all $z > 0$, $i \in \{1, ..., n\}$, and $k \in \{0, ..., k^*\}$, where we refer the reader to the Notations 2.6 and 2.21 for the definitions of $\varkappa \geq 1$ and $\breve{p}_\varkappa \in [0, 1]$. This result implies that the Kullback-Leibler divergence $\mathcal{KL}(\overline{\theta}_i^{(k)}, \mathbb{E}\overline{\theta}_i^{(k)})$ decreases, in probability, at least with rate $\overline{N}_i^{(k)}$. We observe that

$$\mathbb{E}\overline{\theta}_i^{(k)} = \sum_{j=1}^{n} \mathbb{E}\left[\overline{w}_{ij}^{(k)} T(Y_j)/\overline{N}_i^{(k)}\right] = \sum_{j=1}^{n} \overline{w}_{ij}^{(k)} \theta_j / \overline{N}_i^{(k)},$$

whereas

$$\mathbb{E}\tilde{\theta}_i^{(k)} = \sum_{j=1}^{n} \mathbb{E}\left[\tilde{w}_{ij}^{(k)} T(Y_j)/\tilde{N}_i^{(k)}\right] \neq \sum_{j=1}^{n} \tilde{w}_{ij}^{(k)} \theta_j / \tilde{N}_i^{(k)}.$$

Both sums can be considered as an adaptive analog of $\mathbb{E}\overline{\theta}_i^{(k)}$. Since the latter is much easier to compute, we concentrate thereon. Recall that the adaptive weights and their sum are random.

Notation 3.13. *We set*

$$\mathcal{E}\tilde{\theta}_i^{(k)} := \sum_{j=1}^{n} \tilde{w}_{ij}^{(k)} \theta_j / \tilde{N}_i^{(k)}.$$

Next we specify the considered inhomogeneous setting. Following Polzehl and Spokoiny [2006, §5.2], we presume that the variability of the parameter function $\theta(.)$ is smaller in order than the rate of convergence $\overline{N}_i^{(k)}$ in Theorem 2.24. Here, we even require the rate $\max_{j'} \overline{N}_{j'}^{(k)}$ in order to ensure that $\overline{N}_i^{(k)} / \max_{j'} \overline{N}_{j'}^{(k)} \leq 1$ for every $i \in \{1, ..., n\}$. More precisely, we require the existence of a constant $\varphi \geq 0$ such that

$$\mathcal{KL}(\theta_i, \theta_j) \leq \varphi^2 / \max_{j'} \overline{N}_{j'}^{(k)} \quad \text{for all } X_j \in U_i^{(k)} := \{X_j \in \mathcal{X} : \overline{w}_{ij}^{(k)} > 0\} \tag{3.2}$$

for every $i \in \{1, ..., n\}$ and each $k \in \{0, ..., k^*\}$. In this chapter, we require Equation (3.2) with $k := k^*$ for all $i, j \in \{1, ..., n\}$, but, in the next chapter, we will only consider the points in a certain neighborhood, for instance all $X_j \in U_i^{(k)}$ with $k \in \{0, ..., k^*\}$. For brevity, we denote $\varphi_0 := \varphi / \max_i (\overline{N}_i^{(k^*)})^{1/2}$.

We proceed as under homogeneity, see § 3.1.1 for comparison. Hence, we aim to choose the adaptation bandwidth independent of the data at hand. We introduce the following function, which relates to Notation 3.1.

Notation 3.14. *For every $\lambda > 0$, we consider the function*

$$\hat{\mathfrak{Z}}_\lambda : \{0, ..., k^*\} \times (0, 1) \times \Theta^n \times \{1, ..., n\} \to [0, \infty)$$

defined by

$$\hat{\mathfrak{Z}}_\lambda(k, p; \{\theta_j\}_{j=1}^n, i) := \inf\left\{z > 0 : \mathbb{P}\left(\overline{N}_i^{(k)} \mathcal{KL}(\tilde{\theta}_i^{(k)}(\lambda), \mathcal{E}\tilde{\theta}_i^{(k)}(\lambda)) > z\right) \leq p\right\},$$

where $\mathcal{E}\tilde{\theta}_i^{(k)}$ is as in Notation 3.13, and $\tilde{\theta}_i^{(k)}(\lambda)$ denotes the adaptive estimator at the position $X_i \in \mathcal{X}$, resulting from the simplified algorithm in Notation 2 with the adaptation bandwidth λ and the statistically independent observations $Y_j \sim \mathbb{P}_{\theta_j} \in \mathcal{P}$, $j \in \{1, ..., n\}$.

In order to enable the application of Equation (2.2) and Lemma 2.8, we restrict the range of the parameter function $\theta(.)$. Thus, we introduce a subset $\Theta^* \subseteq \Theta$, and we require that $\{\theta_j\}_{j=1}^n \in (\Theta^*)^n$.

Definition 3.15 (Inhomogeneous propagation condition)**.** *The adaptation bandwidth $\lambda > 0$ satisfies the inhomogeneous propagation condition at probability level $\epsilon > 0$ and variability level $\varphi_0 > 0$ for the parameter set $\Theta^* \subseteq \Theta$ if the function $\hat{\mathfrak{Z}}_\lambda(.,p;\{\theta_j\}_{j=1}^n,i)$ is non-increasing for all $p \in (\epsilon,1)$, every $i \in \{1,...,n\}$, and each parameter function $\theta(.)$ with $\{\theta_j\}_{j=1}^n \in (\Theta^*)^n$ and $\mathcal{KL}(\theta_j,\theta_{j'}) \leq \varphi_0^2$ for all $j,j' \in \{1,...,n\}$.*

Remark 3.16. For $\varphi_0 := 0$ and $\Theta^* := \{\theta\}$, the inhomogeneous propagation condition equals the homogeneous propagation condition in Definition 3.2.

3.2.2 The inhomogeneous propagation condition in practice

In § 3.1.3, we explained how the homogeneous propagation condition can be applied in practice. In contrast, the inhomogeneous propagation condition cannot be applied directly if $\varphi_0 > 0$. Here, we need to ensure that the criterion is fulfilled for *every* parameter function satisfying $\{\theta_i\}_{i=1}^n \in (\Theta^*)^n$ and $\mathcal{KL}(\theta_i,\theta_j) \leq \varphi_0^2$ for all $i,j \in \{1,...,n\}$. Therefore, we recommend to choose some $\lambda > 0$ in accordance with the homogeneous propagation condition and to increase it such that the inhomogeneous propagation condition holds as well. Apart from the Gaussian and log-normal distribution, the practical use of our precise choice is questionable due to the size of the involved constants. Nevertheless, it suggests the *existence* of an appropriate value. Hence, the inhomogeneous propagation condition is in the first instance of theoretical interest. In Section 4.2, it will provide the desired extension of the propagation and the stability property to (piecewise) bounded functions. The justification of our choice will be based on a comparison of the homogeneous and the inhomogeneous propagation condition. In order to avoid confusion, we introduce the following notation.

Notation 3.17. *Let the parametric family $\mathcal{P}$ satisfy Assumption A1 (page 10) with a strictly monotonic sufficient statistic T. We fix some constant $\varphi_0 > 0$ and a subset $\Theta^* \subseteq \Theta$. Then, we consider two data sets $\{(X_i,Y_i)\}_{i=1}^n$ and $\{(X_i,\mathcal{Y}_i)\}_{i=1}^n$, where*

- *$Y_i \sim \mathbb{P}_{\theta_i} \in \mathcal{P}$ with $\{\theta_i\}_{i=1}^n \in (\Theta^*)^n$ and $\mathcal{KL}(\theta_i,\theta_j) \leq \varphi_0^2$ for all $i,j \in \{1,...,n\}$,*
- *$\mathcal{Y}_i \sim \mathbb{P}_{\vartheta_i} \in \mathcal{P}$ with $\vartheta_i \equiv \vartheta$ for some $\vartheta \in \Theta^*$ (homogeneity).*

In the rest of this chapter, we will write $\mathcal{Y}$ and ϑ whenever we restrict to the special case of a homogeneous setting. Else, we will write Y and θ, explicitly allowing locally varying parameter functions which satisfy the variability bound in Notation 3.17. Now we look for a description of the homogeneous propagation condition which enables an extension to the inhomogeneous setting. For this purpose, we introduce some auxiliary functions.

Notation 3.18. *Let the functions $p_\theta^{(l)} : (0,\infty) \to [0,1]$, $l = 1,2,3$ and $\theta \in \Theta$, be given as*

$$\begin{aligned} p_\theta^{(1)}(z) &:= \mathbb{P}(\{T(Y) > \theta\} \cap \{\mathcal{KL}(T(Y),\theta) > z\}), \\ p_\theta^{(2)}(z) &:= \mathbb{P}(\{T(Y) \leq \theta\} \cap \{\mathcal{KL}(T(Y),\theta) > z\}), \quad z > 0 \\ p_\theta^{(3)}(z) &:= \mathbb{P}(\{T(Y) \leq \theta\} \cap \{\mathcal{KL}(T(Y),\theta) \leq z\}), \end{aligned}$$

where $Y \sim \mathbb{P}_\theta$.

Lemma 3.19. *The functions $p_\theta^{(l)}$, $l = 1,2,3$, in Notation 3.18 are invariant with respect to the parameter $\theta \in \Theta$ for the Gaussian, log-normal, Gamma, Erlang, scaled chi-squared, exponential, Rayleigh, Weibull, and Pareto distributions.*

The study in § 3.1.2 suggests the invariance of the homogeneous propagation condition with respect to the parameter θ for the Gaussian, log-normal, exponential, Rayleigh, Weibull, and Pareto distribution. In the following lemma, we take advantage of this invariance. There, we completely determine the corresponding function $\mathfrak{Z}_\lambda$ via the distribution of the positions of the observations around the respective parameter $\vartheta \in \Theta$ given by the functions $p_\vartheta^{(l)}$, $l = 1, 2, 3$.

Lemma 3.20. *Assume the setting of Notation 3.17. If the homogeneous propagation condition is invariant with respect to the parameter $\vartheta \in \Theta$, then the corresponding function $\mathfrak{Z}_\lambda$ is uniquely determined by the functions $p_\vartheta^{(l)}$, $l = 1, 2, 3$, for every $\vartheta \in \Theta$.*

In other words, the homogeneous propagation condition is determined by the probability distributions of $\mathcal{KL}(T(\mathcal{Y}), \vartheta)$ on $\{T(\mathcal{Y}) > \vartheta\}$ and on $\{T(\mathcal{Y}) \leq \vartheta\}$. Under inhomogeneity, we have to additionally compensate for the local variability of the parameter function. We investigate the interplay of the observations via the distribution of $\mathcal{KL}(Y_i, Y_j)$, which we compare with its homogeneous counterpart $\mathcal{KL}(\mathcal{Y}_i, \mathcal{Y}_j)$. For simplicity, we presume the sufficient statistic T in Assumption A1 to equal the identity. Instead of that, we could replace in the following all observations Y_i and $\mathcal{Y}_i$ by the transformed observations $T(Y_i)$ and $T(\mathcal{Y}_i)$, leading, for every strictly monotonic T, to the same results but more tedious terms. We restrict to the favorable realizations, where the corresponding event M_0 is related to the event $\Omega_\varkappa$ in Lemma 2.23.

Proposition 3.21. *Suppose the setting of Notation 3.17, $T = \mathrm{Id}$, and let the functions $p_\theta^{(l)}$, $l = 1, 2, 3$, be invariant with respect to the parameter $\theta \in \Theta^*$. Additionally, recall Notation 2.6, and let $\varkappa \geq 1$ satisfy $\{\vartheta\} \cup \{\theta_i\}_{i=1}^n \in (\Theta_\varkappa)^{n+1}$, where $\Theta_\varkappa \subseteq \Theta$ maximizes the probability of the event*

$$M_0 := \bigcap_{i=1}^n \{Y_i, \mathcal{Y}_i \in \Theta_\varkappa\}.$$

Then, for all $z > \varkappa^2 \varphi_0^2$ and every $i, j \in \{1, ..., n\}$, it holds

$$\mathbb{P}\left(\{\mathcal{KL}(Y_i, Y_j) > z\} | M_0\right) \leq \mathbb{P}\left(\left\{\varkappa^2 [\varkappa \mathcal{KL}^{1/2}(\mathcal{Y}_i, \mathcal{Y}_j) + \varphi_0]^2 > z\right\} | M_0\right).$$

Now we propose a precise choice of the adaptation bandwidth for the case of a (piecewise) bounded parameter function.

Claim 3.22. *Let Assumption A1 be satisfied with $T : \mathcal{Y} \to \mathbb{R}$ strictly monotonic, and fix a subset $\Theta^* \subseteq \Theta$ and some constant $\varphi_0 := \varphi / \max_i (\overline{N}_i^{(k^*)})^{1/2}$ with $\varphi > 0$. Additionally, let the homogeneous propagation condition and the functions $p_\theta^{(l)}$, $l = 1, 2, 3$, be invariant with respect to the parameter $\theta \in \Theta$. Finally, we presume the adaptation bandwidth $\lambda > 0$ to be in accordance with the homogeneous propagation condition at level $\epsilon > 0$. Then, the choice*

$$\lambda_\varphi := \varkappa^4 \left[\sqrt{\lambda} + \varphi\right]^2$$

is in accordance with the inhomogeneous propagation condition at a probability level $\epsilon(\lambda_\varphi)$ with $\epsilon(\lambda_\varphi) \leq \epsilon + 2p_\varkappa$ and variability level φ_0 for the parameter set Θ^*, where $p_\varkappa$ is as in Notation 2.16.

Admittedly, the iterative approach of the algorithm impedes a definite proof. Instead, we will present in § 3.3 a justification of Claim 3.22, where we will follow an inductive argumentation in order to overcome the remaining gap at least to a certain extent.

Remark 3.23. Recall that the parameter set $\Theta^* \subseteq \Theta$ and the variability level φ_0 influence the sizes of the constant $\varkappa \geq 1$ and the corresponding probability $p_\varkappa$ in Notation 2.16. Moreover, we point out that the probability level $\epsilon + 2p_\varkappa$ is an upper bound for the actual probability level $\epsilon(\lambda_\varphi)$, but this bound does not have to be sharp. Similarly, the proposed choice of λ_φ is based on a rough estimation, where the effectively required values of $\varkappa$ and φ may be much smaller than supposed. Hence, in practice, one can always use a smaller bandwidth $\lambda^* < \lambda_\varphi$ with an unknown probability level if this seems to be advantageous. This may increase the risk of adaptation to noise as usually $\epsilon(\lambda^*) \geq \epsilon(\lambda_\varphi)$, but the main property, propagation with probability $1 - \epsilon(\lambda^*)$, remains valid. In any case, one should use the homogeneous adaptation bandwidth λ as a lower bound, $\lambda \leq \lambda^*$. Recalling Example 2.17 concerning the trade-off between $\varkappa$ and $p_\varkappa$, we conclude the following.

- Claim 3.22 provides a reasonable choice of the adaptation bandwidth if $\varkappa = 1$ and $p_\varkappa = 0$, such as for Gaussian and log-normal distributed observations.
- For Gamma, Erlang, scaled chi-squared, exponential, Rayleigh, Weibull, and Pareto distribution, $\varkappa$ and $p_\varkappa$ are large. For these distributions, Claim 3.22 justifies the *existence* of an adaptation bandwidth λ_φ which is in accordance with the inhomogeneous propagation condition at level $\epsilon(\lambda_\varphi) \leq \epsilon + 2p_\varkappa$, but its practical use is questionable due to the sizes of $\varkappa$ and $p_\varkappa$.

3.3 Proofs

Proof of Proposition 3.5. Substitution with $y := g_\theta^{-1}(z)$ yields

$$h(g_\theta(y), \theta) = f(y, \theta) \quad \text{for all } (y, \theta) \in \Omega^f.$$

Hence, we get the total derivatives

$$\frac{dh}{d\theta} = \frac{\partial h}{\partial z}\frac{\partial g}{\partial \theta} + \frac{\partial h}{\partial \theta} = \frac{\partial f}{\partial \theta} \quad \text{and} \quad \frac{dh}{dy} = \frac{\partial h}{\partial z}\frac{\partial g}{\partial y} = \frac{\partial f}{\partial y}.$$

Then, it follows $\frac{\partial h}{\partial z} = \frac{\partial f}{\partial y} / \frac{\partial g}{\partial y}$ and furthermore

$$\frac{\partial f}{\partial y}\frac{\partial g}{\partial \theta} + \frac{\partial h}{\partial \theta}\frac{\partial g}{\partial y} = \frac{\partial f}{\partial \theta}\frac{\partial g}{\partial y}.$$

This leads with $\left|\frac{\partial g_\theta}{\partial y}\right| > 0$ to

$$\frac{\partial h}{\partial \theta} = \left(\frac{\partial f}{\partial \theta}\frac{\partial g}{\partial y} - \frac{\partial f}{\partial y}\frac{\partial g}{\partial \theta}\right) \cdot \left(\frac{\partial g}{\partial y}\right)^{-1},$$

and we have

$$\frac{\partial h}{\partial \theta} = 0 \quad \text{if and only if} \quad \frac{\partial f}{\partial \theta}\frac{\partial g}{\partial y} = \frac{\partial f}{\partial y}\frac{\partial g}{\partial \theta}.$$

The chain rule implies with Equation (3.1) (page 34) that indeed

$$\frac{\partial f}{\partial \theta}\frac{\partial g}{\partial y} = \frac{\partial \tilde{f}}{\partial \zeta}\frac{\partial \zeta}{\partial \theta}\frac{\partial \tilde{g}}{\partial \zeta}\frac{\partial \zeta}{\partial y} = \frac{\partial \tilde{f}}{\partial \zeta}\frac{\partial \zeta}{\partial y}\frac{\partial \tilde{g}}{\partial \zeta}\frac{\partial \zeta}{\partial \theta} = \frac{\partial f}{\partial y}\frac{\partial g}{\partial \theta},$$

yielding that h is invariant with respect to θ. □

Proof of Lemma 3.6. We deduce the probability density of the random variable Z from the probability density of Y via the usual change of variable, which leads to

$$f_\theta^Z(z) = \sum_{l=1}^{m(z)} f_\theta^Y\left(\hat{g}_{\theta,l}^{-1}(z)\right) \cdot \left|\frac{\partial g_\theta}{\partial y}\left(\hat{g}_{\theta,l}^{-1}(z)\right)\right|^{-1},$$

where $m(z)$ is the number of solutions for the equation $g_\theta(y) = z$, and $\hat{g}_{\theta,l}^{-1}(z)$ denote these solutions. Equivalently, we can divide the domain of g_θ, that is $\mathcal{Y} \subseteq \mathbb{R}$, into $m := \max_z m(z)$ disjoint regions where g_θ is monotonic, leading to the formula

$$f_\theta^Z(z) = \sum_{l=1}^{m} f_\theta^Y\left(g_{\theta,l}^{-1}(z)\right) \cdot \left|\frac{\partial g_\theta}{\partial y}\left(g_{\theta,l}^{-1}(z)\right)\right|^{-1} \cdot \mathbf{1}_{g_\theta(V_l)}(z),$$

where $g_{\theta,l} : V_l \to g_{\theta,l}(V_l) \subseteq [0,\infty)$ denotes the bijective restriction of g_θ to the region of monotonicity V_l and $\mathbf{1}$ is the indicator function. Then, the invariance with respect to θ follows for each summand $l \in \{1,...,m\}$ and, as a consequence, for f_θ^Z, as a special case of Proposition 3.5 with

$$h(z,\theta) := f_\theta^Y\left(g_{\theta,l}^{-1}(z)\right) \cdot \left|\frac{\partial g_\theta}{\partial y}\left(g_{\theta,l}^{-1}(z)\right)\right|^{-1}, \qquad z \in V_l.$$

This terminates the proof. □

Proof of Theorem 3.8. The non-adaptive estimator is defined as the weighted mean of $T(Y_j)$ with $j = 1,..,n$. We get from Table 2.1 (page 12) that

- $T(Y) = \ln(Y) \sim \mathcal{N}(\mu,\sigma^2)$ if $Y \sim \log\mathcal{N}(\mu,\sigma^2)$ with $\sigma^2 > 0$;
- $T(Y) = Y^2 \sim \mathrm{Exp}\left(\frac{1}{2\theta^2}\right)$ if $Y \sim \mathrm{Rayleigh}(\theta)$;
- $T(Y) = Y^k \sim \mathrm{Exp}\left(\frac{1}{\theta^k}\right)$ if $Y \sim \mathrm{Weibull}(\theta,k)$ with $k > 0$;
- $T(Y) = \ln\left(y/x_m\right) \sim \mathrm{Exp}\left(\frac{1}{\theta}\right)$ if $Y \sim \mathrm{Pareto}(x_m, \frac{1}{\theta})$ with $x_m \geq 1$.

Hence, in each of these cases, the non-adaptive estimator follows the same distribution as for Gaussian or exponentially distributed observations. Additionally, the corresponding Kullback-Leibler divergences coincide with the respective divergences of the Gaussian or exponential distributions. Therefore, it suffices to consider Gaussian and exponential distributed observations.

In the Gaussian case, $Y_j \overset{\text{iid}}{\sim} \mathcal{N}(\theta,\sigma^2)$, it follows from the statistical independence of the observations that

$$\overline{\theta}_i^{(k)} \sim \mathcal{N}\left(\theta, \sigma_i^2\right), \quad \text{where } \sigma_i^2 := \sigma^2 \cdot \sum_{j=1}^{n}\left(\overline{w}_{ij}^{(k)} / \overline{N}_i^{(k)}\right)^2.$$

Hence, the non-adaptive estimator is Gaussian as well, and the invariance with respect to θ follows as in Example 3.7, where ζ and $\tilde{g}$ remain unchanged, and

$$\tilde{f}(\zeta) := \frac{\sigma^2}{\zeta\sigma_i\sqrt{2\pi}} \exp\left(-\frac{\zeta^2}{2\sigma_i^2}\right).$$

Next we consider the exponential distribution, supposing $Y_j \overset{\text{iid}}{\sim} \mathrm{Exp}(1/\theta)$. We distinguish two cases. First, if all non-zero weights are equal and hence $\overline{w}_{ij}^{(k)} \in \{0,1\}$ as $\overline{w}_{ii}^{(k)} = 1$ for all k, then the non-adaptive estimator $\overline{\theta}_i^{(k)}$ is Gamma-distributed, $\overline{\theta}_i^{(k)} \sim \Gamma\left(\overline{N}_i^{(k)}, \theta/\overline{N}_i^{(k)}\right)$. This yields the

desired invariance with respect to θ via Example 3.7, setting $Y := \overline{\theta}_i^{(k)}$. Next in the general case, we require the existence of non-zero weights $\overline{w}_{ij}^{(k)} \neq \overline{w}_{ij'}^{(k)}$ with $j, j' \in \{1, ..., n\}$. For all $a_j > 0$, it holds $a_j Y_j \sim \mathrm{Exp}(1/(\theta a_j))$ if $Y_j \sim \mathrm{Exp}(1/\theta)$, where we denote $a_j := \overline{w}_{ij}^{(k)} / \overline{N}_i^{(k)}$. The linear combination $Y := a_1 Y_1 + a_2 Y_2$ with $a_1 \neq a_2$ has the density

$$\begin{aligned}
f^Y(y) &= \left(f^{a_1 Y_1} * f^{a_2 Y_2}\right)(y) \\
&= \int_0^y \frac{1}{\theta a_1} e^{-\frac{y-z}{\theta a_1}} \frac{1}{\theta a_2} e^{-\frac{z}{\theta a_2}} \, dz \\
&= \frac{e^{-\frac{y}{\theta a_1}}}{\theta^2 a_1 a_2} \int_0^y e^{-z\frac{a_1 - a_2}{\theta a_1 a_2}} \, dz \\
&= \frac{e^{-\frac{y}{\theta a_1}}}{\theta^2 a_1 a_2} \cdot \frac{\theta a_1 a_2}{a_2 - a_1} \left(e^{-y\frac{a_1 - a_2}{\theta a_1 a_2}} - 1\right) \\
&= \frac{1}{\theta(a_1 - a_2)} e^{-\frac{y}{\theta a_1}} - \frac{1}{\theta(a_1 - a_2)} e^{-\frac{y}{\theta a_2}} \\
&= \frac{a_1}{a_1 - a_2} f^{a_1 Y_1}(y) - \frac{a_2}{a_1 - a_2} f^{a_2 Y_2}(y),
\end{aligned}$$

which is a weighted sum of the component densities. Therefore, this extends to the more general case $\overline{Y} := a_1 Y_1 + ... + a_m Y_m$ with $a_j \neq a_{j'}$ for all $j \neq j'$. Subsequently including the observations with equal weights $a_j = a_{j'}$ for some $j, j' \in \{1, ..., n\}$, we conclude by the commutativity, associativity, and distributivity of the convolution that

$$f_\theta^{\overline{\theta}_i^{(k)}} = \sum_{j=1}^{M} c_j f_j, \qquad M \in \mathbb{N},$$

where the constants $c_j \in \mathbb{R}$ again depend on $a_1, ..., a_m$ only. The densities f_j follow the distribution $\Gamma(\hat{m}, \theta a_j)$, where $\hat{m} \leq m_j$ and m_j denotes the number of observations $Y_{j'}$ with weights $a_{j'} = a_j$. Thus, we get from Example 3.7 the invariance with respect to θ for each summand $c_j f_j$, which yields the assertion for weighted sums of exponentials. We again refer the reader to Akkouchi [2008] for further details concerning the density of the convolution of exponential distributions. □

Proof of Lemma 3.19. We recall from the proof of Theorem 3.8 the relationships of the log-normal with the Gaussian distribution, and of the Rayleigh, Weibull, and Pareto distribution with the exponential one. Additionally, it holds $\mathrm{Exp}(1/\theta) = \Gamma(1, \theta)$, $\mathrm{Erlang}(k, \theta) = \Gamma(k, \theta)$, and $Y \sim \Gamma(k/2, 2\theta/k)$ if $kY/\theta \sim \chi^2(k) = \Gamma(k/2, 2)$, where $k \in \mathbb{N}$. Since the associated Kullback-Leibler divergences coincide, it suffices to show the assertion for the Gaussian and the Gamma distribution, which satisfy Assumption A1 with $T = Id$. By Lemma 2.5 and due to Assumption A1, the function $g_\theta(y) = \mathcal{KL}(y, \theta)$, $y \in \mathcal{Y}$, fulfills

$$\frac{dg_\theta}{dy}(y) = C(y) - C(\theta) \begin{cases} < 0 & \text{if } y < \theta, \\ = 0 & \text{if } y = \theta, \\ > 0 & \text{if } y > \theta. \end{cases} \tag{3.3}$$

Hence, $\{Y > \theta\}$ and $\{Y \leq \theta\}$ restrict the random variable $\mathcal{KL}(y, \theta)$ to its regions of monotonicity. On each of these regions the assertion follows from Example 3.7. □

Proof of Lemma 3.20. The parametric family of probability distributions $\mathcal{P}$ is known, and it characterizes, for every $\vartheta \in \Theta$, the function $g_\vartheta : \mathcal{Y} \to [0, \infty)$ given by $g_\vartheta(y) = \mathcal{KL}(T(y), \vartheta)$.

Moreover, the probability distribution of the observations $\mathcal{Y}_i \overset{\text{iid}}{\sim} \mathbb{P}_\vartheta \in \mathcal{P}$ determines the probability distribution of the random quantities $\tilde{N}_i^{(k)}$ and $\tilde{\vartheta}_i^{(k)}$ and consequently of the random variables $\{\tilde{N}_i^{(k)} \mathcal{KL}(\tilde{\vartheta}_i^{(k)}, \vartheta)\}_i$, on which the function $\mathfrak{Z}_\lambda$ is based. The function $\mathfrak{Z}_\lambda$ is invariant with respect to the parameter $\vartheta \in \Theta$ if and only if the homogeneous propagation condition is invariant with respect to ϑ. Therefore, it suffices to show that, for every $\vartheta \in \Theta$, the functions $p_\vartheta^{(l)}$ with $l = 1,2,3$ allow exact reconstruction of $\mathbb{P}_\vartheta$ via the inverse of g_ϑ.

If the sufficient statistic in Assumption A1 satisfies $T = Id$, then it follows from Equation (3.3) that the inverse g_ϑ^{-1} has exactly one solution on $\{y < \vartheta\}$ and $\{y > \vartheta\}$, respectively, and it holds $g_\vartheta^{-1}(0) = \vartheta$. If $T \neq Id$, then we get

$$g_\vartheta(y) := \mathcal{KL}(T(y), \vartheta) = T(y)\,[C(T(y)) - C(\vartheta)] - [B(T(y)) - B(\vartheta)]$$

and

$$\begin{aligned} \frac{dg_\vartheta}{dy}(y) &= T'(y)\,[C(T(y)) - C(\vartheta)] + T(y)C'(T(y))T'(y) - B'(T(y))T'(y) \\ &= T'(y)\,[C(T(y)) - C(\vartheta)] \end{aligned}$$

in place of Equation (3.3). Here, the assumed strict monotonicity of T leads to the same regions of monotonicity as described above, namely $\{T(y) > \vartheta\}$ and $\{T(y) < \vartheta\}$. Furthermore, knowledge of $p_\vartheta^{(l)}$ with $l = 1,2,3$ yields knowledge of $\mathbb{P}(\{T(\mathcal{Y}_i) > \vartheta\} \cap \{\mathcal{KL}(T(\mathcal{Y}_i), \vartheta) \leq z\})$, $i \in \{1, ..., n\}$. Therefore, we can reconstruct $\mathbb{P}_\vartheta$ for every ϑ from $p_\vartheta^{(l)}$, $l = 1,2,3$, which leads to the assertion. □

For the proof of Proposition 3.21, we recall the following basic result.

Lemma 3.24. *Let $a_1, a_2, b_1, b_2 \in \mathbb{R}$ satisfy $(a_1 - b_1) \cdot (a_2 - b_2) \geq 0$. Then it holds*

$$|a_1 - a_2| \leq ||a_1 - b_1| - |a_2 - b_2|| + |b_1 - b_2|.$$

Proof. We start by considering three points $a_l \in \mathbb{R}$, $l = 1,2,3$, with $a_1 \leq a_2 \leq a_3$. Then it holds

$$(a_2 - a_1) = (a_3 - a_1) - (a_3 - a_2) \quad \text{and} \quad (a_3 - a_2) = (a_3 - a_1) - (a_2 - a_1).$$

More generally, for all $a_1, a_2, b \in \mathbb{R}$ with $(b - a_1) \cdot (b - a_2) \geq 0$, we get

$$|a_1 - a_2| = \max\{|a_1 - b|, |a_2 - b|\} - \min\{|a_1 - b|, |a_2 - b|\} = ||a_1 - b| - |a_2 - b||$$

since $(b - a_1) \cdot (b - a_2) \geq 0$ implies that either $a_1, a_2 \leq b$ or $a_1, a_2 \geq b$. In the case where $(a_1 - b_1) \cdot (a_2 - b_2) \geq 0$, we just add the distance $|b_1 - b_2|$ to the right, which leads to the stated upper bound. □

Proof of Proposition 3.21. Lemma 2.8 provides on M_0, for all $i, j \in \{1, ..., n\}$, that

$$\mathcal{KL}(Y_i, Y_j) \leq \varkappa^2 \left[\mathcal{KL}^{1/2}(Y_i, \theta_i) + \mathcal{KL}^{1/2}(\theta_i, \theta_j) + \mathcal{KL}^{1/2}(Y_j, \theta_j)\right]^2. \tag{3.4}$$

On a certain set of realizations this upper bound can be improved by Lemma 3.24. For this purpose, we distinguish the following sets

$$M_1 := \{(Y_i - \theta_i) \cdot (Y_j - \theta_j) \geq 0\} \quad \text{and} \quad M_2 := \{(Y_i - \theta_i) \cdot (Y_j - \theta_j) < 0\} \tag{3.5}$$

and analogously in the homogeneous setting

$$M_3 := \{(\mathcal{Y}_i - \vartheta)\cdot(\mathcal{Y}_j - \vartheta) \geq 0\} \quad \text{and} \quad M_4 := \{(\mathcal{Y}_i - \vartheta)\cdot(\mathcal{Y}_j - \vartheta) < 0\}. \tag{3.6}$$

Now we separately reduce the Kullback-Leibler divergence $\mathcal{KL}(Y_i, Y_j)$ on M_1 and on M_2 to appropriate terms which only depend on the divergences $\mathcal{KL}(Y_i, \theta_i)$ and $\mathcal{KL}(Y_j, \theta_j)$. Then, the invariance of the functions $p_\theta^{(l)}$, $l = 1, 2, 3$, with respect to the parameter θ allows a comparison with the homogeneous analogs, namely $\mathcal{KL}(\mathcal{Y}_i, \vartheta)$ and $\mathcal{KL}(\mathcal{Y}_j, \vartheta)$. Due to the separate handling of the realizations on M_1 and on M_2, the resulting formulas can be reduced to the divergence $\mathcal{KL}(\mathcal{Y}_i, \mathcal{Y}_j)$, which will lead to the assertion.

On the set M_2, we just use the upper bound (3.4). For the set M_1, we recall Equation (2.13) (page 26), where we established by Taylor's Theorem for all $\theta_1, \theta_2 \in \Theta$ the existence of a parameter $\theta^* \in \Theta$ between θ_1 and θ_2 which satisfies

$$\mathcal{KL}(\theta_1, \theta_2) = \tfrac{1}{2I(\theta^*)}\left[C(\theta_1) - C(\theta_2)\right]^2. \tag{3.7}$$

Therefore, on $M_1 \cap M_0$ it follows from Lemma 3.24 and the monotonicity of the function C that

$$\begin{aligned}
\mathcal{KL}(Y_i, Y_j) &\leq \frac{1}{2I(\theta^*)}\left[\,\big||C(Y_i) - C(\theta_i)| - |C(Y_j) - C(\theta_j)|\big| + |C(\theta_i) - C(\theta_j)|\right]^2 \\
&\overset{\text{Eq. (2.2)}}{\leq} \varkappa^2\left[\left|\mathcal{KL}^{1/2}(Y_i, \theta_i) - \mathcal{KL}^{1/2}(Y_j, \theta_j)\right| + \mathcal{KL}^{1/2}(\theta_i, \theta_j)\right]^2.
\end{aligned} \tag{3.8}$$

Then, using the invariance of $p_\theta^{(l)}$, $l = 1, 2, 3$, with respect to the parameter θ and $\mathcal{KL}(\theta_i, \theta_j) \leq \varphi_0^2$ for all $i, j \in \{1, ..., n\}$, we can deduce that

$$\begin{aligned}
&\mathbb{P}\left(\{\mathcal{KL}(Y_i, Y_j) > z\} \cap M_0\right) \\
=\ &\mathbb{P}\left(M_1 \cap M_0 \cap \{\mathcal{KL}(Y_i, Y_j) > z\}\right) + \mathbb{P}\left(M_2 \cap M_0 \cap \{\mathcal{KL}(Y_i, Y_j) > z\}\right) \\
\leq\ &\mathbb{P}\left(M_1 \cap M_0 \cap \left\{\left|\mathcal{KL}^{1/2}(Y_i, \theta_i) - \mathcal{KL}^{1/2}(Y_j, \theta_j)\right|^2 > \left[\sqrt{z}/\varkappa - \varphi_0\right]^2\right\}\right) \\
&+\mathbb{P}\left(M_2 \cap M_0 \cap \left\{\left[\mathcal{KL}^{1/2}(Y_i, \theta_i) + \mathcal{KL}^{1/2}(Y_j, \theta_j)\right]^2 > \left[\sqrt{z}/\varkappa - \varphi_0\right]^2\right\}\right) \\
=\ &\mathbb{P}\left(M_3 \cap M_0 \cap \left\{\left[\mathcal{KL}^{1/2}(\mathcal{Y}_i, \vartheta) - \mathcal{KL}^{1/2}(\mathcal{Y}_j, \vartheta)\right]^2 > \left[\sqrt{z}/\varkappa - \varphi_0\right]^2\right\}\right) \\
&+\mathbb{P}\left(M_4 \cap M_0 \cap \left\{\left[\mathcal{KL}^{1/2}(\mathcal{Y}_i, \vartheta) + \mathcal{KL}^{1/2}(\mathcal{Y}_j, \vartheta)\right]^2 > \left[\sqrt{z}/\varkappa - \varphi_0\right]^2\right\}\right).
\end{aligned}$$

Equation (3.7) leads on $M_3 \cap M_0$ with appropriate parameters $\vartheta_1^*, \vartheta_2^* \in \Theta_\varkappa$ to

$$\begin{aligned}
&\left[\mathcal{KL}^{1/2}(\mathcal{Y}_i, \vartheta) - \mathcal{KL}^{1/2}(\mathcal{Y}_j, \vartheta)\right]^2 \\
\overset{\text{Eq. (3.7)}}{\leq}\ &\max\left\{\frac{1}{2I(\vartheta_1^*)}, \frac{1}{2I(\vartheta_2^*)}\right\}\left[|C(\mathcal{Y}_i) - C(\vartheta)| - |C(\mathcal{Y}_j) - C(\vartheta)|\right]^2 \\
\overset{M_3}{=}\ &\max\left\{\frac{1}{2I(\vartheta_1^*)}, \frac{1}{2I(\vartheta_2^*)}\right\}\left[(C(\mathcal{Y}_i) - C(\vartheta)) - (C(\mathcal{Y}_j) - C(\vartheta))\right]^2 \\
\overset{\text{Eq. (2.2)}}{\leq}\ &\varkappa^2\mathcal{KL}(\mathcal{Y}_i, \mathcal{Y}_j).
\end{aligned}$$

On $M_4 \cap M_0$, we get in a uniform manner

$$\begin{aligned}
&\left[\mathcal{KL}^{1/2}(\mathcal{Y}_i, \vartheta) + \mathcal{KL}^{1/2}(\mathcal{Y}_j, \vartheta)\right]^2 \\
\overset{M_4}{\leq}\ &\max\left\{\frac{1}{2I(\vartheta_1^*)}, \frac{1}{2I(\vartheta_2^*)}\right\}\left[(C(\mathcal{Y}_i) - C(\vartheta)) + (C(\vartheta) - C(\mathcal{Y}_j))\right]^2 \\
\overset{\text{Eq. (2.2)}}{\leq}\ &\varkappa^2\mathcal{KL}(\mathcal{Y}_i, \mathcal{Y}_j).
\end{aligned}$$

Hence, we conclude that

$$\begin{aligned}\mathbb{P}\left(\{\mathcal{KL}(Y_i,Y_j)>z\}\cap M_0\right) &\leq \mathbb{P}\left(\left\{\varkappa^2\mathcal{KL}\left(\mathcal{Y}_i,\mathcal{Y}_j\right)>\left[\sqrt{z}/\varkappa-\varphi_0\right]^2\right\}\cap M_0\right)\\ &= \mathbb{P}\left(\left\{\varkappa^2\left[\varkappa\mathcal{KL}^{1/2}\left(\mathcal{Y}_i,\mathcal{Y}_j\right)+\varphi_0\right]^2>z\right\}\cap M_0\right).\end{aligned}$$

This terminates the proof. □

Justification of Claim 3.22. Recall Notation 3.17. For simplicity, we concentrate on the case $T=\mathrm{Id}$ as presumed in Proposition 3.21. For $T\neq Id$, the assertion follows just as for $T=Id$, replacing in the proof of Proposition 3.21 and in the following formulas the observations Y_i and $\mathcal{Y}_i$ by the transformed observations $T(Y_i)$ and $T(\mathcal{Y}_i)$ for all $i\in\{1,...,n\}$. Recall from the proof of Lemma 3.20 that the assumed strict monotonicity of T ensures that the regions of monotonicity in Equation (3.3) remain valid.

We know from Theorems 2.13 and 2.24 that the Kullback-Leibler divergence between the non-adaptive estimator and its expectation converges, in probability, at least with rate $\overline{N}_i^{(k)}$ under homogeneity and under inhomogeneity. The choice of λ is in accordance with the homogeneous propagation condition by assumption. Hence, it compensates with probability $1-\epsilon$ the impact of the adaptivity under homogeneity, and it only depends on the functions $p_\vartheta^{(l)}$, $l=1,2,3$, see Lemma 3.20. Due to the assumed invariance of $p_\vartheta^{(l)}$ with respect to ϑ, it holds $p_{\theta_i}^{(l)}=p_\vartheta^{(l)}$ for every $l=1,2,3$ and all $i\in\{1,...,n\}$. Therefore, it suffices to increase the homogeneous bandwidth λ pursuant to the maximal impact of the local variability of $\theta(.)$, but independent of the precise definition of $\theta(.)$.

The local variability of $\theta(.)$ effects the interplay of the observations and hence the adaptive weights, where we consider the random variables $\{s_{ij}^{(k)}\}_{i,j}$, see Algorithm 2 (page 18). Proposition 3.21 provides, on the set M_0, an upper bound for the augmentation of the random variable $\mathcal{KL}\left(Y_i,Y_j\right)$ compared to $\mathcal{KL}(\mathcal{Y}_i,\mathcal{Y}_j)$. It justifies the given choice of λ_φ for the iteration step $k=1$ if $h^{(0)}$ satisfies $\overline{w}_{ij}^{(0)}=0$ for all $i,j\in\{1,...,n\}$ with $i\neq j$. We seek for a generalization to other choices of $h^{(0)}$ and the subsequent iteration steps.

The adaptive estimator is defined as a weighted mean of the observations. Therefore, for all $k\in\{0,...,k^*\}$, it holds

$$M_0\subseteq\bigcap_{i=1}^{n}\{\tilde{\theta}_i^{(k)},\tilde{\vartheta}_i^{(k)}\in\Theta_\varkappa\},$$

where M_0 is as in Proposition 3.21. This enables on M_0 the application of Equation (2.2) (page 11) and Lemma 2.8 with respect to the adaptive estimates. We distinguish the same cases as in the proof of Proposition 3.21, recall Equations (3.5) and (3.6) and the corresponding upper bounds in Equations (3.4) and (3.8). For the sake of brevity, we summarize both cases in one equation, using the operation $\pm$. Then, we get on the set M_0 in a uniform manner as in the proof of Proposition 3.21 that

$$\begin{aligned}s_{ij}^{(k)} &= \tilde{N}_i^{(k-1)}\mathcal{KL}\left(\tilde{\theta}_i^{(k-1)},\tilde{\theta}_j^{(k-1)}\right)\\ &\leq \varkappa^2\tilde{N}_i^{(k-1)}\Big[\Big|\mathcal{KL}^{1/2}\left(\tilde{\theta}_i^{(k-1)},\mathcal{E}\tilde{\theta}_i^{(k-1)}\right)\pm\mathcal{KL}^{1/2}\left(\tilde{\theta}_j^{(k-1)},\mathcal{E}\tilde{\theta}_j^{(k-1)}\right)\Big|\\ &\qquad +\mathcal{KL}^{1/2}\left(\mathcal{E}\tilde{\theta}_i^{(k-1)},\mathcal{E}\tilde{\theta}_j^{(k-1)}\right)\Big]^2,\end{aligned}\tag{3.9}$$

where $\mathcal{E}\tilde{\theta}_i^{(k)}$ is as in Notation 3.13. The variability of the parameter function $\theta(.)$ effects the third summand, which satisfies by Equation (2.2) (page 11) and the convexity of the Kullback-Leibler divergence with respect to the first argument that

$$\max_{i,j} \mathcal{KL}\left(\mathcal{E}\tilde{\theta}_i^{(k-1)}, \mathcal{E}\tilde{\theta}_j^{(k-1)}\right) \leq \varkappa^2 \max_{i,j} \mathcal{KL}\left(\theta_i, \theta_j\right) \leq \varkappa^2 \varphi_0^2$$

with $\varphi_0 = \varphi / \max_i \sqrt{\overline{N}_i^{(k^*)}}$. The remaining term

$$\sqrt{\tilde{N}_i^{(k-1)}} \left| \mathcal{KL}^{1/2}\left(\tilde{\theta}_i^{(k-1)}, \mathcal{E}\tilde{\theta}_i^{(k-1)}\right) \pm \mathcal{KL}^{1/2}\left(\tilde{\theta}_j^{(k-1)}, \mathcal{E}\tilde{\theta}_j^{(k-1)}\right) \right| \tag{3.10}$$

forms the inhomogeneous analog of

$$\sqrt{\tilde{N}_i^{(k-1)}} \left| \mathcal{KL}^{1/2}\left(\tilde{\vartheta}_i^{(k-1)}, \vartheta\right) \pm \mathcal{KL}^{1/2}\left(\tilde{\vartheta}_j^{(k-1)}, \vartheta\right) \right|. \tag{3.11}$$

However, the corresponding probability distributions cannot be compared as for the single observations since the probability distributions of $\tilde{\vartheta}_l^{(k-1)}$ and $\tilde{\theta}_l^{(k-1)}$, $l = i, j$, may differ. Nevertheless, it follows in the same lines as at the end of the proof of Proposition 3.21 that

$$\left| \mathcal{KL}^{1/2}\left(\tilde{\vartheta}_i^{(k-1)}, \vartheta\right) \pm \mathcal{KL}^{1/2}\left(\tilde{\vartheta}_j^{(k-1)}, \vartheta\right) \right| \leq \varkappa \mathcal{KL}^{1/2}\left(\tilde{\vartheta}_i^{(k-1)}, \tilde{\vartheta}_j^{(k-1)}\right).$$

Hence, Equation (3.11) is controlled by $\sqrt{\lambda}$, up to the factor $\varkappa$. Similarly, Equation (3.10) mainly depends on the randomness of the observations. Admittedly, this cannot be proven due to the impact of the adaptive weights which are influenced by the variability of the inhomogeneous parameter function. For instance, for $\mathcal{P} = \{\mathcal{N}(\theta, \sigma^2)\}_{\theta \in \Theta}$, it holds

$$\mathcal{KL}\left(\tilde{\theta}_i^{(k)}, \mathcal{E}\tilde{\theta}_i^{(k)}\right) = \left(\sum_{j=1}^n (Y_j - \theta_j) \tilde{w}_{ij}^{(k)} / \tilde{N}_i^{(k)} \right)^2 / 2\sigma^2$$

and, for $\mathcal{P} = \{\mathrm{Exp}(1/\theta)\}_{\theta \in \Theta}$, we get

$$\mathcal{KL}\left(\tilde{\theta}_i^{(k)}, \mathcal{E}\tilde{\theta}_i^{(k)}\right) = \left(\frac{\sum_{j=1}^n Y_j \tilde{w}_{ij}^{(k)} / \tilde{N}_i^{(k)}}{\sum_{j=1}^n \theta_j \tilde{w}_{ij}^{(k)} / \tilde{N}_i^{(k)}} \right) - 1 - \ln \left(\frac{\sum_{j=1}^n Y_j \tilde{w}_{ij}^{(k)} / \tilde{N}_i^{(k)}}{\sum_{j=1}^n \theta_j \tilde{w}_{ij}^{(k)} / \tilde{N}_i^{(k)}} \right).$$

These terms cannot be compared with their homogeneous counterparts, impeding a theoretical proof of the parameter choice in Claim 3.22.

Instead, we follow an inductive argumentation, considering the relation to the non-adaptive estimator. The initialization of the algorithm by the non-adaptive estimator serves as the base clause. Assuming that the adaptive weights in iteration step k are, with high probability, similar to the non-adaptive ones, we get that the divergence $\mathcal{KL}(\tilde{\theta}_i^{(k)}, \mathcal{E}\tilde{\theta}_i^{(k)})$ behaves similar to $\mathcal{KL}(\overline{\theta}_i^{(k)}, \mathbb{E}\overline{\theta}_i^{(k)})$. Additionally, we know from Theorems 2.13 and 2.24 that $\mathcal{KL}(\overline{\vartheta}_i^{(k)}, \vartheta)$ and $\mathcal{KL}(\overline{\theta}_i^{(k)}, \mathbb{E}\overline{\theta}_i^{(k)})$ satisfy, in probability, the same rate of convergence. The divergence $\mathcal{KL}(\overline{\vartheta}_i^{(k)}, \vartheta)$ relates via the homogeneous propagation condition to the divergence $\mathcal{KL}(\tilde{\vartheta}_i^{(k)}, \vartheta)$ and, as a consequence, to Equation (3.11), which we controlled by the constant $\varkappa\sqrt{\lambda}$. In other words, we construct a relation between the homogeneous and the inhomogeneous setting via the non-adaptive estimates. This motivates together with Proposition 3.21 and the invariance of the functions $p_\vartheta^{(l)}$, $l = 1, 2, 3$, with respect to $\theta \in \Theta$, the supposition that the impact of the

variability of the parameter function on Equation (3.10) is sufficiently small such that $\varkappa\sqrt{\lambda}$ can still control it. Then, we may conclude that the choice

$$\lambda_\varphi \geq \varkappa^4 \left[\sqrt{\lambda}+\varphi\right]^2$$

ensures in the next iteration step $k+1$ the similarity of the adaptive and the non-adaptive weights, yielding on M_0 the desired behavior of $\hat{\mathfrak{Z}}_{\lambda_\varphi}$. The restriction to the set M_0 leads to an increased probability level of $\epsilon + 2p_\varkappa$ since $\mathbb{P}(M_0^c) \leq 2p_\varkappa$. □

Chapter 4

The simplified algorithm under the propagation condition

Next we will analyze the behavior of the simplified Propagation-Separation Approach, recall Algorithm 2 (page 18). First, we will show the separation property, which illustrates the impact of the adaptive weights and particularly of the adaptation bandwidth λ. Then, we will consider a homogeneous setting, where the propagation property and a certain stability of estimates will follow as direct consequences of the homogeneous propagation condition in Section 3.1. These are the main properties of the algorithm, where the subsequent results will be based on.

For piecewise constant parameter functions with sufficiently sharp discontinuities, the separation property restricts smoothing to the respective homogeneous regions. By the propagation condition, this will again yield a propagation property. This illustrates the coaction of propagation and separation during iteration. We will need to ensure the separation property as well as propagation on the whole design space and for all iteration steps. In practice, the effects may be negligible, but the consequences of a failed separation at a certain number of design points cannot be quantified exactly, which will lead to restrictive conditions. Taking advantage of the inhomogeneous propagation condition in Section 3.2, we will establish similar results for (piecewise) bounded parameter functions.

Then, we will consider the case of a misspecified structural assumption. In Section 4.3, we will introduce an *associated step function*. The numerical examples in Chapter 5 suggest for any piecewise smooth parameter function that the estimation function, which results from the pointwise estimates of Algorithm 2, approaches the associated step function during iteration. However, there are several reasons which impede a theoretical proof of the convergence of the algorithm as we will discuss in Section 5.4.

Most of the results will be based on Assumption A1 (page 10). Therefore, in Section 4.4, we will generalize the assumed exponential family model, and we will provide further details for the application of the Propagation-Separation Approach on observations which satisfy the local likelihood model in Notation 2.1, but where the parametric family of probability distribution does not form an exponential family. We will assume throughout this chapter that we have identified λ and ϵ such that the propagation condition holds. Longer proofs are collected in the final Section 4.5.

4.1 Separation

For considerably different parameter values, the corresponding adaptive weights become zero, see Proposition 4.1 below. The result is similar to the first part of [Polzehl and Spokoiny, 2006, Thm. 5.9]. It implies that different homogeneous regions with sufficiently sharp discontinuities will be separated by the algorithm. In particular, we will see that the lower bound for the discontinuities which allows exact separation of the distinct regions mainly depends on the adaptation bandwidth λ and the achieved quality of estimation in the previous iteration step. Recall that the adaptive weights $\tilde{w}_{ij}^{(k)}$ and their sum $\tilde{N}_i^{(k)}$ are random. In the proof, we apply Lemma 2.8, which requires that $\tilde{\theta}_i^{(k)} \in \Theta_{\varkappa}$.

Proposition 4.1 (Separation property)**.** *Suppose Assumption A1, and consider two design points $X_{i_1}, X_{i_2} \in \mathcal{X}$. Assume that the realization at hand satisfies, at these points in iteration step k, the estimation accuracy $\mathcal{KL}(\tilde{\theta}_{i_m}^{(k)}, \theta_{i_m}) \leq z_m^{(k)} := z/\overline{N}_{i_m}^{(k)}$ for some constant $z > 0$ and $\theta_{i_m}, \tilde{\theta}_{i_m}^{(k)} \in \Theta_{\varkappa}$, $m = 1, 2$, with $\varkappa \geq 1$ fixed and $\Theta_{\varkappa}$ as in Notation 2.6. If additionally*

$$\mathcal{KL}^{1/2}(\theta_{i_1}, \theta_{i_2}) > \varkappa \left(\sqrt{\lambda / \tilde{N}_{i_1}^{(k)}} + \sqrt{z_1^{(k)}} + \sqrt{z_2^{(k)}} \right), \tag{4.1}$$

then it holds $\tilde{w}_{i_1 i_2}^{(k+1)} = 0$.

Proof. The support of the adaptation kernel K_{ad} equals $[0, 1)$, see Notation 2.14. Therefore, it suffices to show that the statistical penalty in Algorithm 2 satisfies $s_{i_1 i_2}^{(k+1)} > \lambda$. Lemma 2.8 yields, for $\mathcal{KL}(\tilde{\theta}_{i_m}^{(k)}, \theta_{i_m}) \leq z_m^{(k)}$ with $m = 1, 2$, that

$$\mathcal{KL}^{1/2}\left(\tilde{\theta}_{i_1}^{(k)}, \tilde{\theta}_{i_2}^{(k)}\right) \geq \varkappa^{-1} \mathcal{KL}^{1/2}(\theta_{i_1}, \theta_{i_2}) - \sqrt{z_1^{(k)}} - \sqrt{z_2^{(k)}},$$

and we get

$$s_{i_1 i_2}^{(k+1)} \geq \tilde{N}_{i_1}^{(k)} \left[\varkappa^{-1} \sqrt{\mathcal{KL}(\theta_{i_1}, \theta_{i_2})} - \sqrt{z_1^{(k)}} - \sqrt{z_2^{(k)}} \right]^2 > \lambda$$

by Equation (4.1). □

Remark 4.2. The lower bound (4.1) holds if

$$\mathcal{KL}^{1/2}(\theta_{i_1}, \theta_{i_2}) > 3\varkappa \cdot \frac{\max\left\{\sqrt{\lambda}, \sqrt{z}\right\}}{\min\left\{\sqrt{\tilde{N}_{i_1}^{(k)}}, \sqrt{\overline{N}_{i_1}^{(k)}}, \sqrt{\overline{N}_{i_2}^{(k)}}\right\}}.$$

This emphasizes the impact of the involved sample sizes. In Section 5.3.1, we will show some numerical examples which illustrate the separation property.

Example 4.3. For $\mathcal{P} := \{\mathcal{N}(\theta, \sigma^2)\}_{\theta \in \mathbb{R}}$, we know from Example 2.17 that $\varkappa = 1$. Moreover, it follows from Example 2.7 that the lower bound (4.1) is satisfied if and only if

$$|\theta_{i_1} - \theta_{i_2}| > \sigma\sqrt{2} \left(\sqrt{\lambda / \tilde{N}_{i_1}^{(k)}} + \sqrt{z_1^{(k)}} + \sqrt{z_2^{(k)}} \right),$$

where we used the notation of Proposition 4.1. Obviously, this scales with the noise standard deviation σ.

4.2 Propagation and stability in the case of sharp discontinuities

We start by considering a homogeneous setting, that is the case of a parametric model. Then, we will study piecewise constant functions with sharp discontinuities, and finally we will extend our results to piecewise bounded functions.

4.2.1 Homogeneous setting

For a homogeneous setting, we show that the propagation condition in Definition 3.2 yields with Theorem 2.13 an exponential bound for $\mathbb{P}(\overline{N}_i^{(k)}\mathcal{KL}(\tilde{\theta}_i^{(k)},\theta) > z)$, the excess probability of the Kullback-Leibler divergence between the adaptive estimator $\tilde{\theta}_i^{(k)}$ and the true parameter θ.

Proposition 4.4 (Propagation and stability under homogeneity)**.** *Suppose $\theta(.) \equiv \theta$, Assumption A1, and let the adaptation bandwidth λ be chosen in accordance with the homogeneous propagation condition at level ϵ for $\theta \in \Theta$. Then, for each $i \in \{1,...,n\}$, $k \in \{0,...,k^*\}$, and all $z > 0$, it holds*

$$\mathbb{P}\left(\overline{N}_i^{(k)}\mathcal{KL}\left(\tilde{\theta}_i^{(k)},\theta\right) > z\right) \leq \max\left\{2e^{-z},\epsilon\right\}.$$

In particular, for all $k' \leq k$, we get

$$\mathbb{P}\left(\overline{N}_i^{(k)}\mathcal{KL}\left(\tilde{\theta}_i^{(k)},\theta\right) > z\right) \leq \max\left\{\mathbb{P}\left(\overline{N}_i^{(k')}\mathcal{KL}\left(\tilde{\theta}_i^{(k')},\theta\right) > z\right),\epsilon\right\}. \tag{4.2}$$

Proof. Equation (4.2) follows from the homogeneous propagation condition, which ensures that the function $\mathfrak{Z}_\lambda(.,p;\theta,i)$ in Notation 3.1 is non-increasing for all $p \in (\epsilon,1)$ and every $i \in \{1,...,n\}$. Since, see (2) in Algorithm 2, we have $\tilde{\theta}_i^{(0)} = \overline{\theta}_i^{(0)}$, this yields

$$\begin{aligned}\mathbb{P}\left(\overline{N}_i^{(k)}\mathcal{KL}\left(\tilde{\theta}_i^{(k)},\theta\right) > z\right) &\overset{\text{Eq. (4.2)}}{\leq} \max\left\{\mathbb{P}\left(\overline{N}_i^{(0)}\mathcal{KL}\left(\overline{\theta}_i^{(0)},\theta\right) > z\right),\epsilon\right\} \\ &\overset{\text{Thm. 2.13}}{\leq} \max\left\{2e^{-z},\epsilon\right\},\end{aligned}$$

which leads to the assertion. □

Basically, this result provides the root-n consistency of $\tilde{\theta}_i^{(k)}$ up to a log-factor if $z := \mu\log(n)$ and $\epsilon := c_\epsilon n^{-\mu}$, where $c_\epsilon > 0$ and $\mu > 2$, and with $h^{(k^*)}$ sufficiently large such that $\overline{N}_i^{(k^*)}$ is of order n. More precisely, these conditions provide with Proposition 4.4 that

$$\mathbb{P}\left(\exists i : \mathcal{KL}\left(\tilde{\theta}_i^{(k)},\theta\right) > \mu\log(n)/\overline{N}_i^{(k)}\right) \leq \max\{2,c_\epsilon\}\cdot n^{-1},$$

and we recall the quadratic approximation of the Kullback-Leibler divergence in Equation (2.4) (page 11). However, asymptotic results are problematic in this context as we will discuss in § 5.4.2. Therefore, we prefer to consider Proposition 4.4 as an error bound for the fixed iteration step k. It provides exponential bounds for the estimation bias and the variance of the relative frequency which we proposed in § 3.1.3 as estimator for the propagation condition.

Lemma 4.5. *We consider the same setting as in § 3.1.3, and presume the conditions of Proposition 4.4 to be satisfied. Then, for every $i \in \{1,...,n_0\}$, it holds*

$$\left|\mathbb{E}\left[\hat{p}_\lambda^{(k)}(z)\right] - \mathbb{P}\left(\overline{N}_i^{(k)}\mathcal{KL}(\tilde{\theta}_i^{(k)}(\lambda),\theta) > z\right)\right| \leq \max\{2e^{-z},\epsilon\}$$

and

$$\mathrm{Var}\left[\hat{p}_\lambda^{(k)}(z)\right] \leq \max\{2e^{-z},\epsilon\}. \tag{4.3}$$

4.2.2 Piecewise constant functions

Next we consider a piecewise homogeneous setting with sharp discontinuities, formally described in Assumption A2, below. In this case, the separation property restricts smoothing to the homogeneous regions, and we get a certain propagation property. We introduce some auxiliary notations.

Notation 4.6.

- *For any set M, let $\mathfrak{C}(M)$ be the smallest connected space which includes the respective set M, that is*

$$\mathfrak{C}(M) := \bigcap \{M_c : M_c \text{ is a connected space and } M \subseteq M_c\}.$$

- *We call the discrete set $M := \{X_j\}_{j=1}^m \subseteq \mathcal{X}$ connected if we have*

$$X_j \in M \text{ if and only if } X_j \in \mathfrak{C}(M), \qquad X_j \in \mathcal{X}.$$

- *We call the connected set $M := \{X_j\}_{j=1}^m \subseteq \mathcal{X}$ convex if $\mathfrak{C}(M)$ is convex.*

Then, the setting is described by the following structural assumption.

Assumption A2. *There is a non-trivial partition $\mathcal{V} := \{\mathcal{V}_i\}_i$ of $\mathcal{X}$ into maximal homogeneity regions, that is for each $X_i \in \mathcal{X}$, there are a convex neighborhood $\mathcal{V}_i \subseteq \mathcal{X}$ and a constant $\phi_i > 0$ such that*

$$\begin{cases} \mathcal{KL}(\theta_i, \theta_j) = 0 & \text{for all } X_j \in \mathcal{V}_i, \\ \mathcal{KL}(\theta_i, \theta_j) > \phi_i^2 & \text{for all } X_j \notin \mathcal{V}_i. \end{cases}$$

The convexity of the neighborhoods $\{\mathcal{V}_i\}_{i=1}^n$ ensures the comparability of the homogeneous setting in Proposition 4.4 and the setting within each of these neighborhoods. A violation of this condition may lead to a different behavior of the adaptive estimator due to the changed impact of the non-adaptive weights. The specific form of the homogeneity regions does not matter since Theorem 2.13 and hence the probability condition do not depend thereon.

We deduce the propagation property for the present case of piecewise homogeneity. Here, we should take into account that the considered neighborhood $U_i^{(k)} = \{X_j \in \mathcal{X} : \overline{w}_{ij}^{(k)} > 0\}$ might be much larger than the respective homogeneity region $\mathcal{V}_i$. Obviously, the divergence $\mathcal{KL}(\tilde{\theta}_i^{(k)}, \theta_i)$ cannot converge with rate $\overline{N}_i^{(k)}$ in this case. Therefore, we introduce the notion of the effective sample size $\overline{n}_i^{(k)}$.

Notation 4.7. *We define, for all $i \in \{1, ..., n\}$ and every $k \in \{0, ..., k^*\}$, the effective sample size and its local minimum*

$$\overline{n}_i^{(k)} := \sum_{X_j \in \mathcal{V}_i \cap U_i^{(k)}} \overline{w}_{ij}^{(k)} \qquad \text{and} \qquad n_i^{(k)} := \min_{X_j \in U_i^{(k)}} \overline{n}_j^{(k)}.$$

As it turns out, the quantities $n_i^{(k)}$ determine a lower bound for the stepsizes ϕ_i which allows the detection of the associated discontinuity by the algorithm. During the first iteration steps, it holds $\overline{n}_i^{(k)} = \overline{N}_i^{(k)}$. The quotient $\overline{n}_i^{(k)} / \overline{N}_i^{(k)}$ decreases when $U_i^{(k)}$ becomes larger than $\mathcal{V}_i$. In the following theorem, we consider two events. On the first one, $\mathcal{B}^{(k)}(z)$, the estimation error is bounded from above, and on the second one, $M^{(k)}(z)$, the discontinuities are sufficiently sharp for separation.

Notation 4.8. *Let the constants $\phi_i > 0$, $i \in \{1,...,n\}$, be as in Assumption A2, and fix $\lambda > 0$ and $z > 0$. Additionally, we recall Notation 2.6 and choose $\varkappa \geq 1$ such that $\{\theta_i\}_{i=1}^n \in (\Theta_\varkappa)^n$. Then, we set*

$$\mathcal{B}^{(k)}(z) := \bigcap_{i=1}^{n} \left\{ \overline{n}_i^{(k)} \mathcal{KL}(\tilde{\theta}_i^{(k)}, \theta_i) \leq z \right\}, \qquad k \in \{0,...,k^*\}, \tag{4.4}$$

and

$$M^{(k)}(z) := \begin{cases} \bigcap_{k'=0}^{k-1} \bigcap_{i=1}^{n} \left\{ \phi_i > \varkappa \left[\sqrt{\lambda / \tilde{N}_i^{(k')}} + 2\sqrt{z / n_i^{(k')}} \right] \right\}, & k \in \{1,...,k^*\}, \\ \Omega, & k = 0, \end{cases} \tag{4.5}$$

where Ω is as in Notation 2.1.

Theorem 4.9 (Propagation property under piecewise homogeneity)**.** *Suppose Assumptions A1 and A2, and let the bandwidth λ be chosen in accordance with the propagation condition at level ϵ for all θ_i, $i \in \{1,...,n\}$. Additionally, we choose $h^{(0)} > 0$ sufficiently small such that $\overline{w}_{ij}^{(k)} = 0$ for all $i, j \in \{1,...,n\}$ with $i \neq j$. If $z > 0$ and $k \in \{0,...,k^*\}$ satisfy that $\mathbb{P}(M^{(k)}(z)) > 0$, then it holds*

$$\mathbb{P}\left(\mathcal{B}^{(k)}(z) | M^{(k)}(z) \right) \geq 1 - \frac{p_\varkappa + (k+1)\max\{2ne^{-z}, n\epsilon\}}{\mathbb{P}\left(M^{(k)}(z)\right)}, \tag{4.6}$$

where $p_\varkappa$ is as in Notation 2.16 with $\Theta^* \subseteq \Theta$ and $\varphi_0 > 0$ appropriate.

Remark 4.10.

- In Equation (4.6), we observe an additional factor $(k+1)$, which appeared in the propagation property of Polzehl and Spokoiny [2006] as well, see Equation (5.4) in § 5.4.2, below. This factor results from the proof only, and it might be avoidable. In particular, we notice that the given bound does not need to be sharp. The above theorem provides a meaningful result for $z \geq q\log(n)$ and small values of ϵ or, at best, $\epsilon := c_\epsilon n^{-q}$ with $c_\epsilon > 0$ and $q > 1$. It is the better the larger $\mathbb{P}(M^{(k)}(z))$, which increases with the discontinuities ϕ_i and the minimal effective sample size $n_i^{(k)}$. Additionally, the lower bound improves the smaller $p_\varkappa$. Recall from Example 2.17 that $p_\varkappa = 0$ for Gaussian and log-normal distributed observations.
- The applied separation property depends via the statistical penalty on the estimation quality of all data within the local neighborhood $U_i^{(k)}$. Therefore, the extension of the smallest homogeneous region, measured by the minimal effective sample size $n_i^{(k)}$, determines the lower bound (4.5) for the discontinuities which provides an exact separation of the distinct homogeneous regions. This bound is closely related to Equation (4.1) (page 52), which only involves two points and where the term $2/(n_i^{(k)})^{1/2}$ from Equation (4.5) is replaced by
$$\left(1/\sqrt{\overline{N}_{i_1}^{(k)}} + 1/\sqrt{\overline{N}_{i_2}^{(k)}} \right),$$
having the same effect.

4.2.3 Piecewise bounded functions

By means of the inhomogeneous propagation condition in Definition 3.15 and Theorem 2.24, we can establish similar results for (piecewise) bounded parameter functions. The following

proposition provides the inhomogeneous analog of Proposition 4.4. It requires slightly different assumptions and yields a different exponent in the exponential bound of the excess probability, see Equation (4.7).

Proposition 4.11 (Propagation and stability under bounded variability)**.** *Let Assumption A1 be fulfilled, and let the adaptation bandwidth λ be chosen in accordance with the inhomogeneous propagation condition at probability level $\epsilon > 0$ and variability level $\varphi_0 > 0$ for some set $\Theta^* \subseteq \Theta$ satisfying $\{\theta_i\}_{i=1}^n \in (\Theta^*)^n$. Additionally, we recall Notation 2.6, and we choose $\varkappa \geq 1$ sufficiently large such that $\Theta^* \subseteq \Theta_\varkappa$. If $\mathcal{KL}(\theta_i, \theta_j) \leq \varphi^2 / \max_{j'} \overline{N}_{j'}^{(k_0)} = \varphi_0^2$ holds for all $i, j \in \{1, ..., n\}$ and some fixed $k_0 \in \{1, ..., k^*\}$, then we get*

$$\mathbb{P}\left(\overline{N}_i^{(k)} \mathcal{KL}\left(\tilde{\theta}_i^{(k)}, \theta_i\right) > z\right) \leq \max\left\{2e^{-\left[\sqrt{z}/\varkappa - \varphi\right]^2/\varkappa^2}, \epsilon\right\} + \breve{p}_{\varkappa,0} \tag{4.7}$$

for each $i \in \{1, ..., n\}$, $k \in \{0, ..., k_0\}$, and all $z > \varkappa^2 \varphi^2$, where $\breve{p}_{\varkappa,0}$ is as in Notation 2.21. In particular, for all $k_1 \leq k_2 \leq k_0$, it holds

$$\mathbb{P}\left(\overline{N}_i^{(k_2)} \mathcal{KL}\left(\tilde{\theta}_i^{(k_2)}, \mathcal{E}\tilde{\theta}_i^{(k_2)}\right) > z\right) \leq \max\left\{\mathbb{P}\left(\overline{N}_i^{(k_1)} \mathcal{KL}\left(\tilde{\theta}_i^{(k_1)}, \mathcal{E}\tilde{\theta}_i^{(k_1)}\right) > z\right), \epsilon\right\}, \tag{4.8}$$

where $\mathcal{E}\tilde{\theta}_i^{(k)} = \sum_j \tilde{w}_{ij}^{(k)} \theta_j / \tilde{N}_i^{(k)}$ as in Notation 3.13.

Next let us consider piecewise bounded functions with sharp discontinuities.

Assumption A3. *Suppose the existence of a non-trivial partition $\mathcal{V} := \{\mathcal{V}_i\}_i$ of $\mathcal{X}$ such that, for every $X_i \in \mathcal{X}$, there are constants $\phi_i > \varphi_0 \geq 0$ and a convex neighborhood $\mathcal{V}_i \subseteq \mathcal{X}$ which satisfy*

$$\begin{cases} \mathcal{KL}(\theta_i, \theta_j) \leq \varphi_0^2 & \text{for all } X_j \in \mathcal{V}_i, \\ \mathcal{KL}(\theta_i, \theta_j) > \phi_i^2 & \text{for all } X_j \notin \mathcal{V}_i. \end{cases}$$

Recall the effective sample size $\overline{n}_i^{(k)}$ and its minimum $n_i^{(k)}$ in Notation 4.7, and the events $\mathcal{B}^{(k)}(z)$ and $M^{(k)}(z)$ in the subsequent Equations (4.4) and (4.5).

Theorem 4.12 (Propagation property under piecewise boundedness)**.** *Suppose Assumptions A1 and A3 to be satisfied. Additionally, let the adaptation bandwidth λ, the constant $\varkappa \geq 1$, and the corresponding set $\Theta_\varkappa \subseteq \Theta$ be as in Proposition 4.11 and $h^{(0)} > 0$ as in Theorem 4.9. Finally, we fix some iteration step $k_0 \in \{0, ..., k^*\}$ and some constant $\varphi \geq 0$ such that $\varphi^2 / \max_i \overline{n}_i^{(k_0)} = \varphi_0^2$. If $z > \varkappa^2 \varphi^2$ satisfies $\mathbb{P}\left(M^{(k_0)}(z)\right) > 0$, then it holds*

$$\mathbb{P}\left(\mathcal{B}^{(k_0)}(z) | M^{(k_0)}(z)\right) \geq 1 - \frac{\breve{p}_{\varkappa,0} + (k_0 + 1) \max\{2ne^{-\left[\sqrt{z}/\varkappa - \varphi\right]^2/\varkappa^2}, n\epsilon\}}{\mathbb{P}\left(M^{(k_0)}(z)\right)},$$

where $\breve{p}_{\varkappa,0}$ is as in Notation 2.21.

We skip the proof since the assertion follows in the same manner as in the proof of Theorem 4.9, using Proposition 4.11 instead of Proposition 4.4.

Remark 4.13. Theorem 4.12 yields a meaningful result for $z \geq \varkappa^2 [\varkappa \sqrt{q \log(n)} + \varphi]^2$ and small values of ϵ or, at best, $\epsilon := c_\epsilon n^{-q}$ with $q > 1$ and $c_\epsilon > 0$. The established lower bound is not sharp.

4.3 Consequences of a violated structural assumption

All previous results only hold for parameter functions with sharp discontinuities. What happens in the case of a violated structural assumption? In § 2.2.4, we observed for simulated examples with Gaussian distributed observations that the estimation function resulted in a step function in the case of a piecewise constant parameter function and as well for a piecewise smooth function. Therefore, we will introduce a specific step function, that we will call the *associated step function* of the Propagation-Separation Approach. Then, we will establish an upper bound for the pointwise Kullback-Leibler divergence between the adaptive estimator in Algorithm 2 (page 18) and the corresponding value of the associated step function.

Applying the simplified Propagation-Separation Approach with some fixed adaptation bandwidth $\lambda > 0$ provides, for every $k \in \{0, ..., k^*\}$, a set of adaptive weights $\{\tilde{w}_{ij}^{(k)}\}_{i,j=1}^{n}$. In particular, this yields the subsets

$$\mathcal{H}_i^{(k)} := \left\{ \mathcal{X}_j \in \mathcal{X} : \tilde{w}_{il}^{(k+1)} > 0 \text{ if and only if } \tilde{w}_{jl}^{(k+1)} > 0 \text{ for all } X_l \in \mathcal{X} \right\}, \tag{4.9}$$

where we set $\tilde{w}_{ij}^{(k^*+1)} := \overline{w}_{ij}^{(k^*)} \cdot K_{\mathrm{ad}}(s_{ij}^{(k^*)}/\lambda)$. They are based on an equivalence relation, and we get, for every parameter function $\theta(.)$, a well-defined partition $\{H_l^{(k)}\}_{l=1}^{m}$ of the design space $\mathcal{X}$ into $m \leq n$ regions. We introduce a step function whose steps match this partition $\{H_l^{(k)}\}_{l=1}^{m}$.

Definition 4.14. *Let* $\mathbf{1}$ *denote the indicator function, and let* $\theta_l^{(k)}$ *be the mean value of the* n_l *estimates* $\tilde{\theta}_{l_j}^{(k)}$ *resulting from Algorithm 2 at the design points* $\{X_{l_j}\}_{j=1}^{n_l}$ *which form the region* $H_l^{(k)}$*. Then, we call the piecewise constant function*

$$\breve{\theta}^{(k)}(X_i) := \sum_{l=1}^{m} \theta_l^{(k)} \mathbf{1}_{H_l^{(k)}}(X_i) \quad \textit{with} \quad \theta_l^{(k)} := \frac{1}{n_l} \sum_{j=1}^{n_l} \tilde{\theta}_{l_j}^{(k)} \tag{4.10}$$

the associated step function of $\theta(.)$ *in step* k*. For* $i \in \{1, ..., n\}$ *and* $k \in \{1, ..., k^*\}$*, we set* $\breve{\theta}_i^{(k)} := \breve{\theta}^{(k)}(X_i)$*.*

Lemma 4.15. *Under the notation in Definition 4.14, it holds*

$$\mathcal{KL}\left(\breve{\theta}_i^{(k)}, \tilde{\theta}_i^{(k)}\right) \leq \max\{\lambda / \tilde{N}_j^{(k)} : X_j \in \mathcal{H}_i^{(k)}\}$$

for all $i \in \{1, ..., n\}$ *and* $k \in \{0, ..., k^*\}$*.*

Proof. We know from Lemma 2.5 that the Kullback-Leibler divergence is convex with respect to the first argument. Therefore, it holds

$$\mathcal{KL}\left(\breve{\theta}_i^{(k)}, \tilde{\theta}_i^{(k)}\right) \leq \max\left\{\mathcal{KL}\left(\tilde{\theta}_j^{(k)}, \tilde{\theta}_i^{(k)}\right) : X_j \in \mathcal{H}_i^{(k)}\right\}.$$

Since $X_j \in \mathcal{H}_i^{(k)}$ implies $\tilde{w}_{ji}^{(k+1)} > 0$, we have $\mathcal{KL}\left(\tilde{\theta}_j^{(k)}, \tilde{\theta}_i^{(k)}\right) \leq \lambda / \tilde{N}_j^{(k)}$, which leads to the assertion. □

In § 5.3.2, we will illustrate the formation of the associated step function during iteration. The corresponding simulations suggest its immutability for sufficiently large bandwidths. Additionally, we will see that, in the presented examples, the sets $\{\mathcal{X}_j \in \mathcal{X} : \tilde{w}_{ij}^{(k)} > 0\}$ with $i \in \{1, ..., n\}$ form a well-defined partition of the design space $\mathcal{X}$ if $h^{(k)}$ is sufficiently large. However, both heuristic observations could not be theoretically justified for reasons that we will discuss in § 5.4.1.

4.4 Generalization of the setting

We will extend our theoretical results to the case of a parametrized exponential family model. Then, we will discuss further details concerning the application of the Propagation-Separation Approach in the case of a local likelihood model without restrictions.

4.4.1 Parametrized exponential families

Assumption A1 (page 10) and hence the whole study were restricted to the parametrization where $\mathbb{E}_\theta[T(Y)] = \theta$. Which modifications and additional assumptions are required in order to take the previous results over to the case where $t(\theta) := \mathbb{E}_\theta[T(Y)]$ is some invertible function?

Assumption A4 (Parametrized exponential family model)**.** *Consider the setting in Notation 2.1 and let $t : \Theta \to \Theta$ denote an invertible and continuously differentiable function. We assume that the parametric family $\mathcal{P}^{(t)} = \{\mathbb{P}_\vartheta^{(t)}\}_{\vartheta\in\Theta}$ is an exponential family with a convex parameter set Θ and two functions $C_t, B_t \in C^2(\Theta, \mathbb{R})$ such that*

$$p_t(y,\vartheta) := d\mathbb{P}_\vartheta^{(t)}/d\mathbb{P}(y) = p(y)\exp\left[T(y)C_t(\vartheta) - B_t(\vartheta)\right], \qquad \vartheta \in \Theta,$$

where $\mathbb{P}$ denotes a dominating σ-finite measure, $T : \mathcal{Y} \to \mathbb{R}$ is a sufficient statistic, p is some non-negative function on $\mathcal{Y}$, and C_t is strictly monotonic increasing. For the parameter ϑ, it holds

$$\int p_t(y,\vartheta)\mathbb{P}(dy) = 1 \quad \text{and} \quad \mathbb{E}_\vartheta^{(t)}[T(Y)] = \frac{B_t'(\vartheta)}{C_t'(\vartheta)} = t(\vartheta).$$

Corollary 4.16. *Let Assumption A4 be satisfied.*

- *Reparametrization with $\theta := t(\vartheta)$ yields*

$$\mathcal{KL}(\vartheta_1,\vartheta_2) = \mathcal{KL}(\theta_1,\theta_2) \qquad \text{for all } \vartheta_1,\vartheta_2 \in \Theta.$$

- *If $t(\vartheta)$ is linear in ϑ, then it follows, for $\tilde{\vartheta} := t^{-1}\left(\tilde{\theta}\right)$, that*

$$\mathcal{KL}(\tilde{\vartheta}, \mathbb{E}\tilde{\vartheta}) = \mathcal{KL}(\tilde{\theta}, \mathbb{E}\tilde{\theta}),$$

  where $\tilde{\theta}$ denotes an estimate of θ.
- *It holds for the weighted MLE $\theta_i^{(\mathrm{MLE})}$ in Equation (2.5) (page 13) that $\theta_i^{(\mathrm{MLE})} = t^{-1}(\overline{\theta}_i)$, where $\overline{\theta}_i = \sum_{j=1}^n \overline{w}_{ij}T(Y_j)/\overline{N}_i$ as in Equation (2.6).*

Remark 4.17. In Assumption A1, we required the unbiasedness of the parameter, $\theta = \mathbb{E}_\theta[T(Y)]$ for all $\theta \in \Theta$. Under Assumption A4, this can be achieved via the reparametrization $\theta := t(\vartheta)$, where $t(\vartheta) = \mathbb{E}_\vartheta[T(Y)]$. Alternatively, the Propagation-Separation Approach can be used for the estimation of the biased parameter ϑ, replacing the adaptive estimator $\tilde{\theta}_i^{(k)}$ in Equation (2.7) (page 19) by $\tilde{\vartheta}_i^{(k)} := t^{-1}(\tilde{\theta}_i^{(k)})$ for all $i \in \{1,...,n\}$ and $k \in \{0,...,k^*\}$. If $t(\vartheta)$ is linear in ϑ, then it follows from Corollary 4.16 that all results in the Chapters 3 and 4 remain valid under Assumption A4. In contrast, for non-linear functions t, the extension of the results is not trivial. Many of our results relate to Theorems 2.13 and 2.24, which we recalled from Polzehl and Spokoiny [2006, Thm. 2.1 & 2.2]. The proof of the underlying Theorem 2.19 (page 28) is based on the identity $\overline{N}_i\overline{\theta}_i = S_i$. Therefore, we lose the exponential bounds in Theorems 2.13 and 2.24 and in all related results if t is not linear.

4.4.2 Violation of the exponential family model

The presented theoretical results require a local exponential family model. Nevertheless, the Propagation-Separation Approach can be applied on any local likelihood model, recall Notation 2.1. For this purpose, some crucial decisions have to be made.

First, we need to choose an appropriate estimator. Basically, the Propagation-Separation Approach can be based on any weighted pointwise estimator, replacing the non-adaptive weights by the corresponding iteratively updated structural adaptive weights. The precise choice of the estimator depends on the respective application. Under the local exponential family model we established, for the case of non-adaptive weights, the equivalence of the weighted MLE and the corresponding weighted mean, see Lemma 2.10. This equivalence may be violated in the case of an unrestricted local likelihood model. Here, the weighted MLE provides an estimate of the parameter θ, while the weighted mean refers to the expected value, where possibly $\mathbb{E}_\theta[T(Y)] \neq \theta$. Additionally, the MLE does not need to have a (unique) solution or an explicit representation. Then its computation may be impossible or computationally demanding. Moreover, we should ensure that the estimator of the Propagation-Separation Approach remains in the same distribution class as the observations. For instance, in § 6.2.3, we will use a weighted quadratic mean instead of a weighted arithmetic mean.

Next we need a distance function on the observation space that can be used for the statistical penalty, which determines the adaptive weights. Under the local exponential family model, we use the Kullback-Leibler divergence. This relates, under Assumption A1 (page 10), to the fitted log-likelihood. Additionally, in § 3.1.2, it provided the invariance of the homogeneous propagation condition with respect to the fixed parameter $\theta \in \Theta$ for several probability distributions. For other distribution classes, the application of the Kullback-Leibler divergence requires an explicit representation or an appropriate approximation. Alternatively, one could use another f-divergence or even a possibly asymmetric distance function on the observation space. Here, a detailed evaluation of the impact of the Kullback-Leibler divergence and of possible consequences of a replacement would be interesting. The basic properties of the algorithm can be extended to any distance which satisfies a generalized triangle inequality as established for the Kullback-Leibler divergence in Lemma 2.8. Only the exponential bounds in Theorems 2.13 and 2.24 are restricted to the local exponential family model. In Section 4.2, these were used to bound the probabilities

$$\mathbb{P}\left(\overline{N}_i^{(0)} \mathcal{KL}\left(\overline{\theta}_i^{(0)}, \theta\right) > z\right) \quad \text{and} \quad \mathbb{P}\left(\overline{N}_i^{(0)} \mathcal{KL}\left(\overline{\theta}_i^{(0)}, \mathbb{E}\overline{\theta}_i^{(0)}\right) > \left[\sqrt{z}/\varkappa - \varphi\right]^2\right)$$

in the proofs of Propositions 4.4 and 4.11.

Finally, the choice of the adaptation bandwidth needs to be revisited in the case of a violated Assumption A1. Our new formulation of the propagation condition in Chapter 3 is motivated by Theorems 2.13 and 2.24. Nevertheless, it can be applied whenever the local likelihood model is satisfied. If no choice of λ leads to the desired monotonicity of the corresponding function $\mathfrak{Z}_\lambda$, the propagation condition can be weakened. For instance, the adaptation bandwidth can be chosen as the smallest value of λ which ensures similar isolines as the non-adaptive estimator for all $p \in (\epsilon, 1)$, where $\epsilon > 0$ denotes the corresponding *weak propagation level*. This relates to the basic idea of the propagation condition, to ensure under homogeneity that the adaptive estimator behaves similar to the non-adaptive one. The alternative requirement that $\mathfrak{Z}_\lambda(0, p, \theta, i) \approx \mathfrak{Z}_\lambda(k^*, p, \theta, i)$ holds for all $p \in (\epsilon, 1)$ and every $i \in \{1, ..., n\}$ ensures a certain stability of estimates. For practical purposes, this condition is usually appropriate. Anyway, a careful evaluation of the variability with respect to the parameter $\theta \in \Theta$ is recommendable.

4.5 Proofs

Proof of Lemma 4.5. Recall the notations in § 3.1.3, and note that the adaptive estimators $\{\tilde{\theta}_i^{(k)}\}_{i=1}^{n_0}$ are not identically distributed. For every $j \in \{1, ..., n_0\}$, Proposition 4.4 yields

$$\begin{aligned}
& \left| \mathbb{E}\left[\hat{p}_\lambda^{(k)}(z)\right] - \mathbb{P}\left(\overline{N}_j^{(k)} \mathcal{KL}(\tilde{\theta}_j^{(k)}(\lambda), \theta) > z\right) \right| \\
\leq \; & n_0^{-1} \sum_{i=1}^{n_0} \left| \mathbb{E}\left[\mathbf{1}_{M_{(\lambda)}^{(k)}(z)}(X_i)\right] - \mathbb{P}\left(\overline{N}_j^{(k)} \mathcal{KL}(\tilde{\theta}_j^{(k)}(\lambda), \theta) > z\right) \right| \\
\leq \; & \max_{i \in \{1,...,n_0\}} \left\{ \left| \mathbb{P}\left(\overline{N}_i^{(k)} \mathcal{KL}(\tilde{\theta}_i^{(k)}(\lambda), \theta) > z\right) - \mathbb{P}\left(\overline{N}_j^{(k)} \mathcal{KL}(\tilde{\theta}_j^{(k)}(\lambda), \theta) > z\right) \right| \right\} \\
\leq \; & \max_{i \in \{1,...,n_0\}} \mathbb{P}\left(\overline{N}_i^{(k)} \mathcal{KL}(\tilde{\theta}_i^{(k)}(\lambda), \theta) > z\right) \\
\leq \; & \max\{2e^{-z}, \epsilon\}.
\end{aligned}$$

Furthermore, the variance of any random variable can be described as an $\mathbb{L}_2$-norm. The corresponding triangle inequality leads to

$$\begin{aligned}
\mathrm{Var}\left[\hat{p}_\lambda^{(k)}(z)\right] &= \left\| n_0^{-1} \sum_{i=1}^{n_0} \left(\mathbf{1}_{M_\lambda^{(k)}(z)}(X_i) - \mathbb{E}\left[\mathbf{1}_{M_\lambda^{(k)}(z)}(X_i)\right] \right) \right\|_{\mathbb{L}_2}^2 \\
&\leq \left(n_0^{-1} \sum_{i=1}^{n_0} \left\| \mathbf{1}_{M_\lambda^{(k)}(z)}(X_i) - \mathbb{E}\left[\mathbf{1}_{M_\lambda^{(k)}(z)}(X_i)\right] \right\|_{\mathbb{L}_2} \right)^2 \\
&\leq \max_{i \in \{1,...,n_0\}} \mathrm{Var}\left[\mathbf{1}_{M_\lambda^{(k)}(z)}(X_i)\right].
\end{aligned}$$

For any random variable X with values in $[0,1]$, we have $\mathrm{Var}[X] \leq \mathbb{E}[X]$. We get by Proposition 4.4 and the definition of $M_\lambda^{(k)}(z)$ that

$$\max_{i \in \{1,...,n_0\}} \mathbb{E}\left[\mathbf{1}_{M_\lambda^{(k)}(z)}(X_i)\right] = \max_{i \in \{1,...,n_0\}} \mathbb{P}\left(\overline{N}_i^{(k)} \mathcal{KL}(\tilde{\theta}_i^{(k)}(\lambda), \theta) > z\right) \leq \max\{2e^{-z}, \epsilon\},$$

which leads to Equation (4.3). □

Proof of Theorem 4.9. Recall the event $\Omega_\varkappa$ in Lemma 2.23. The adaptive estimator is defined as a weighted mean of the observations. Therefore, for all $k \in \{0, ..., k^*\}$, we get

$$\Omega_\varkappa \subseteq \bigcap_{i=1}^{n} \left\{ \tilde{\theta}_i^{(k)} \in \Theta_\varkappa \right\}.$$

Now let M^c denote the complement of the set M. We construct a disjoint union

$$\left[\mathcal{B}^{(k)}(z)\right]^c = \bigcup_{k'=0}^{k} \left(\left[\mathcal{B}^{(k')}(z)\right]^c \cap \left[\bigcap_{k''=0}^{k'-1} \mathcal{B}^{(k'')}(z) \right] \right),$$

where we set $\bigcap_{k''=0}^{k'-1} \mathcal{B}^{(k'')}(z) := \Omega$ if $k' = 0$. Then, we get

$$\begin{aligned}
\mathbb{P}\left(\mathcal{B}^{(k)}(z)\right) &\geq 1 - \mathbb{P}\left(\Omega_\varkappa^c\right) - \mathbb{P}\left(\Omega_\varkappa \cap \bigcup_{k'=0}^{k} \left[(\mathcal{B}^{(k')}(z))^c \cap \left(\bigcap_{k''=0}^{k'-1} \mathcal{B}^{(k'')}(z) \right) \right] \right) \\
&\geq 1 - p_\varkappa - \sum_{k'=0}^{k} \mathbb{P}\left(\Omega_\varkappa \cap \left[\mathcal{B}^{(k')}(z)\right]^c \cap \left[\bigcap_{k''=0}^{k'-1} \mathcal{B}^{(k'')}(z) \right] \right)
\end{aligned}$$

and analogously for the conditional probability

$$\begin{aligned} & \mathbb{P}\left(\mathcal{B}^{(k)}(z)|M^{(k)}(z)\right) \qquad (4.11)\\ \geq\ & 1-\left[p_{\varkappa}+\sum_{k'=0}^{k}\mathbb{P}\left(M^{(k')}(z)\cap\Omega_{\varkappa}\cap\left[\mathcal{B}^{(k')}(z)\right]^{c}\cap\left[\bigcap_{k''=0}^{k'-1}\mathcal{B}^{(k'')}(z)\right]\right)\right]/\mathbb{P}\left(M^{(k)}(z)\right), \end{aligned}$$

where we used that $M^{(k)}(z)\subseteq M^{(k')}(z)$ for $k'\leq k$. The choice of $h^{(0)}$ ensures, for every $i\in\{1,...,n\}$, that $U_i^{(0)}\setminus\mathcal{V}_i=\emptyset$. Moreover, it holds $\tilde{\theta}_i^{(0)}=\overline{\theta}_i^{(0)}$ see Algorithm 2 (2) (page 18), and it follows from Theorem 2.13 that

$$\begin{aligned} \mathbb{P}\left(M^{(0)}(z)\cap\Omega_{\varkappa}\cap\left[\mathcal{B}^{(0)}(z)\right]^{c}\right) \quad & \overset{\overline{n}_i^{(0)}=\overline{N}_i^{(0)}}{\leq} \quad n\cdot\mathbb{P}\left(\overline{N}_i^{(0)}\mathcal{KL}(\overline{\theta}_i^{(0)},\theta_i)>z\right)\\ & \leq \quad 2ne^{-z}. \qquad (4.12) \end{aligned}$$

By definition of the events $\mathcal{B}^{(k)}(z)$, $M^{(k)}(z)$, and $\Omega_{\varkappa}$, the conditions of Proposition 4.1 are satisfied on the intersection

$$M^{(k')}(z)\cap\Omega_{\varkappa}\cap\left[\bigcap_{k''=0}^{k'-1}\mathcal{B}^{(k'')}(z)\right]$$

for all $k'\in\{1,...,k\}$. There, it follows that $\tilde{w}_{ij}^{(k')}=0$ for all $X_j\notin U_i^{(k')}\cap\mathcal{V}_i$. Hence, smoothing is restricted to the homogeneous neighborhood $\mathcal{V}_i$, and $\mathbb{E}[T(Y_j)]=\theta_i$ for every X_j with $\tilde{w}_{ij}^{(k')}>0$. Then, we get by Proposition 4.4 that

$$\mathbb{P}\left(\{\overline{n}_i^{(k')}\mathcal{KL}\left(\tilde{\theta}_i^{(k')},\theta_i\right)>z\}\cap M^{(k')}(z)\cap\Omega_{\varkappa}\cap\left[\bigcap_{k''=0}^{k'-1}\mathcal{B}^{(k'')}(z)\right]\right)\leq\max\left\{2e^{-z},\epsilon\right\} \quad (4.13)$$

for all $k'\in\{1,...,k\}$. Finally, Equations (4.11), (4.12), and (4.13) lead to

$$\mathbb{P}\left(\mathcal{B}^{(k)}(z)|M^{(k)}(z)\right)\geq 1-\left[p_{\varkappa}+(k+1)\max\left\{2ne^{-z},n\epsilon\right\}\right]/\mathbb{P}\left(M^{(k)}(z)\right).$$

This terminates the proof. □

Proof of Proposition 4.11. The inhomogeneous propagation condition yields the monotonicity of the function $\hat{\mathfrak{Z}}_{\lambda}\left(k,p;\{\theta_j\}_{j=1}^{n},i\right)$ in $k\leq k_0$ for all $p\in(\epsilon,1)$ and every $i\in\{1,...,n\}$. This implies Equation (4.8). We turn to Equation (4.7), and we consider the event $\breve{\Omega}_{\varkappa}$ in Lemma 2.23, which satisfies

$$\breve{\Omega}_{\varkappa}\subseteq\bigcap_{i=1}^{n}\left\{\tilde{\theta}_i^{(k)}\in\Theta_{\varkappa}\right\}$$

for the same reasons as in the proof of Theorem 4.9. Then, we use the convexity of the Kullback-Leibler divergence with respect to the first argument, see Lemma 2.5. Denoting the complement

of the set M by M^c, it follows from the initialization of Algorithm 2 (page 18) with $\tilde{\theta}_i^{(0)} = \overline{\theta}_i^{(0)}$ that

$$
\begin{aligned}
& \mathbb{P}\left(\overline{N}_i^{(k)} \mathcal{KL}\left(\tilde{\theta}_i^{(k)}, \theta_i\right) > z\right) \\
\overset{\text{Lem. 2.8}}{\leq} \; & \mathbb{P}\left(\left\{\varkappa^2 \overline{N}_i^{(k)} \left[\mathcal{KL}^{1/2}\left(\tilde{\theta}_i^{(k)}, \mathcal{E}\tilde{\theta}_i^{(k)}\right) + \mathcal{KL}^{1/2}\left(\mathcal{E}\tilde{\theta}_i^{(k)}, \theta_i\right)\right]^2 > z\right\} \cap \breve{\Omega}_\varkappa\right) + \mathbb{P}\left(\breve{\Omega}_\varkappa^c\right) \\
\overset{\text{Lem. 2.5}}{\leq} \; & \mathbb{P}\left(\left\{\overline{N}_i^{(k)} \mathcal{KL}\left(\tilde{\theta}_i^{(k)}, \mathcal{E}\tilde{\theta}_i^{(k)}\right) > \left[\sqrt{z}/\varkappa - \varphi\right]^2\right\} \cap \breve{\Omega}_\varkappa\right) + \breve{p}_{\varkappa,0} \\
\overset{\text{Eq. (4.8)}}{\leq} \; & \max\left\{\mathbb{P}\left(\left\{\overline{N}_i^{(0)} \mathcal{KL}\left(\overline{\theta}_i^{(0)}, \mathbb{E}\overline{\theta}_i^{(0)}\right) > \left[\sqrt{z}/\varkappa - \varphi\right]^2\right\} \cap \breve{\Omega}_\varkappa\right), \epsilon\right\} + \breve{p}_{\varkappa,0} \\
\overset{\text{Thm. 2.24}}{\leq} \; & \max\left\{2e^{-\left[\sqrt{z}/\varkappa - \varphi\right]^2/\varkappa^2}, \epsilon\right\} + \breve{p}_{\varkappa,0}
\end{aligned}
$$

since the event $\breve{\Omega}_\varkappa$ is independent of the iteration step k. □

Chapter 5

Simulations and discussion

In the last chapter, we established several theoretical properties of the simplified Propagation-Separation Approach in Algorithm 2 (page 18). Here, we will illustrate these properties by simulated examples with Gaussian and exponentially distributed observations. In particular, we will compare the results of the simplified algorithm with the original procedure in Algorithm 1 (page 16). Furthermore, we will simulate non-central chi-distributed observations, which violate the local exponential family model in Assumption A1. Nevertheless, the algorithm yields, for the considered examples, the same heuristic behavior as for exponential families. We will close with a discussion of our theoretical and numerical results on the simplified Propagation-Separation Approach. In Chapter 6, we will apply the simplified algorithm on experimental magnetic resonance data.

5.1 Test functions

In this section, we will present all test functions that we will consider in the following numerical study. The applied parameter choices will be specified together with the numerical results in Section 5.3. Usually, we simulated data with $n = 1000$ observations. In some examples, we changed the sample size $n \in \mathbb{N}$. Then, we increased the cardinality of each region of the introduced parameter functions by the same factor such that the design portions remain unchanged.

First, we will demonstrate the separation and the propagation property. For this purpose, we will consider a shifted and scaled indicator function, where we will vary, for Gaussian and exponentially distributed observations, the step size, the variance, the sample size, and the choice of the adaptation bandwidth λ. This piecewise constant setting coincides with the setting which the original Propagation-Separation Approach in Algorithm 1 and its simplified version in Algorithm 2 assume. Therefore, we expect the procedure to provide propagation, separation, and a certain stability of estimates, see Sections 4.1 and 4.2. Let the sample size be even, $n \in 2\mathbb{N}$. Then, we split the design into two parts with equal cardinality, $\mathcal{X}_1 := \{X_i\}_{i=1}^{n/2}$ and $\mathcal{X}_2 := \{X_i\}_{i=n/2+1}^{n}$. We will consider the test function

$$\theta(x) := a_1 + (a_2 - a_1) \cdot \mathbf{1}_{\mathcal{X}_2}(x), \qquad x \in \mathcal{X}, \tag{5.1}$$

where $\mathbf{1}$ denotes the indicator function, and $a_1, a_2 \in \mathbb{R}$ are constants.

Next we will illustrate the formation of the associated step function in Definition 4.14. Here, we applied three test functions where the structural Assumptions A2 and A3 are violated. First, we reused the piecewise smooth function (2.9) (page 19). The corresponding smoothing results at three different location bandwidths are shown in Figure 2.2 (page 20). Second, we will consider a piecewise constant function with small discontinuities and three different regions of monotonicity,

$$\begin{array}{llll}
\theta(x) := 0, & x \in \{1,...,50\}, & \theta(x) := 2.2, & x \in \{451,...,500\}, \\
\theta(x) := 0.5, & x \in \{51,...,100\}, & \theta(x) := 1.7, & x \in \{501,...,550\}, \\
\theta(x) := 1, & x \in \{101,...,150\}, & \theta(x) := 1.2, & x \in \{551,...,600\}, \\
\theta(x) := 1.5, & x \in \{151,...,200\}, & \theta(x) := 0.7, & x \in \{601,...,650\}, \\
\theta(x) := 2, & x \in \{201,...,250\}, & \theta(x) := 0.9, & x \in \{651,...,750\}, \\
\theta(x) := 2.5, & x \in \{251,...,300\}, & \theta(x) := 1.6, & x \in \{751,...,800\}, \\
\theta(x) := 3, & x \in \{301,...,350\}, & \theta(x) := 2.6, & x \in \{801,...,900\}, \\
\theta(x) := 3.2, & x \in \{351,...,400\}, & \theta(x) := 2.9, & x \in \{901,...,1000\}. \\
\theta(x) := 2.7, & x \in \{401,...,450\}, & &
\end{array} \tag{5.2}$$

This function is constructed to especially illustrate the consequences of close steps in distant locations. Third, we will study the behavior of the simplified Propagation-Separation Approach for the logarithmic function

$$\theta(x) := \log(x), \qquad x \in \mathcal{X}. \tag{5.3}$$

Here, the parameter values change slowly. Polzehl and Spokoiny [2006] introduced the memory step in order to avoid the increase of the estimation bias for large location bandwidths, which occurs in this case.

In § 5.3.3, we will concentrate on the impact of the memory step. For this purpose, we will compare the results of the original and the simplified Propagation-Separation Approach applied to the above mentioned test functions. In particular, we will vary the memory bandwidth in order to illustrate the effects of an increasing amount of aggregation. Finally, we will show some results of the simplified procedure for non-central chi-distributed observations. Here, we will reuse the piecewise constant and the piecewise smooth functions in § 2.2.4, Equations (2.8) and (2.9).

5.2 Methods

We used the implementation of the Propagation-Separation Approach in the **R**-package **aws** by Polzehl [2012]. Here, the memory step is omitted by default. If desired, it can be included in the procedure, setting `memory=TRUE`. For the sake of simplicity, we will only show univariate examples where $\mathcal{X} \subseteq \mathbb{R}$. However, the **R**-package **aws** as well allows higher dimensions, where $\mathcal{X} \subseteq \mathbb{R}^d$ with $d = 1,2,3$.

Let `n` denote the sample size. For `t` $:= \theta \in \mathbb{R}^{\mathtt{n}}$, we generated the random observations via the commands

```
tnoise <- t + rnorm(n, sd=1)          if P := {N(θ,1)}_{θ∈Θ},
tnoise <- rexp(n,1/t)                 if P := {Exp(1/θ)}_{θ∈Θ},
tnoise <- sqrt(rchisq(n, df, ncp=t^2))  if P := {χ_df(θ)}_{θ∈Θ},
```

where `df` is the assumed degree of freedom of the non-central chi-distribution, here $\texttt{df} = 4$. For the non-central chi-distribution, the expectation of the observations does not equal the non-centrality parameter. Using the Equation (6.7), we calculate the associated expected values via

```
et  <-  sqrt(pi/2)*hyperg_1F1(-0.5, df/2, -t^2/2)*gamma(df/2+.5)
        /gamma(1.5)/gamma(df/2),
```

which requires the **R**-package **gsl**.

We applied the **R**-function `aws`, setting

```
that  <-  aws(tnoise, u=t, hmax=10000, lkern="Triangle",
          homogen=FALSE, maxni=TRUE).
```

This function depends on the following arguments.

- The n-dimensional vector `tnoise` contains a realization of the random observations.
- The n-dimensional vector `t` contains the corresponding expected values, which are used for comparison with the estimate `that`. If $\mathcal{P} := \{\chi_{\mathrm{df}}(\theta)\}_{\theta\in\Theta}$, then we use the choice `u=et` in the function `aws` in order to compare the adaptive estimate `that` with the expectation `et` in place of the true parameter `t`.
- The argument `hmax` sets an upper bound for the location bandwidth.
- The choice `lkern="Triangle"` yields the location kernel in Equation (2.10) (page 19).
- The option `homogen=TRUE` leads to a modification of the algorithm, which reduces the computation time. We set `homogen=FALSE` in order to avoid artifacts which result from these modification, but not from the Propagation-Separation Algorithm itself.
- Via `maxni=TRUE`, we use a modified statistical penalty, setting

  $$s_{ij}^{(k)} := \max_{k'\le k} \hat{N}_i^{(k'-1)} \mathcal{KL}(\hat{\theta}_i^{(k-1)}, \hat{\theta}_j^{(k-1)})$$

  in Algorithm 1 (page 16) and analogously in Algorithm 2 (page 18). This modification preserves an achieved quality of estimation during iteration. It will be further discussed in § 5.4.1.

For some examples, we used the following additional arguments for further adjustment.

- The argument `family` determines the underlying family of probability distributions $\mathcal{P}$. Here, we used `family="Gaussian"`, `family="exponential"`, and `family="NCchi"`.
- The parameter `ladjust` scales the default choice of the adaptation bandwidth λ and hence allows for an increase or reduction of the amount of adaptation.
- For Gaussian distributed observations, the assumed standard deviation can be specified via `sigma2`$=\sigma^2$, which influences the Kullback-Leibler divergence in the statistical penalty. Else, the algorithm estimates the variance from the observations `tnoise`.
- In § 5.3.3 and § 5.3.4, we will include the memory step via the argument `memory=TRUE`. This is implemented in the package **aws** for two different memory kernels, which can be specified by `aggkern="Triangle"` or `aggkern="Uniform"`.

For the other arguments of the function `aws`, we used the default values. The corresponding adaptation bandwidths λ are in accordance with the propagation condition of the propagation levels ϵ_λ given in Table 5.1.

By means of an additionally included function `awsweights`, we visualized the weighting schemes of the resulting non-adaptive weights $\{\overline{w}_{ij}^{(k)}\}_{i,j}$, the adaptive weights $\{\tilde{w}_{ij}^{(k)}\}_{i,j}$, and

$\mathcal{P}$	$\{\mathcal{N}(\theta,1)\}_{\theta\in\Theta}$	$\{\mathrm{Exp}(1/\theta)\}_{\theta\in\Theta}$	$\{\chi_4(\theta)\}_{\theta\in\Theta}$
λ	14.6	12.8	30.2
ϵ_λ	0.001	$5\cdot 10^{-4}$	0.001

Table 5.1: Propagation levels of the default adaptation bandwidths in the **R**-package **aws**.

the adaptation kernel $\{K_{\mathrm{ad}}(s_{ij}^{(k)}/\lambda)\}_{i,j}$, which equals the quotient $\tilde{w}_{ij}^{(k)}/\overline{w}_{ij}^{(k)} = K_{\mathrm{ad}}(s_{ij}^{(k)}/\lambda)$ if $\overline{w}_{ij}^{(k)} > 0$. The values of these quantities are shown in grey scales, where zero corresponds to black and one to white, respectively. Moreover, we included a scaling factor `tadjust`, which allows a manipulation of the memory bandwidth τ, see Algorithm 1 (page 16) for the application of τ and § 2.2.2 for its choice. In the package **aws**, this is given as

$$\tau^{(k)} := (2 * \tau_1 + \tau_1 * \max\{k_{\text{star}} - \log(h^{(k)}), 0\}),$$

where the constant k_{star} depends on the family of probability distributions $\mathcal{P}$, and

$$\tau_1 := \begin{cases} \texttt{ladjust*tadjust*20} & \text{if } \texttt{aggkern="Triangle"}, \\ \texttt{ladjust*tadjust*8} & \text{if } \texttt{aggkern="Uniform"} \end{cases}$$

with `ladjust=1` and `tadjust=1` by default.

5.3 Numerical results

Here, we will list our parameter choices, and we will show the results of our numerical study in order to illustrate the separation and propagation property, the formation of the associated step function, the impact of the memory step, and the stability of estimates. More precisely, we will present several example plots for the realization `seed=1`, the corresponding weighting schemes, and boxplots of the mean absolute error (MAE) over 1000 realizations, where `seed=l` and $l \in \{1,...,1000\}$. This set of realizations was chosen in order to ensure the reproducibility of our results. The MAE is automatically provided by the function `aws`.

5.3.1 Separation and propagation

In Figure 5.1, we show four typical results, to which the Propagation-Separation Approach may lead on a locally constant parameter function. From left to right we observe adaptation to outliers (a), an (almost) perfect separation of the two homogeneity regions (b), a shifted step (c), and over-smoothing (d), where the two different regions are treated as one. Results are shown for Gaussian observations of the test function (5.1) with $a_1 := 1$, and $a_2 := 2$, using increasing values of the adaptation bandwidth $\lambda > 0$ by `ladjust=0.25,1,7,20`. For a sufficiently large adaptation bandwidth λ, the first case can be avoided with high probability. Then, the algorithm yields similar plots as in Figure 5.1 (b-d) for decreasing step sizes (not shown).

The boxplots in Figure 5.2 support this heuristic observation. In the top left, we show the change of the MAE for an increasing adaptation bandwidth, where we set `ladjust=0.1, 0.25,0.5,1,4,6,10,20`, considering the same test function as in Figure 5.1. At the boundaries, the MAE increases considerably due to under-smoothing for small values of `ladjust` and due to over-smoothing for large values. Similarly, we observe an increasing MAE in the top right

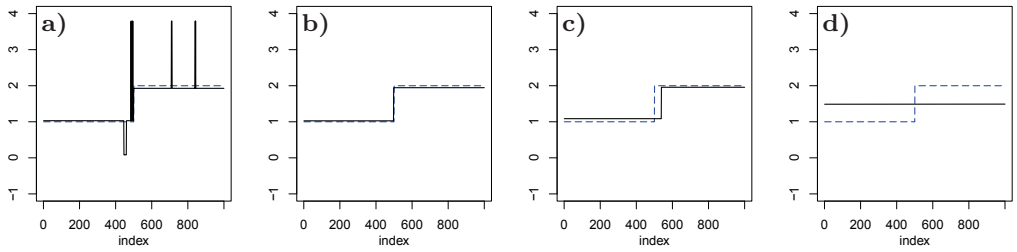

Figure 5.1: The separation property on Gaussian observations for increasing adaptation bandwidths (from left to right), where the blue dashed line represents the true parameter function, and the adaptive estimate is shown as black solid line.

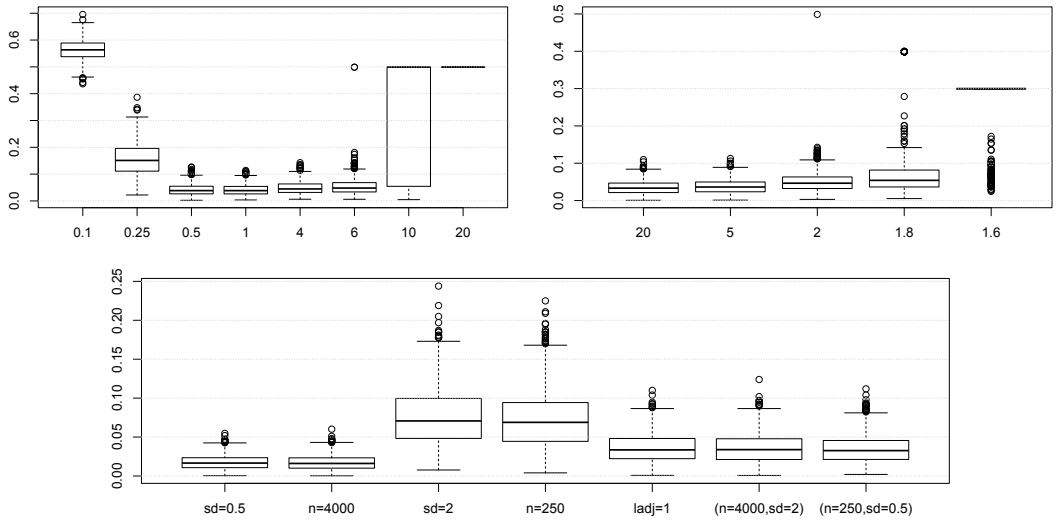

Figure 5.2: MAE-boxplots concerning the separation property on Gaussian observations for increasing adaptation bandwidths (upper left), decreasing step sizes (upper right), and varying sample sizes and noise standard deviations (bottom).

panel of Figure 5.2, which we computed with `ladjust=5`, $a_1 := 1$, and $a_2 := 20, 5, 2, 1.8, 1.6$, where small steps may assimilate. Finally, for $a_2 := 5$ and `ladjust=1`, we illustrate the impact of the sample size and the noise standard deviation, where we chose $n = 250, 1000, 4000$ and $\sigma = 0.5, 1, 2$. In the bottom panel of Figure 5.2, we only assign the choices $n \neq 1000$ and $\sigma \neq 1$. As suggested by Example 4.3, a smaller standard deviation σ as well as an increased sample size n yield improved results, while an increased standard deviation or a reduced sample size lead (on average) to a larger MAE. To some extent, these effects can be compensated by a simultaneous change of σ^2 and n, using the same scaling factor for both.

We produced similar plots for exponentially distributed observations, using different parameter choices. For the boxplot in the top left panel of Figure 5.3, we set $a_1 := 1$, $a_2 := 1/10$, and `ladjust=0.2,0.25,0.5,1,4,6,10,20`. For the top right panel, we used `ladjust=1`, $a_1 := 1$, and $a_2 := 1/20, 1/10, 1/2, 1/1.6, 1/1.4$. The sample sizes were chosen as `n=2000` (top) and `n=250,1000,4000` (bottom). The standard deviation depends on the parameter values θ, which we adjusted by the scaling factors 0.5 and 2. This choice led to the pairs $(a_1, a_2) \in \{(1/2, 1/10), (1, 1/5), (2, 2/5)\}$. We observe that the effect of the standard deviation

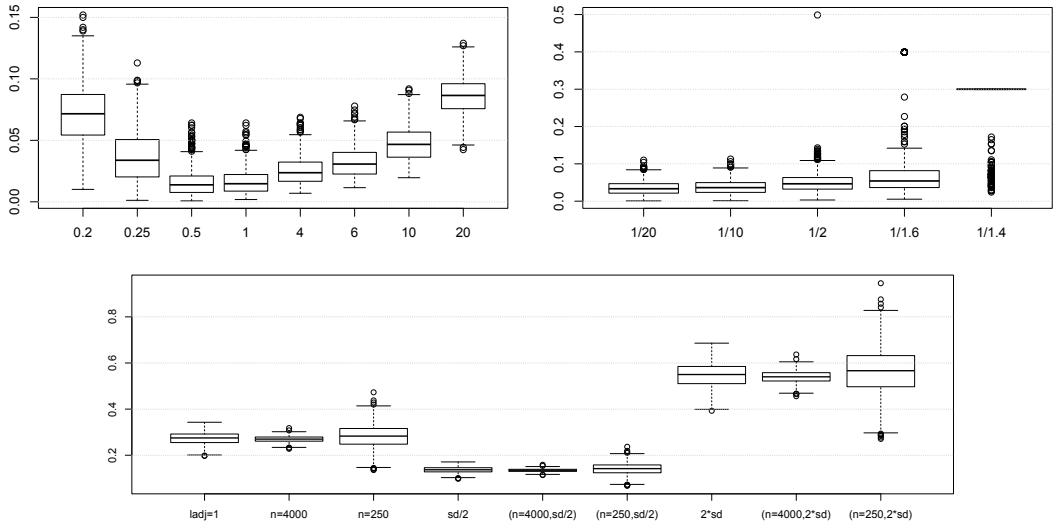

Figure 5.3: MAE-boxplots concerning the separation property on exponentially distributed observations for increasing adaptation bandwidths (top left), decreasing step sizes (top right), and varying sample sizes and noise standard deviations (bottom).

is similar as for Gaussian observations, while the sample size has minor impact in this example. The variance of the MAE decreases with increasing sample sizes, while its median remains almost unchanged. Probably, much larger sample sizes would show stronger effects, but this would lead to a considerable increase of the computation time.

5.3.2 Formation of the associated step function

Next we study the formation of the associated step function of the Propagation-Separation Approach, which we introduced in Section 4.3. For this purpose, we visualize the resulting weighting schemes.

In Figure 5.4, we consider the piecewise smooth function (2.9) (page 19) with Gaussian observations. In the first row, we provide the weighting schemes of the iteration step where the MAE is minimized. The product of the adaptive term $\{K_{\mathrm{ad}}(s_{ij}^{(k)}/\lambda)\}_{i,j}$ (a) and the non-adaptive weights $\{\overline{w}_{ij}^{(k)}\}_{i,j}$ (b) results in the adaptive weights $\{\tilde{w}_{ij}^{(k)}\}_{i,j}$ (c). This illustrates the interaction of adaptation and location. For `hmax=2000`, the algorithm results in the associated step function (d). Here, the adaptive weights (f) and the weighting scheme of the corresponding adaptive term $\{K_{\mathrm{ad}}(s_{ij}^{(k)}/\lambda)\}_{i,j}$ (not shown) were visually indistinguishable due to the large size of the considered local neighborhood, which is determined by the non-adaptive weighting scheme (e).

We present in Figure 5.5 the example plots and corresponding weighting schemes $\{\tilde{w}_{ij}^{(k)}\}_{i,j}$ of the step function (5.2) (page 64). The discontinuities are too small for separation. Therefore, the algorithm forms a step function which differs from the original parameter function. The minimal MAE is provided for `hmax=30`, where the considered local neighborhood is small and separation does not yet occur (a+e). For `hmax=120`, we observe, in the example plot (b) as well as in the weighting scheme (f), that the estimation function starts to form a step function. In (c), the estimation function of `hmax=600` resembles a step function, but the weighting scheme (g) already indicates that the formed steps may change with increasing location bandwidths. Indeed,

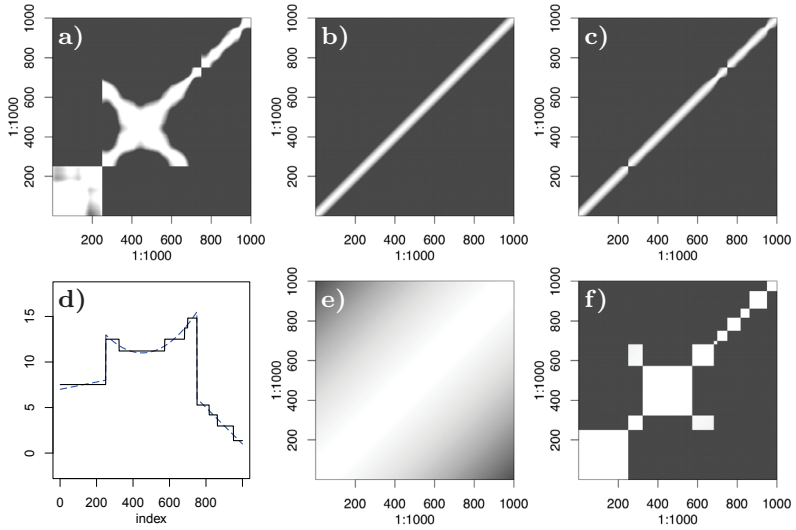

Figure 5.4: Formation of the associated step function for the piecewise smooth function (2.9) with Gaussian observations.

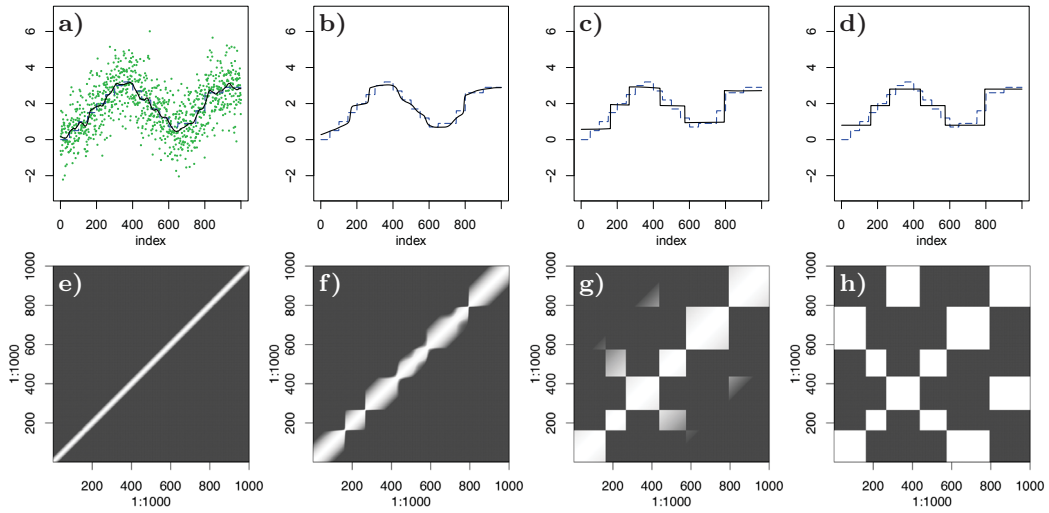

Figure 5.5: Formation of the associated step function for the step function (5.2) with Gaussian observations.

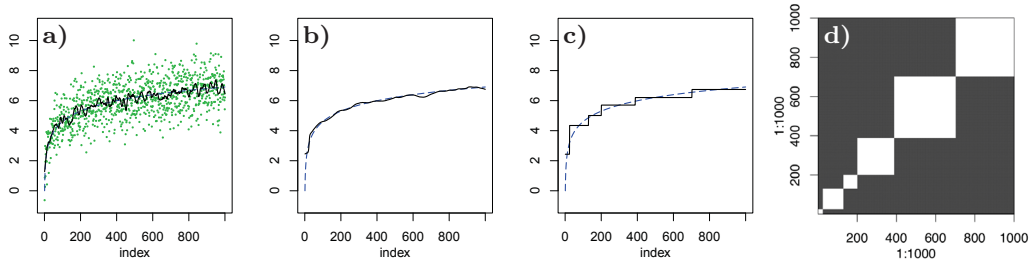

Figure 5.6: Formation of the associated step function for the logarithmic function (5.3) with Gaussian observations.

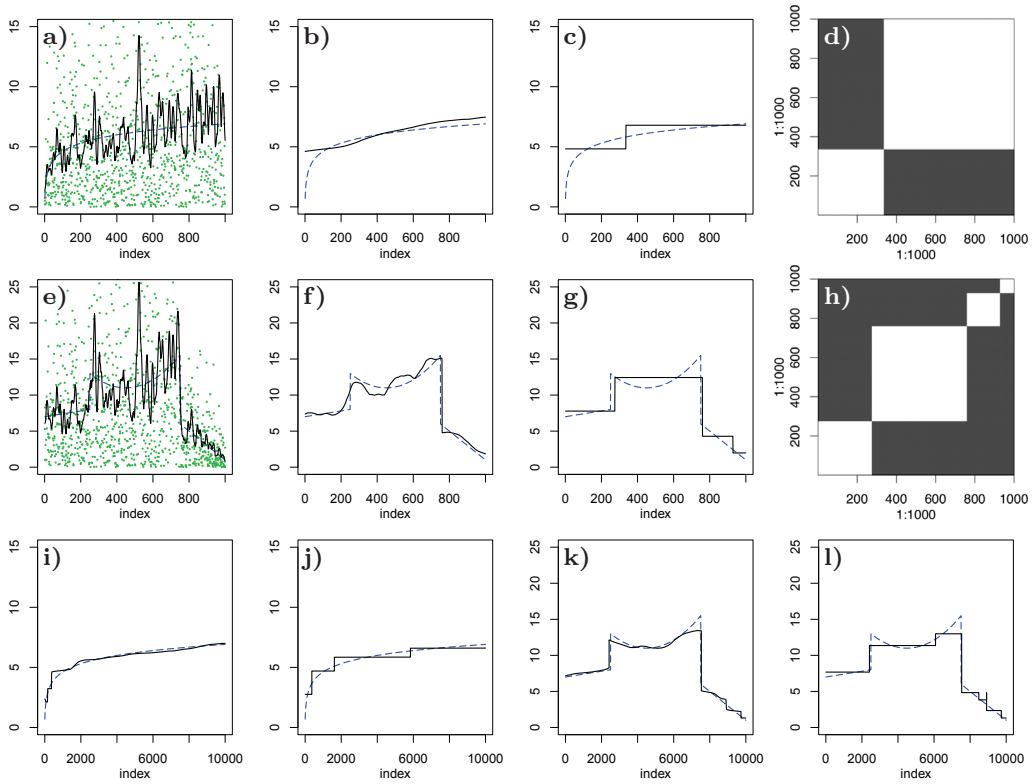

Figure 5.7: Formation of the associated step function for the piecewise smooth function (2.9) and the logarithmic function (5.3) with exponentially distributed observations, where `n=1000` (top) and `n=10000` (last row).

in (d), several steps in different locations have been assimilated as the weighting scheme (h) points out. For the plots in the last column, we set `hmax=20000`.

Even for the logarithmic function (5.3), the simplified algorithm results in a step function with disjoint regions (d). In Figure 5.6, we show the example plots for a small location bandwidth `hmax=10` (a), at `hmax=60` (b), where the MAE is minimal, and at `hmax=2000` (c).

Additionally, we studied the formation of the associated step function for exponentially distributed observations on several test functions. In Figure 5.7, we provide the results for the parameter functions in Equations (5.3) (first row) and (2.9) (second row). Here again, for sufficiently large location bandwidths, the algorithm results in the associated step function with disjoint regions of non-zero adaptive weights (d+h). We show the example plots for a small location bandwidth (a+e), where `hmax=10`, an intermediate iteration step with minimized MAE (b+f), and a large location bandwidth `hmax=20000` (c+g). As the resulting plots look considerably worse than in the case of Gaussian observations, we repeated the computation on a larger sample size, setting `n=10000` instead of `n=1000`. In the last row, we show the results with minimal MAE (i+k) and for `hmax=20000` (j+l).

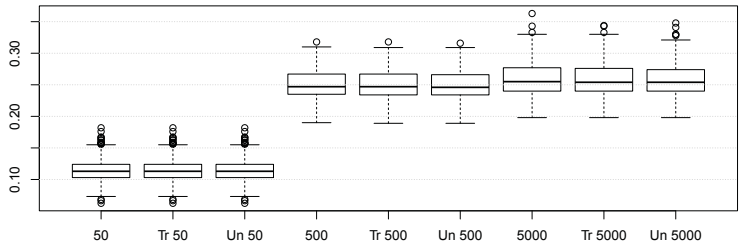

Figure 5.8: MAE-boxplots for `hmax=50,500,5000` with and without memory step, setting `aggkern="Triangle"` (Tr) and `aggkern="Uniform"` (Un).

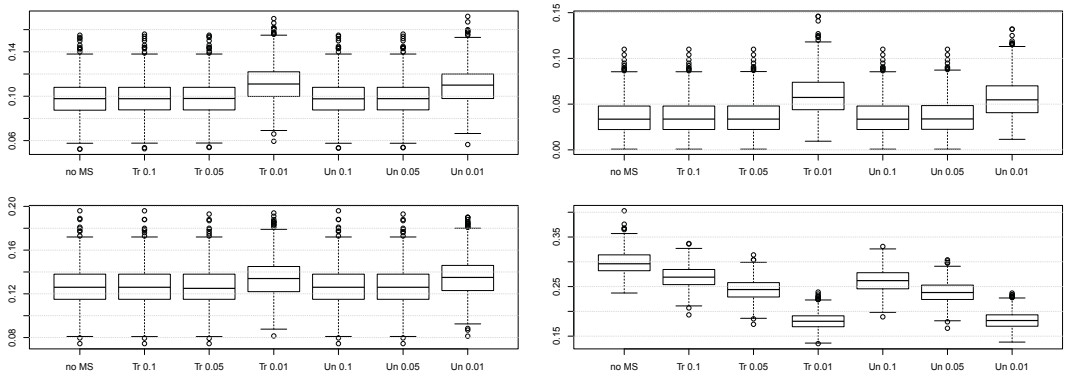

Figure 5.9: MAE-boxplots for `hmax=50` (left) and `hmax=10000` (right) for the indicator function (top) and for the piecewise smooth function (bottom) with Gaussian observations. We applied the algorithm without (no MS) and with memory step, setting `aggkern="Triangle"` (Tr) and `aggkern="Uniform"` (Un), where `tadjust=0.1,0.05,0.01` increases the amount of aggregation.

5.3.3 Impact of the memory step

In [Polzehl and Spokoiny, 2006, Thm. 5.7], the memory step provided a general result on the stability of estimates, up to some constant. However, its practical use is questionable. No situation has been reported to date where the memory step considerably improved the results of the Propagation-Separation Approach. Therefore, we aim for a better understanding of its impact on the resulting estimates. For this purpose, we compared the results of the original and the simplified algorithm on the test functions in Section 5.1 for Gaussian and exponentially distributed observations.

In Figure 5.8, we show the results for the piecewise smooth function (2.9) (page 19) with Gaussian distributed observations. We applied three location bandwidths, `hmax=50,500,5000`, each of them without memory step (`memory=FALSE`), with memory step, using a triangular kernel (`memory=TRUE, aggkern="Triangle"`), and with memory step, using a uniform kernel (`memory=TRUE, aggkern="Uniform"`). As for all other test functions with Gaussian or exponentially distributed observations, there is (almost) no difference between the resulting boxplots with and without memory step and for the two memory kernels. This raises the question whether the memory step itself does not have any effect, or whether the default parameter choices in the **R**-package **aws** are unfavorable.

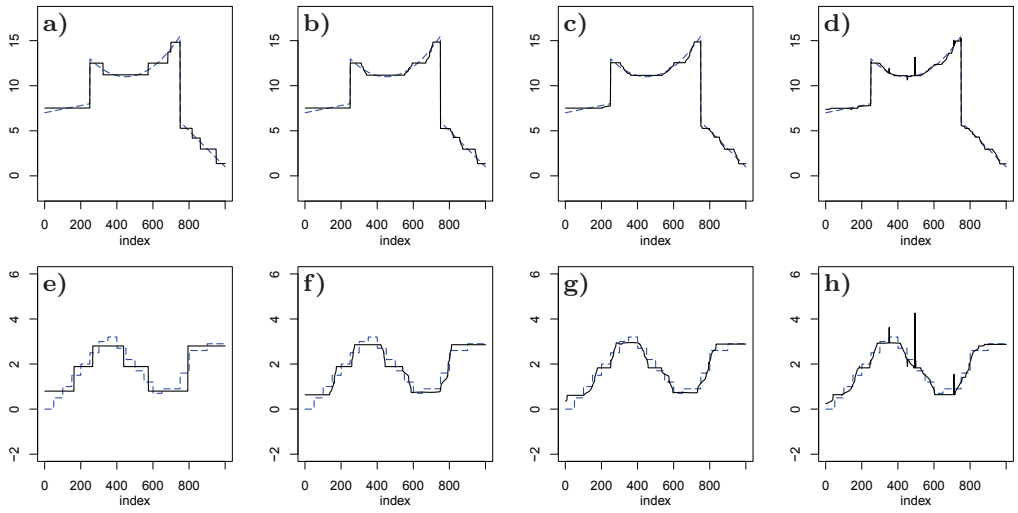

Figure 5.10: Example plots for the piecewise smooth function (top) and the step function (bottom) at `hmax=100000`. We applied the algorithm without (a+e) and with memory step, setting `aggkern="Triangle"` and (from left to right) `tadjust=0.05,0.02,0.01`.

In order to provide a deeper insight into the mode of action of the memory step, we increased the amount of aggregation by means of the additionally implemented scaling factor `tadjust` of the memory bandwidth $\tau > 0$. In Figure 5.9, we present, for the indicator function (5.1) with $a_1 := 1$ and $a_2 := 5$ (top) and for the piecewise smooth function (2.9) (bottom), the MAE-boxplots at some early iteration step with `hmax=50` (left) and for `hmax=10000` (right), assuming Gaussian observations. The amount of aggregation increases due to the choices `tadjust=0.1,0.05,0.01` with `aggkern="Triangle"` and `aggkern="Uniform"`. For comparison, we show the result of the simplified procedure as well, where `memory=FALSE`. For `hmax=50`, we observe an increase of the MAE for both test functions at `tadjust=0.01`, while the MAE without memory step coincides with the results for `tadjust=0.1,0.05`. For `hmax=10000`, this observation remains valid for the indicator function (top right), where the locally constant model of the Propagation-Separation Approach is satisfied.

In contrast, for the piecewise smooth function, we know from § 5.3.2 that the estimation function approaches the associated step function, which leads to an increase of the MAE. As demonstrated in the bottom right of Figure 5.9, the MAE decreases with increasing amount of aggregation, that is with decreasing `tadjust`. Unfortunately, this increases the risk of adaptation to noise as we illustrate on some example plots in Figure 5.10 for the piecewise smooth function (2.9) (top) and for the step function (5.2) (bottom), both with Gaussian observations. Without memory step as well as with memory step and `tadjust=1`, the algorithm results in the associated step function (a+e). To some extent, this effect can be attenuated by increasing the amount of aggregation, setting `tadjust=0.05` (b+f) or even `tadjust=0.02` (c+g). For `tadjust=0.01` (d+h), we observe adaptation to noise, which indicates the increased risk of adaptation to outliers due to the decreased memory bandwidth. Naturally, for other realizations, larger sample sizes or different test functions, this could happen for larger values of `tadjust` as well. We got similar results for the other test functions in Section 5.1 with Gaussian and as well with exponentially distributed observations (not shown).

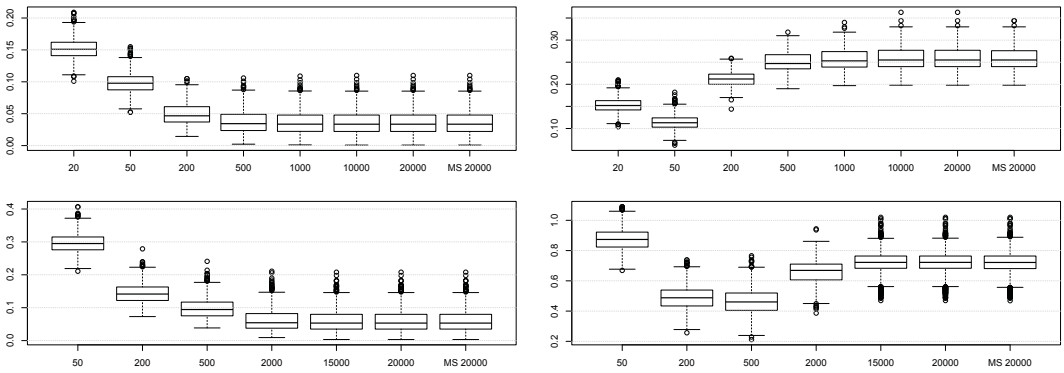

Figure 5.11: Stability of estimates for Gaussian (top) and exponentially (bottom) distributed observations for the indicator function (left) and for the piecewise smooth function (right).

5.3.4 Stability of estimates

The numerical results in § 5.3.2 suggest that the simplified Propagation Separation Approach provides a certain stability of estimates, where the associated step function acts as an intrinsic stopping criterion. In § 5.4.1, we will discuss the reasons which impede a theoretical proof of this heuristic property. Here, we present some boxplots which indicate the immutability of the MAE for sufficiently large location bandwidths.

We show results for Gaussian and exponentially distributed observations (Figure 5.11). We set `n=1000` and `hmax=20,50,200,500,1000,10000,20000` for the former, and `n=4000` and `hmax=50,200,500,2000,15000,20000` for the latter. Here again, we consider the indicator function (5.1) with $a_1 := 1$ and $a_2 := 5$, where the structural assumption of the Propagation-Separation Approach is satisfied. This leads to a decreasing MAE during iteration. As an example for the case of a misspecified model, we again apply the piecewise smooth function (2.9). Here, the MAE increases for larger location bandwidth as the estimator is forced into a step function. Nevertheless, for both test functions and both probability distributions, the MAE stabilizes for sufficiently large location bandwidths. For comparison, we show the MAE which results from the choices `memory=TRUE, aggkern="Triangle", tadjust=1`.

5.3.5 Non-central chi-distributed observations

Finally, we apply the simplified algorithm to non-central chi-distributed observations. This class of distributions violates Assumption A1, which is required in our theoretical results. Therefore, it is no longer ensured that the algorithm provides the same behavior as for Gaussian or exponentially distributed observations. We consider the same test functions as in § 2.2.4, that is the piecewise constant function (2.8) and the piecewise smooth function (2.9), where `n=2000`. In Figures 5.12 and 5.13, we observe the same behavior as before, namely the formation of the associated step function, a final weighting scheme that forms a partition of the design space, and the stability of the MAE for sufficiently large iteration steps.

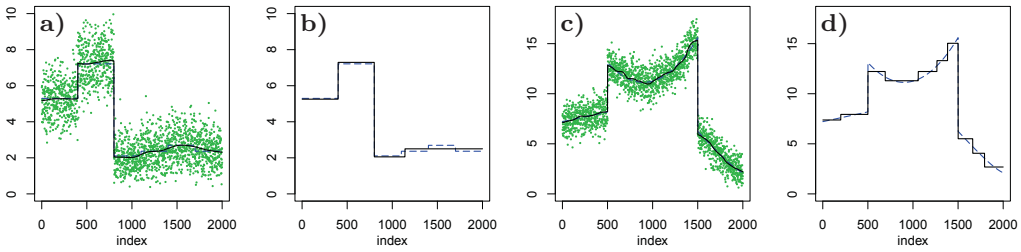

Figure 5.12: Results of the simplified Propagation-Separation Approach on non-central chi-distributed observations for an optimal and for an extremely large bandwidth, where the true parameter function is shown as blue dashed line and the corresponding estimation function is represented by a black solid line. In (a) and (c), the observations are visualized by small circles.

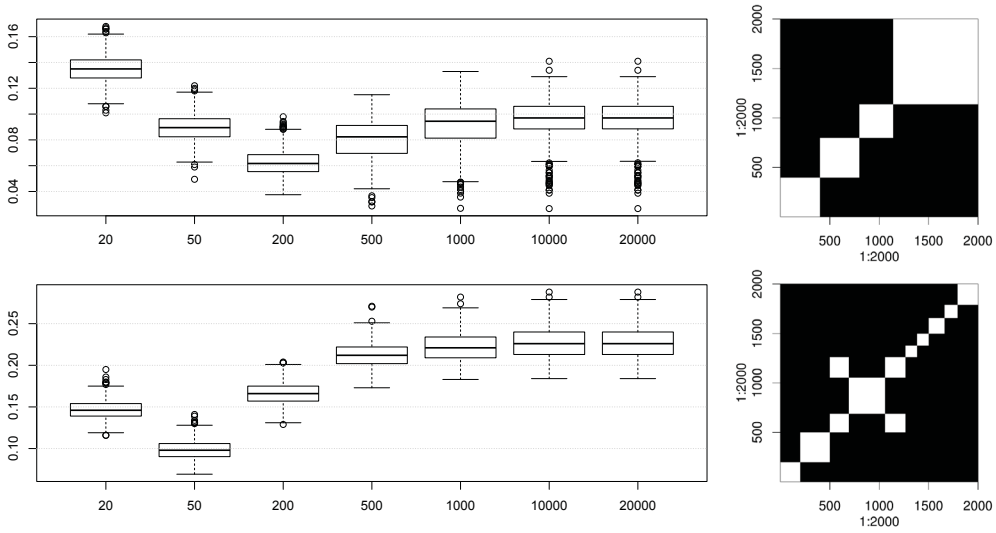

Figure 5.13: Stability of estimates and final weighting schemes for non-central chi-distributed observations using the same test functions as in Figure 5.12.

5.4 Discussion of the simplified Propagation-Separation Approach

Our study provides theoretical and numerical results for the simplified Propagation-Separation Approach in Algorithm 2 (page 18), where the memory step is omitted. This helps for a better understanding of the procedure as the impact and interaction of the involved components is clarified. Furthermore, the presented results substantiate the reasons for omitting the memory step and provide a detailed study of its impact.

Our theoretical results, see Chapter 4, rely on an advanced parameter choice for the adaptation bandwidth λ, which we introduced and justified in Chapter 3. In practice, the corresponding propagation condition yields a better interpretability of the adaptation bandwidth λ due to the precise information of the propagation level. In theory, the propagation condition leads to the propagation property and a certain stability of estimates for (piecewise) constant and (piecewise) bounded functions. These results demonstrate that the behavior of the algorithm, and hence the achievable quality of estimation, mainly depend on the sample size of the homogeneous regions, on the local smoothness of the parameter function $\theta(.)$, on the contrast between different regions,

and via the adaptation bandwidth λ on the parametric family $\mathcal{P} = \{\mathbb{P}_\theta\}_{\theta\in\Theta}$ of probability distributions.

Next we will discuss the following two questions.

- Does the simplified Propagation-Separation Approach converge?
- (Where) do we need the memory step?

Finally, we will give a brief overview on possible topics for future research.

5.4.1 Does the Propagation-Separation Approach converge?

In Section 4.3, we introduced a specific step function, which approximates the estimation function from the Propagation-Separation Approach. The formation of this *associated step function* can be explained as follows.

Since $\text{support}(K_{ad}) = [0,1)$, the statistical penalty $s_{ij}^{(k)}$ defined in Algorithm 2 (page 18) ensures zero weights $\tilde{w}_{ij}^{(k)} = 0$ if the Kullback-Leibler divergence of the estimators from the last iteration step $\mathcal{KL}(\tilde{\theta}_i^{(k-1)}, \tilde{\theta}_j^{(k-1)})$ exceeds some lower bound $\lambda/\tilde{N}_i^{(k-1)}$. Let us consider the case where $\overline{w}_{ij}^{(k^*)} > 0$ implies $\tilde{w}_{ij}^{(k^*)} > 0$ for all $i, j \in \{1, ..., n\}$. This means that separation did not occur. The monotonicity of the sequence of location bandwidths $\{h^{(k)}\}_{k=0}^{k^*}$ ensures that the non-adaptive weights increase during iteration, and, without separation, all estimators approach each other. With $h^{(k^*)}$ sufficiently large, this results in an almost constant estimation function. However, in many cases, there are $X_i, X_j \in \mathcal{X}$ such that $\overline{w}_{ij}^{(k^*)} > 0$, but $\tilde{w}_{ij}^{(k^*)} = 0$.

We know from Proposition 4.1 that separation occurs if $\max\{\mathcal{KL}(\theta_i, \theta_j) : X_i, X_j \in \mathcal{X}\}$ is sufficiently large, or if the algorithm adapts to outliers. The latter leads to separation of single observations, probably together with some local neighborhood. If separation happens due to the variability of the true parameter function $\theta(.)$, then it starts either at the boundaries of the design space $\mathcal{X}$ or close to discontinuities and local extrema of $\theta(.)$ and $\theta'(.)$. This leads, by subsequent attraction and repulsion of the estimators, to the formation of a step function which approximates the associated step function in Definition 4.14.

On the test functions in Section 5.1, we observed for Gaussian, exponential, and non-central chi-distributed observations that after separation has started, the algorithm behaves within each separated region similar as under homogeneity as long as the increasing local neighborhood does not reach a distant region with similar values. Additionally, for sufficiently large location bandwidths, the algorithm resulted for every test function in an adaptive weighting scheme, whose disjoint regions define a partition of the design space. This indicates the immutability of the associated step function for sufficiently large location bandwidths.

Hence, the presented numerical results suggest the convergence of the algorithm, but we lack for a theoretical justification. There are three main reasons for this.

- Each realization may yield another associated step function with slightly shifted steps.
- The improvement of the estimation quality during iteration is not ensured to be monotonic, neither for the non-adaptive nor for the adaptive estimates. Several other iterative methods, such as the expectation-maximization algorithm or the conjugate gradient method, rely on the minimization or maximization of a certain criterion. This provides a monotonic improvement of some quality criterion, which ensures the convergence of the algorithm. In contrast, the Propagation-Separation Approach considers an increasing local neighbor-

hood, where unfavorable, newly included or stronger weighted observations may worsen the estimation quality in comparison to a previous iteration step.

- The immutability of the associated step function for sufficiently large iteration steps requires the existence of some iteration step $k_0 < \infty$ such that the considered neighborhood equals the complete design, that is $\overline{w}_{ij}^{(k_0)} > 0$ for all $X_i, X_j \in \mathcal{X}$, and

$$\{X_j \in \mathcal{X} : s_{ij}^{(k_0)} \leq \lambda\} = \{X_j \in \mathcal{X} : s_{ij}^{(k)} \leq \lambda\}$$

for all $k > k_0$ and every $X_i \in \mathcal{X}$.

Let us consider the last reason in more detail. We know from the definition of the statistical penalty, see Algorithm 2 (page 18), that a violation of the above condition can arise from

1. a reunion of previously separated regions due to a decrease of the factor $\tilde{N}_i^{(k-1)}$;
2. a reunion of previously separated regions due to a decrease of the Kullback-Leibler divergence $\mathcal{KL}(\tilde{\theta}_i^{(k-1)}, \tilde{\theta}_j^{(k-1)})$;
3. a subsequent segmentation of a before created step due to an increase of the Kullback-Leibler divergence $\mathcal{KL}(\tilde{\theta}_i^{(k-1)}, \tilde{\theta}_j^{(k-1)})$;
4. a too strong intensification of the statistical penalty by the factor $\tilde{N}_i^{(k-1)}$.

We discuss these events case by case.

Recall that the non-adaptive sequence $\{\overline{N}_i^{(k)}\}_{k=0}^{k^*}$ is monotonically increasing, whereas its adaptive counterpart $\{\tilde{N}_i^{(k)}\}_{k=0}^{k^*}$ does not need to be monotonic. Therefore, we propose a slight modification of the statistical penalty in Algorithm 2 (page 18), setting

$$s_{ij}^{(k)} := \max_{k' \leq k} \tilde{N}_i^{(k'-1)} \mathcal{KL}(\tilde{\theta}_i^{(k-1)}, \tilde{\theta}_j^{(k-1)}).$$

This modification preserves an already achieved adaptation quality. As a consequence, it avoids that a design point switches all the time between two steps due to oscillation of the value of $\tilde{N}_i^{(k-1)}$ during iteration.

A late segmentation and a reunion as described in (2) and (3) could be imposed by an appropriate upper bound of

$$\max\left\{\mathcal{KL}\left(\tilde{\theta}_{i_1}^{(k)}, \tilde{\theta}_{i_2}^{(k)}\right) : X_{i_1}, X_{i_2} \in \mathcal{H}_i^{(k)}\right\}$$

and a lower bound of

$$\min\left\{\mathcal{KL}\left(\tilde{\theta}_{i_1}^{(k)}, \tilde{\theta}_{j_1}^{(k)}\right) : X_{i_1} \in \mathcal{H}_i^{(k)}, X_{j_1} \in \mathcal{H}_j^{(k)} \neq \mathcal{H}_i^{(k)}\right\},$$

where $\mathcal{H}_i^{(k)}$ is as in Equation (4.9) (page 57). Due to the factor $\varkappa$ in Lemma 2.5, the corresponding discussion in Section 2.4, and the missing monotonicity of the Kullback-Leibler divergences $\mathcal{KL}(\tilde{\theta}_i^{(k-1)}, \tilde{\theta}_j^{(k-1)})$ in $k > k_0$, this may lead to a criterion which is too restrictive to be satisfiable with $k_0 < \infty$.

However, the main impediment of a theoretical proof results from (4). The statistical penalty becomes more restrictive during iteration by the factor $\tilde{N}_i^{(k)}$, but this factor is not guaranteed to be always appropriate. For statistically independent observations $\{Y_j\}_j$ with expected values $\{\theta_j\}_j$ and variance σ^2, it may be explained as an upper bound of the achieved variance reduction, see Lemma 2.10 for the case of the non-adaptive estimator. A generalization to the adaptive estimator may be prohibitive due to the randomness of the adaptive weights. Additionally, for other classes of probability distributions than the Gaussian one, the relation

between the variance and the Kullback-Leibler divergence is complicated, and the variance may be heteroscedastic. For instance, for exponentially distributed observations, the variance depends on the locally varying parameter θ.

Therefore, we prefer to consider $\tilde{N}_i^{(k)}$ as the achieved improvement of the estimation quality in terms of the Kullback-Leibler divergence. This is motivated by the Theorems 2.13 and 2.24 and the propagation condition, which yield with high probability and in case of sharp discontinuities for $\mathcal{KL}(\tilde{\theta}_i^{(k)}, \mathcal{E}\tilde{\theta}_i^{(k)})$ the rate of convergence $\tilde{N}_i^{(k)}$, see Propositions 4.4 and 4.11. Nevertheless, there remains a certain probability of unfavorable realizations, for which the intensification of the statistical penalty is not justified. Furthermore, the mentioned propagation results do not generalize to the case of model misspecification. They are based on the propagation condition and this requires well separated regions. If the corresponding structural assumptions are violated, the impact of the adaptivity may change such that propagation cannot be ensured anymore. In fact, model misspecification leads to a decrease of the probability for propagation. Therefore, we may still observe propagation in practice, but its probability cannot be quantified as the established exponential bounds do not hold under model misspecification. As a consequence, we cannot ensure neither the immutability of the associated step function nor the convergence of the simplified Propagation Separation Approach.

5.4.2 (Where) do we need the memory step?

In order to clarify the impact of the memory step, we compare our theoretical results with the original study by Polzehl and Spokoiny [2006]. There, the authors demonstrated propagation, separation, and stability of estimates up to some constant. We will summarize these results briefly. Here, we have shown similar properties for the simplified algorithm, where the memory step is omitted.

Both studies include a certain separation property, see Polzehl and Spokoiny [2006, §5.5] and Proposition 4.1. This justifies that, in case of sufficiently large discontinuities, smoothing is restricted to the homogeneity regions.

For the propagation property, Polzehl and Spokoiny supposed, among other assumptions, the statistical independence of the adaptive weights from the observations. Then, for $\theta(.) \equiv \theta$ and all $k \in \{0, ..., k^*\}$, they showed that

$$\mathbb{P}\left(\overline{N}_i^{(k)} \mathcal{KL}\left(\hat{\theta}_i^{(k)}, \theta\right) \leq \mu \log(n) \quad \forall i\right) > 1 - 2k/n, \qquad \mu \geq 2, \tag{5.4}$$

where $\hat{\theta}_i^{(k)}$ denotes the adaptive estimator after modification by the memory step, see Algorithm 1 (page 16). For locally bounded parameter functions, the authors established a similar result. Equation (5.4) could be improved by Proposition 4.4, taking advantage of the new propagation condition in Section 3.1. For $z := \mu \log(n)$ and $\epsilon := c_\epsilon n^{-q}$, Proposition 4.4 implies

$$\mathbb{P}\left(\overline{N}_i^{(k)} \mathcal{KL}\left(\tilde{\theta}_i^{(k)}, \theta\right) \leq \mu \log(n) \quad \forall i\right) > 1 - \max\{2, c_\epsilon\}/n, \qquad \mu, q \geq 2,$$

where the additional factor k is avoided.

Theorems 4.9 and 4.12 shed light on the interplay of propagation and separation during iteration. Here, we do not restrict the analysis to the respective homogeneity region as in [Polzehl and Spokoiny, 2006] and Propositions 4.4 and 4.11. Instead, we use the separation property to verify the propagation property for piecewise constant and piecewise bounded functions with sharp discontinuities. Setting $z \geq \mu \log(n)$ and $\epsilon := c_\epsilon n^{-\mu}$ with $c_\epsilon > 0$ and $\mu \geq 2$, the resulting

exponential bound (4.6) (page 55) differs from Equation (5.4) by the terms $p_{\varkappa}$ and $\mathbb{P}(M^{(k)}(z))$. These are required for the separation of distinct homogeneity regions.

The results on the stability of estimates are difficult to compare. Our corresponding results are stated in Proposition 4.4 for constant parameter functions and in Proposition 4.11 for bounded parameter functions. Under weak assumptions, Polzehl and Spokoiny proved stability of estimates up to some constant. More precisely, they showed that

$$\overline{N}_i^{(k)} \mathcal{KL}\left(\hat{\theta}_i^{(k)}, \theta_i\right) \leq \mu \log(n)$$

implies with probability one that

$$\overline{N}_i^{(k)} \mathcal{KL}\left(\hat{\theta}_i^{(k^*)}, \theta_i\right) \leq c \log(n), \qquad c := \varkappa^2 \left(\sqrt{c_1 C_\tau} + \sqrt{\mu}\right)^2, \tag{5.5}$$

where $\varkappa$ is as in Notation 2.6, $\tau := C_\tau \log(n)$ denotes the bandwidth of the memory kernel, and $c_1 := \varkappa^2 \nu (1 - \sqrt{\nu})^{-2}$ depends on the constant ν satisfying $\nu_1 \leq \overline{N}_i^{(k-1)} / \overline{N}_i^{(k)} \leq \nu$ for every $k \in \{1, ..., k^*\}$ and $\nu_1, \nu \in (2/3, 1)$. Hence, the constant c might be large. Under smoothness conditions on the parameter function $\theta(.)$, this result allowed Polzehl and Spokoiny [2006] the verification of the optimal rate of convergence, up to a log-factor. Equation (5.5) is based on Lemma 2.8 and consequently requires that $\theta_i, \hat{\theta}_i^{(k)}, \hat{\theta}_i^{(k^*)} \in \Theta_{\varkappa}$ for some compact and convex subset $\Theta_{\varkappa} \subseteq \Theta$ as in Notation 2.6. This again leads to the discussion in Section 2.4, not mentioned by Polzehl and Spokoiny [2006].

Here, we did not study the asymptotic behavior of the Propagation-Separation Approach. This has the following reason. An asymptotic study requires to decrease the propagation level ϵ with increasing sample size n, $\lim_{n \to \infty} \epsilon(n) = 0$. However, the adaptation bandwidth λ depends on the propagation level ϵ. For a fixed sample size, the simulations in § 3.1.2 suggest that $\lim_{\epsilon \to 0} \lambda(\epsilon) = \infty$ holds under weak conditions. As large values of λ yield similar results as non-adaptive smoothing, this leads to a setting which is not convenient to study properties of the Propagation-Separation Approach. The adaptation bandwidth is the crucial parameter which distinguishes the Propagation-Separation Approach from non-adaptive smoothing. Hence, an asymptotic study provides little insight if $\lim_{n \to \infty} \lambda(n) = \infty$ or if we do not know how the increasing sample size affects the choice of the adaptation bandwidth given by the propagation condition.

In summary, there are two theoretical properties of the original Propagation-Separation Approach which could not be justified for the simplified version yet. First, our approach is not constructed to provide asymptotic results. Second, our stability results hold for constant and bounded parameter functions only. In other words, we lose the general stability of estimates in Equation (5.5). Nevertheless, the essential properties of the algorithm remain valid as these are propagation and separation. Both properties follow from the adaptivity of the estimator, but not from the memory step. Hence, for a piecewise bounded parameter function with sufficiently sharp discontinuities, the memory step is not needed.

From a practical point of view, the benefits of the memory step are still questionable. In § 5.3.3, we illustrated the impact of the memory step for several test functions. Using the default parameter choices of the **R**-package **aws** by Polzehl [2012], we could not observe any effect of the memory step. However, these choices are not arbitrary. The memory bandwidth was chosen in accordance with a former version of the propagation condition, and, indeed, we observed an increased risk of adaptation to noise for considerably smaller bandwidths. On the one hand, this emphasizes the importance of a sufficiently large memory bandwidth to avoid adaptation to

outliers. On the other hand, we got a smaller MAE by increasing the amount of aggregation, which slightly attenuated the formation of a step function during iteration. In any case, we found the best results by restricting the maximal location bandwidths appropriately. The omittance of the memory step provides a better interpretability of the procedure and, as a consequence, of the results since the memory step introduces additional interactions between the involved components, which are not fully understood yet.

5.4.3 Future research

There are several topics for future research that arise from this thesis. For instance, one could study the impact of the Kullback-Leibler divergence. Especially in Chapter 3, we took advantage of its properties. Are there other (possibly asymmetric) distance functions or f-divergences which provide similar results? Moreover, it could be interesting to study consequences of a random or irregular design. Which additional assumptions are required to extend the propagation condition from the artificial data set to the data at hand in this case? Is it still reasonable to estimate the propagation level on the basis of a single realization as proposed in Remark 3.12 for regular designs?

Here, we would like to concentrate on another question concerning the consequences of a violated structural assumption. As indicated by our numerical results in Section 5.3, an appropriate stopping of the iterative procedure may reduce the resulting estimation bias considerably by avoiding the formation of a step function. For the presented univariate examples, a choice by visual inspection seems to be promising. In all observed cases, the iteration step where the formation of the step function started to dominate the smoothing result could be easily identified. Additionally, we always observed a certain range of iteration steps where the estimation quality is very similar. However, on more complicated test functions or for higher dimensional design spaces, an automatic choice of the maximal number of iterations is desired or even required. In the context of local polynomial regression and locally weighted maximum likelihood estimation, there is a large amount of literature concerning the choice of the location bandwidth. For instance, the maximal location bandwidth $h^{(k^*)}$ could be chosen such that the non-adaptive estimator in Notation 2.14 behaves well within regions without discontinuities. Then, assuming an appropriate choice of the adaptation bandwidth λ, Algorithm 2 would yield similar results as non-adaptive smoothing within these regions, while smoothing among distinct regions would be avoided as sharp discontinuities could be detected by the adaptive weights. To evaluate the appropriateness of the different approaches for the Propagation-Separation Algorithm would form a promising research project for the future. Alternatively, one could search for a criterion which takes advantage of the involved components of the method. The evaluation of the behavior of the statistical penalty or of the sum of the adaptive weights could provide useful information about the iteration step where the formation of a step function negatively affects the smoothing results.

Chapter 6

Application to magnetic resonance imaging

In this chapter we will demonstrate the practical value of the Propagation-Separation Approach with the example of diffusion-weighted magnetic resonance images (dMRI). This is a versatile technique for the in-vivo examination of microstructures in the body, such as muscle tissue [Sinha et al., 2006], the spinal cord [Clark et al., 1999], or neuronal fibers in the human brain [Le Bihan, 2003]. In medicine, it is used for the diagnosis of diseases and for presurgical planning. The neuroscientific community is interested, for instance, in the connectivities of the brain and in changes due to aging and diseases [Johansen-Berg and Behrens, 2009; Jones, 2010; Mori, 2007].

The application of the Propagation-Separation Approach requires a description of the data in accordance with the local likelihood model in Notation 2.1. Therefore, we need some information about the measurement process. DMRI data are acquired on a three-dimensional grid of volume elements, called voxels, applying varying diffusion-weighting magnetic field gradients. These are fully described by their direction and the corresponding b-value $b > 0$ which relates to the magnetic field strength, its duration, and the diffusion time before application of a transverse gradient [Johansen-Berg and Behrens, 2009, Eq. (3.2)]. In practice, measurements are usually performed with a small number of b-values. For each b-value, the data can be described as a real valued function on the measurement space $\mathbb{R}^3 \times \mathbb{S}^2$, specifying the (voxel) position and (gradient) direction, see § 6.1.2 for more details. This is a specific perspective on the data. Its advantages will be discussed in § 6.6.4.

The diffusion-weighted images provide information about the diffusion profile in each voxel. This reveals intra-voxel information at a micron level, although the data are acquired at a millimeter scale, see for instance Mitra and Sen [1992]. For the analysis of dMRI data, a wide range of diffusion models have been introduced, such as the diffusion tensor model (DTI), tensor mixture models, diffusion spectrum imaging (DSI), and many more. Some of them can be evaluated based on a single q-shell, that is on measurements with a single b-value. Other models require multi-shell data, where varying b-values are applied. This kind of data is becoming increasingly popular for the exploration of the white matter anatomy, where diffusion models beyond the diffusion tensor should be used, see Johansen-Berg and Behrens [2009, Ch. 6] and Jones et al. [2013]. To date, there is no model which is generally satisfactory. In § 6.1.3, we will give an overview of the most common diffusion models.

As all imaging techniques, dMRI suffers from random noise. This may render the subsequent analysis and medical decisions more difficult by creating artifacts or a systematical bias of model features [Basser and Pajevic, 2000; Jones and Basser, 2004]. Recently, the interest in high spatial resolution imaging has increased as this is believed to improve the resolution of complex fiber bundles in the human body [Heidemann et al., 2010; Kamali et al., 2013; Kleinnijenhuis et al., 2012; Zhan et al., 2012]. Increasing the spatial resolution inherently reduces the signal-to-noise ratio (SNR), which is defined as the mean signal value divided by the noise standard deviation. Hence, attempts to achieve a higher image resolution collide with the deterioration of the acquired images [Johansen-Berg and Behrens, 2009; Lohmann et al., 2010]. Similarly, large b-values, as used for multi-shell measurements, lead to a very low SNR due to the resulting signal attenuation. Thus, the noise negatively affects the modeling and analysis of dMRI data in general, but for high spatial resolution and multi-shell data in particular.

In order to reduce the noise in dMRI data, a number of different approaches have been developed. The most common and simplest method is Gaussian filtering [Westin et al., 1999]. However, isotropic smoothing blurs fine structures. Therefore, more sophisticated methods have been developed, which we will discuss in § 6.6.5. It is generally preferable to smooth the diffusion-weighted images directly, rather than model dependent derived quantities. This has several advantages. First, direct smoothing of the MR images avoids the bias of model dependent estimates induced by the noise. Second, this approach enables a subsequent analysis with any diffusion model. Additionally, a previous noise reduction can stabilize modeling, such as for DTI, where smoothing reduces the probability of estimating a degenerated tensor [Tabelow et al., 2008].

In this chapter, we develop a position-orientation adaptive smoothing method (POAS) for single-shell data and a generalization for multi-shell data, called msPOAS. These methods directly smooth the diffusion-weighted images prior to any modeling. Both procedures benefit from the whole information of the data by considering the measurement space $\mathbb{R}^3 \times \mathbb{S}^2$, that is positions and orientations. The method msPAOS gains additional efficiency by a simultaneous smoothing of all q-shells, that is for all applied b-values. To our best knowledge, this is the first algorithm which is constructed to smooth the diffusion-weighted images of multi-shell dMRI data, using the whole information in position and orientation and a vector structure of the data from the different shells.

The algorithm of (ms)POAS is based on the simplified Propagation-Separation Approach, see Algorithm 2 (page 18). In the context of dMRI, the Propagation-Separation Approach is of special interest as it is constructed to smooth piecewise polynomial functions without blurring at discontinuities. Moreover, it applies to any real valued function whose domain is endowed with a distance function, see § 2.1.1. Hence, in order to apply the Propagation-Separation Approach on dMRI data, we need a metric on $\mathbb{R}^3 \times \mathbb{S}^2$ which we can use for the location kernel.

The outline of the chapter is as follows. The first section will summarize the general concepts of dMRI. Readers who feel familiar with this topic may concentrate on the graphical abstract in Figure 6.4 (page 91). In Section 6.2, we will introduce our smoothing method msPOAS and its special case POAS. Then, we will follow an approach by Duits and Franken [2011], embedding the space $\mathbb{R}^3 \times \mathbb{S}^2$ into the special Euclidean motion group $\mathrm{SE}(3)$. This approach enables the application of orientation score theory, and it provides a criterion for the appropriateness of an operator for processing dMRI data. We will prove (ms)POAS to satisfy the required properties. In Section 6.4, we will present two distance functions on $\mathbb{R}^3 \times \mathbb{S}^2$ which can be used in (ms)POAS. This is the crucial point to make (ms)POAS applicable. Then, we will demonstrate on simulated and experimental data that (ms)POAS significantly improves the

quality of the diffusion-weighted data. As it turns out, the method is suitable even for very low SNR. We will finish with a discussion of the presented results, a brief comparison to other smoothing methods and further comments on advantages, drawbacks and pitfalls occurring when using (ms)POAS. The chapter is closely related to the recent publications [Becker et al., 2012] and [Becker et al., 2013].

6.1 General concepts of diffusion-weighted MRI

We aim to provide a better understanding of dMRI. For this purpose, we will review the physical foundation of the measurement process and recent modeling concepts. In particular, we will present a specific description of the measured data and discuss their probability distribution. For the smoothing method (ms)POAS, Equations (6.2), (6.6), and (6.7) will be of particular interest.

6.1.1 Physical foundation and data acquisition

DMRI measures the diffusion of water in the human body to deduce the underlying tissue structure. The sensitivity on molecular diffusion results from a signal attenuation as diffusion disturbs the reproduction of the signal during the pulsed-gradient spin echo (PGSE) sequence, which we will introduce in the next subsection. First, we take a very basic look at the underlying physics. In principle, magnetic resonance imaging (MRI) is based on quantum mechanics, but Bohr's correspondence principle allows a semi-classical description [Oppelt, 2005, §7.2.2]. The signal acquisition is consistent for all MR techniques, such as diffusion-weighted MRI (dMRI), T_1-, and T_2-images, functional MRI (fMRI), and MR angiography (MRA), see Oppelt [2005, §7 & 15]. The specifics of dMRI follow from the mentioned PGSE sequence. For more details, we refer the reader to Johansen-Berg and Behrens [2009, §I] and in addition to Callaghan [2007], and Minati and Weglarz [2007].

The human body consists, at great expense, of water and consequently of protons. These protons possess a spin, creating a local magnetic field. In the case of an additional external magnetic field B, the spins of the protons align themselves with its main direction. Their precession is then effected at an angular frequency of $\omega = \gamma B$, called the Lamor frequency. The constant γ denotes the gyromagnetic ratio which depends on the respective atomic nucleus, here the proton.

By excitation with a transverse $\pi/2$ ($90°$) radiofrequency magnetic wave (rf-pulse) at the Lamor frequency of the spins, they can be tilted into the plane which is orthogonal to the magnetic field B. The resulting synchronized precession around the magnetic field direction is known as the Lamor precession phenomenon. The excited spins realign themselves with the main direction of the magnetic field B in exponential time with constant T_1, and the precession part of the spin magnetization decays exponentially with a time constant T_2. This is called relaxation. Both, T_1 and T_2 depend on the tissue [Bottomley et al., 1984]. The former provides excellent morphological information, describing the form and structure of organisms, such as organs or tumors. The latter is often used for lesion characterization, having broad application in the context of diseases with accumulation of fluid, such as in edemas or cysts. More details and references can be found in [Oppelt, 2005, §15.2].

Additional to the relaxation, the synchronized spins start to dephase. The resulting dephasing can be compensated via a subsequent application of a π ($180°$) rf-pulse, which changes the spin direction. Then, under consistent conditions and in the case of only small inhomogeneities, the spins refocus with the result that the original signal magnitude is reproduced, except for the T_1

and T_2 decay. Coherently precessing spins induce a current which may be measured in one or more receiver antennas, known as magnetic resonance coils. From this spin echo the MR image is generated.

In MRI, we take advantage of the fact that excitation of the spins only occurs if the rf-pulse is applied at the respective Lamor frequency. This allows to relate the signals to their corresponding volume elements (voxels) by labeling the spins in dependence of their location. The MR tomograph generates a homogeneous magnetic field B_0 with the result that the proton spins precess at a homogeneous frequency. Then, we apply three magnetic field gradients. Each of them creates a linear change of the homogeneous field B_0 and consequently implies a linear change in the precessional frequency ω of the spins along the direction of the respective gradient.

One gradient is applied simultaneously to the $\pi/2$ rf-pulse. Then, the given frequency of this pulse only corresponds to one plane (slice) of spins. This plane is perpendicular to the applied gradient and contains all spins which precess with the respective Lamor frequency. Hence, a change of the rf-pulse frequency enables the selection of varying slices. The slice thickness is determined by the gradient strength and the frequency bandwidth of the rf-pulse. By convention, the excited slice is perpendicular to the z-axis. For each selected slice, that is for each rf-pulse frequency, two other gradients are applied in directions of the x- and y-axes, labeling the in-plane locations by a dephasing of the spin precession. The three gradients vary over time. More precisely, they are rapidly turned on and off and lead to an inhomogeneous magnetic field [Johansen-Berg and Behrens, 2009, Eq. (2.2)],

$$B = B_0 + G_x(t)x + G_y(t)y + G_z(t)z.$$

For a fixed slice (one rf-pulse frequency), each signal measured in a receiver coil equals the integral over the signals from all excited spins in the (x, y)-plane [Johansen-Berg and Behrens, 2009, Eq. (2.6)],

$$F(k_x, k_y) = \int f(x, y)\, e^{i2\pi(k_x x + k_y y)}\, dx\, dy, \tag{6.1}$$

where $f(x, y)$ denotes the net magnetization across the excited slice, and $2\pi(k_x x + k_y y)$ equals the location dependent phase which each excited spin accumulates over time with

$$k_x := \gamma \int G_x(t)dt \quad \text{and} \quad k_y := \gamma \int G_y(t)dt.$$

The function F equals the Fourier transform of the net magnetization f, which can be reconstructed by an inverse Fourier transform if $F(k_x, k_y)$ was measured for sufficiently many values. The domain of the measured signal F as a function of k_x and k_y is called k-space, and f is known as MR image. Here, the edges and details in the MR image f are represented by the values of F at higher values of k, while low values of k relate to regional values of f.

Hence, a magnetic resonance scan does not directly yield the MR image. Instead, it generates complex valued data in k-space. For measurements with a single receiver coil, this leads to the MR image via the inverse Fourier transform. In the case of more than one receiver coil, that is for parallel imaging, the image reconstruction combines the signals of all coils. The details depend on the respective acquisition method, such as SENSE [Pruessmann et al., 1999] or GRAPPA [Griswold et al., 2002] among others. In the image space, one typically considers the absolute value of the resulting complex signal [Aja-Fernández et al., 2009; Dietrich et al., 2008; Lohmann et al., 2010]. This equals the magnitude and neglects the phase information. In summary, we get for each volume element (voxel), described by its center $\vec{v} \in \mathbb{R}^3$, a real valued signal $S(\vec{v}) \in \mathbb{R}$.

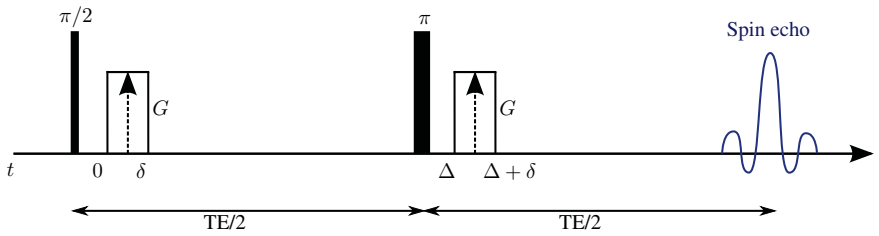

Figure 6.1: Schematic of the pulsed-gradient spin echo (PGSE) sequence.

6.1.2 Measurement process

The power of magnetic resonance imaging lies in the variability of the MR pulse sequence. This is a unique feature in medical imaging which enables easy adjustment of the measurement process. Many modifications have been designed which optimize the image quality and provide different information, such as tissue contrast, blood flow or diffusion properties. In diffusion-weighted MRI, we aim to examine not only the underlying tissue, but the direction and connectivity of possibly existing fibers. This is enabled by the fact that, within a fibrous structure, the diffusion of water is anisotropic with diffusion maxima in direction of the present fibers, see Moseley et al. [1990] or Johansen-Berg and Behrens [2009, Ch. 6]. Hence, we use additional diffusion-weighting gradients, applied in various directions, to attenuate the signal $S(\vec{v})$ in dependence of the diffusion properties within the respective voxel.

The measurement process of dMRI follows the pulsed-gradient spin echo (PGSE) sequence, introduced by Stejskal and Tanner [1965] and summarized, for example, by Johansen-Berg and Behrens [2009, Ch. 2]. Its schematic is shown in Figure 6.1. We go through the sequence step by step. As before, the MR tomograph generates a homogeneous magnetic field B_0. A $\pi/2$ rf-pulse excites the proton spins within the thereby selected slice in dependence of the simultaneously applied localization gradient G_z. Then, a diffusion-weighting magnetic field gradient G is applied for a short time δ. Similar to the localization gradients, this diffusion-weighting gradient introduces a phase shift in the precession of the excited spins. Subsequent excitation with a π $(180°)$ rf-pulse changes as before the spin direction. Then, reapplication of the diffusion-weighting gradient G reverses the dephasing by the magnetic field inhomogeneity. The spin precession is refocused as in the spin echo experiment. Thus, in a completely unchanged setting, the diffusion-weighting gradient would not have any effect. However, the permanent diffusion of the water molecules causes a rearrangement of the labeled spins before the second rf-pulse is applied. The hindered resynchronization after excitation with the π rf-pulse leads to a considerable signal attenuation. In order to minimize the diffusion effect during application of the diffusion-weighting gradients, the application time δ should be much smaller than the diffusion time Δ between the two gradients, $\delta \ll \Delta$. After performance of the whole PGSE sequence, the data are collected by the localization gradients G_x and G_y described above.

Each diffusion-weighting gradient yields a diffusion-weighted image on $\mathbb{R}^3$ whose data acquisition follows the procedure which we described in § 6.1.1. The directions of the diffusion-weighting gradients can be identified with elements of the 2-sphere

$$\mathbb{S}^2 := \{\vec{g} \in \mathbb{R}^3 : \|\vec{g}\| = 1\}.$$

Additionally, the diffusion-weighting gradients depend on the applied magnetic field strength, its duration, and the diffusion time δ, which together define the b-value $b > 0$ [Johansen-Berg

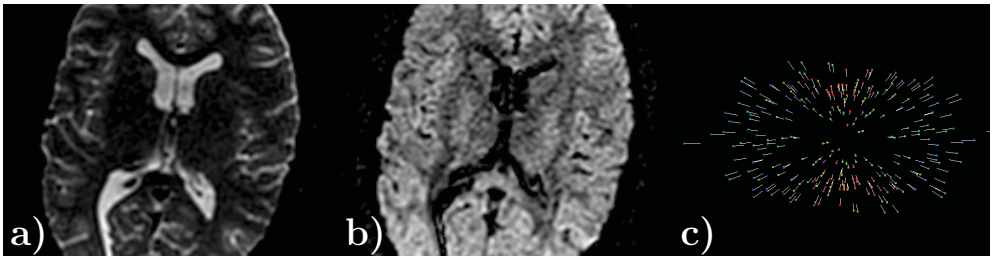

Figure 6.2: Diffusion-weighted data: a) slice of the non-diffusion-weighted image $\{S_0(\vec{v})\}_{\vec{v}\in\mathbb{R}^3}$, b) same slice of the diffusion-weighted image $\{S_b(\vec{v},\vec{g}_*)\}_{\vec{v}\in\mathbb{R}^3}$ for *a single* (arbitrarily selected) diffusion-weighting gradient direction $\vec{g}_* \in G_b$ with b-value $b = 1000s/mm^2$, c) diffusion-weighted data $\{S_b(\vec{v}_*,\vec{g})\}_{\vec{g}\in G_b}$ for all diffusion gradients in a *single voxel* $\vec{v}_* \in \mathbb{R}^3$ in corpus callosum. For better visibility, the diffusion-weighted image in (b) has been scaled. The unscaled image is very dark due to the signal attenuation. The single voxel visualization in (c) is created from a three-dimensional plot of the data, where the diffusion-weighted values are shown in their corresponding gradient direction.

and Behrens, 2009, Eq. (3.2)]. Then, for each b-value, the data can be considered as a real valued sample on the measurement space $\mathbb{R}^3 \times \mathbb{S}^2$, given as

$$S_b : V \times G_b \ni (\vec{v},\vec{g}) \mapsto S_b(\vec{v},\vec{g}) \in \mathbb{R}, \qquad V \times G_b \subseteq \mathbb{R}^3 \times \mathbb{S}^2, \tag{6.2}$$

where V denotes the voxel space and G_b the set of applied diffusion-weighting gradient directions at b-value $b > 0$. Alternative approaches will be discussed in § 6.6.4. We include the non-diffusion-weighted images into this notation, setting $b = 0$, $G_b = \{\vec{0}\}$, and $S_0(\vec{v},\vec{0}) := S_0(\vec{v})$. Then, each three-dimensional MR image corresponds to some data set $\{S_b(\vec{v},\vec{g}_*)\}_{\vec{v}\in\mathbb{R}^3}$, where the b-value $b \geq 0$ and the diffusion-weighting gradient direction $\vec{g}_* \in G_b \subseteq \mathbb{S}^2 \cup \{\vec{0}\}$ are fixed. Conversely, for every voxel $\vec{v}_* \in \mathbb{R}^3$, we have one value per gradient direction $\{S_b(\vec{v}_*,\vec{g})\}_{\vec{g}\in G_b}$, where the b-value $b \geq 0$ is fixed. Both cases are visualized in Figure 6.2. The data set was provided by H.U. Voss (Weill Cornell Medical College, New York, USA).

The PGSE sequence results in a signal attenuation due to relaxation and diffusion. While T_1- and T_2-images concentrate on the relaxation effects, dMRI is targeted on the consequences of diffusion. The relaxation-related attenuation mainly depends on the location, but it is almost independent of the applied diffusion gradients. Therefore, we may (almost) eliminate its effects by considering the quotient $S_b(\vec{v},\vec{g})/S_0(\vec{v})$ of the diffusion-weighted signal $S_b(\vec{v},\vec{g})$ and the non-diffused signal $S_0(\vec{v})$, where no magnetic field gradient was applied. It can be shown [Johansen-Berg and Behrens, 2009, Eq. (1.8)] that for fixed $\Delta, \delta > 0$ the resulting signal attenuation equals

$$S_b(\vec{v},\vec{g})/S_0(\vec{v}) = \int_{x_1\in\vec{v}} \int_{x_2\in\mathbb{R}^3} e^{-iq_{(\vec{g},b)}(x_2-x_1)} P(x_1,x_2,\Delta) dx_2 \rho(x_1) dx_1. \tag{6.3}$$

This depends on the following components. The spin density ρ describes the probability of finding a proton at location x_1 at the time of application of the first rf-pulse. The diffusion propagator P represents the probability that a proton diffuses from location x_1 to x_2 in time Δ. The Fourier kernel $e^{-iq_{(\vec{g},b)}(x_2-x_1)}$ relates to the frequency response of the signal, depending on the constant $q_{(\vec{g},b)} := \gamma\delta G_{(\vec{g},b)}$ with magnetic field gradient $G_{(\vec{g},b)}$ and γ, δ as introduced above.

Hence, together with at least one non-diffusion-weighted image, each diffusion-weighted image reveals information about the diffusion in the corresponding direction at each voxel. In the starting time of magnetic resonance imaging, a single diffusion-weighted image was acquired. Later on, it was recognized that the derived quantities should be invariant with respect to the

rotation of the sample in the scanner. Additionally, it turned out that it is more practical to uniformly sample the diffusion directions from a small number of spheres with different radii, which depend on the corresponding b-values, than using samples on a Cartesian grid. Then, measuring a certain number of gradient directions allows to examine the fibrous structure at a micron level, although the data are acquired at a millimeter scale. The power of diffusion-weighted MRI lies in this intra-voxel information which may be carved out via appropriate diffusion models, see below. Generally, a local minimum of the signal indicates a diffusion maximum and hence the presence of a fiber bundle which follows the corresponding direction.

6.1.3 Modeling

We are interested in the diffusion propagator P in Equation (6.3), which provides information about the microstructural features of the examined object. For this purpose, we distinguish between model-based and non-parametric approaches. The former models distinct fiber populations separately, and the latter considers the probability distribution of the fiber directions. For recent reviews of MRI modeling, we refer the reader to Assemlal et al. [2011]; Minati and Weglarz [2007] and Johansen-Berg and Behrens [2009, Ch. 4].

The most popular and simple model is called diffusion tensor imaging (DTI). It supposes a single fiber direction per voxel with Gaussian diffusion. This can be justified either by the random walk theory or via Fick's (empirical) laws on water diffusion, supposing free and unrestricted diffusion within the fibers, see Hagmann [2005], Minati and Weglarz [2007] or Johansen-Berg and Behrens [2009, Ch. 3]. Then, it follows [Basser et al., 1994b,a] that

$$S_b(\vec{v},\vec{g})/S_0(\vec{v}) = \exp\left(-b\vec{g}^T\mathbf{D}_{\vec{v}}^{-1}\vec{g}\right),$$

where $\mathbf{D}_{\vec{v}}$ denotes the diffusion tensor in voxel $\vec{v}$, which is represented by a symmetric and positive definite 3×3 matrix. From this equation, the diffusion tensor can be estimated using different methods. This allows to derive important quantities from the corresponding eigenvalues $\lambda_1,\lambda_2,\lambda_3\in\mathbb{R}$, such as the mean diffusivity $D_{\text{mean}} := (\lambda_1+\lambda_2+\lambda_3)/3$ or the fractional anisotropy

$$\text{FA} := \sqrt{\frac{3}{2}}\sqrt{\frac{(\lambda_1-D_{\text{mean}})^2+(\lambda_2-D_{\text{mean}})^2+(\lambda_3-D_{\text{mean}})^2}{\lambda_1^2+\lambda_2^2+\lambda_3^2}}. \tag{6.4}$$

DTI is used for a wide range of clinical and neuroscience applications [Johansen-Berg and Behrens, 2009; Jones, 2010; Mori, 2007]. However, for the more realistic case of inhomogeneous and restricted diffusion, it only provides an approximation, spurring interest in models beyond DTI.

The main drawback of DTI is its restriction to a single fiber per voxel. Neuronal fibers have a diameter at a micrometer scale and a length of up to several centimeter, but the measurements are realized at a millimeter scale, see for instance Mitra and Sen [1992]. Hence, in many voxels, there are not only one fiber, but a whole fiber bundle and possibly even crossings of fiber bundles with distinct directions [Johansen-Berg and Behrens, 2009, Ch. 6]. This is known as the partial volume effect [Johansen-Berg and Behrens, 2009, Ch. 9 Box 9.1]. A direct generalization of DTI takes this effect into account. Assuming that diffusion is Gaussian along all present fibers and that the signal adds independently, the tensor mixture model [Assaf and Basser, 2005; Behrens et al., 2003; Tuch et al., 2002] describes the signal by the weighted sum

$$S_b(\vec{v},\vec{g})/S_0(\vec{v}) = \sum_{i=1}^{n} f_i(\vec{v})\exp\left(-b\vec{g}^T\mathbf{D}_{(\vec{v},i)}^{-1}\vec{g}\right),$$

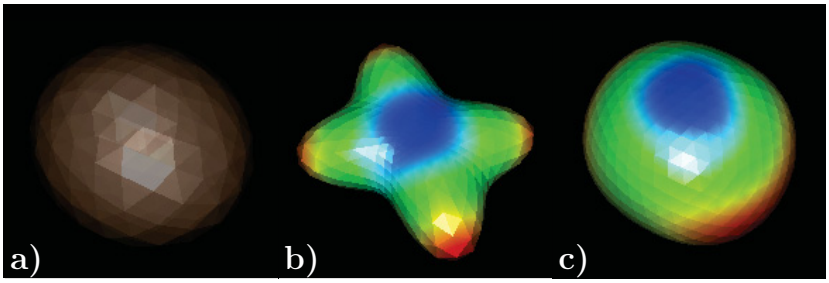

Figure 6.3: Different diffusion models in a fixed voxel of the data set in Figure 6.2: a) Diffusion tensor, b) Mixed tensor, c) Q-ball.

where $f_i(\vec{v})$ and $\mathbf{D}_{(\vec{v},i)}$ are the volume fractions and diffusion tensors corresponding to the n fiber bundles in voxel $\vec{v}$. Often, n is chosen a priori. An adaptive and voxel dependent choice can be found in [Tabelow et al., 2012]. Sometimes, an additional term is added which describes isotropic diffusion.

Higher order tensor models [Liu et al., 2003; Özarslan and Mareci, 2003] and diffusion kurtosis imaging (DKI) [Jensen et al., 2005] treat the restriction of DTI to Gaussian diffusion. DKI considers the cumulant expansion of Equation (6.3), which requires data acquisition with at least two non-zero b-values [Tabesh et al., 2011]. It includes DTI when terminating the expansion at the second term. The additional terms model deviation of Gaussian behavior.

In order to deduce the diffusion propagator P in Equation (6.3), one may fit mathematical models to the acquired data using Equation (6.3). The missing knowledge of the spin density ρ can be handled by introducing a net displacement variable $\vec{x} = x_2 - x_1 \in \mathbb{R}^3$ with the result, see Johansen-Berg and Behrens [2009, Eqs. (1.12) & (1.13)], that

$$S_b(\vec{v},\vec{g})/S_0(\vec{v}) = \int_{\mathbb{R}^3} \overline{P}(\vec{x},\Delta) e^{-iq\vec{x}} d\vec{x} =: E(q), \tag{6.5}$$

where the ensemble average propagator (EAP) in voxel $\vec{v}$ equals

$$\overline{P}(\vec{x},\Delta) := \int_{x_1 \in \vec{v}} P(x_1, x_1 + \vec{x}, \Delta)\rho(x_1) d(x_1) \qquad \vec{x} \in \mathbb{R}^3, \Delta > 0.$$

Then, the EAP follows by inverting the Fourier transformation in Equation (6.5), providing the required information for the reconstruction of the diffusion spectra. However, diffusion spectrum imaging (DSI) requires, for the inverse Fourier transform, data acquisition with diffusion-weighting gradients on a sufficiently large three-dimensional grid of points, that is with varying directions $\vec{g} \in \mathbb{S}^2$ and varying b-values $b > 0$. This leads to an overly long acquisition time.

In contrast, QBall imaging (QBI) estimates the well-known orientation distribution function (ODF) on the basis of data measured with a single b-value. Here, we use the Funk Radon transform and spherical interpolation to estimate the ODF which equals the radial summation of the above EAP,

$$ODF(\vec{u}) = \int_0^\infty \overline{P}(r\vec{u},\Delta) dr, \qquad u \in \mathbb{S}^2, \Delta > 0,$$

as introduced by Tuch [2002, 2004], but see also Johansen-Berg and Behrens [2009, Eq. (4.4)]. In other words, the ODF measures the amount of diffusion for directions $\vec{u} \in \mathbb{S}^2$. However, this approach relies on the unrealistic assumption of infinitely short pulses, leading to a considerable blurring of the derived ODF [Johansen-Berg and Behrens, 2009, Ch. 4 p. 62]. Hence, the

acquisition requirements are more manageable than for DSI, while the precision of the peak direction is questionable.

Other approaches use spherical deconvolution [Tournier et al., 2004] or other projections onto the sphere as the persistent angular structure (PAS) introduced by Jansons and Alexander [2003]. The methods by Aganj et al. [2010], Cheng et al. [2010], and Descoteaux et al. [2011] again require multi-shell data for estimation of the full diffusion propagator or its radial part.

Then, the deduced diffusion models can be used for the analysis of the data. This is often done by assigning a local anisotropy or diffusivity measure or by determining the main diffusion direction to produce fiber tracks. In Figure 6.3, we show a visualization of the diffusion tensor model, the mixed tensor model, and QBI for a single voxel of the dataset in Figure 6.2.

6.1.4 Random noise and other artifacts

The data acquisition suffers from several artifacts which require an appropriate pre-processing before the modeling and analysis of the data. For instance, motion, magnetic field inhomogeneities, eddy currents, and random noise may disturb the MR image [Johansen-Berg and Behrens, 2009, Ch. 3]. The respective methods are applied at different points of the processing pipeline. In this study, we will concentrate on the reduction of the random noise. This requires the specification of the probability distribution of the observations.

For simplicity, we denote both the realizations of our observations and the associated random variables by S_b with $b \geq 0$ fixed. The complex dMRI signal in k-space is usually assumed to be Gaussian distributed. In general experimental setups, this is justified by the central limit theorem since we measure integral values on the respective voxels. Hence, the final measurement error equals the sum of many independent random errors. The Gaussian distribution is invariant with respect to the inverse Fourier transformation [Bracewell, 1978, Ch. 7 Ex. 3]. Therefore, as well the complex reconstructed MR image f is Gaussian distributed. We assume the variance σ^2 of the imaginary part $\mathrm{Im}[f(\vec{v},\vec{g})]$ and of the real part $\mathrm{Re}[f(\vec{v},\vec{g})]$ of the signal f to be the same. With a single receiver antenna (MR coil), $L = 1$, this leads, for the standardized magnitude image,

$$S_b(\vec{v},\vec{g})/\sigma = \sqrt{\mathrm{Im}[f(\vec{v},\vec{g})]^2 + \mathrm{Re}[f(\vec{v},\vec{g})]^2}/\sigma,$$

to a non-central chi-distribution with two degrees of freedom and non-centrality parameter $\theta = \sqrt{\mathbb{E}\left[S_b^2(\vec{v},\vec{g})\right]/\sigma^2 - 2}$.

In the case of parallel imaging, that is for $L > 1$ MR coils, the probability distribution of the reconstructed signals S follows from the applied reconstruction algorithm. The simplest one is the sum-of-squares (SoS) formula [Aja-Fernández et al., 2009; Roemer et al., 1990] defined by

$$S_b(\vec{v},\vec{g}) := \sqrt{\sum_{c=1}^{L} S_{b,c}^2(\vec{v},\vec{g})} \quad \text{with} \quad S_{b,c}(\vec{v},\vec{g}) = \sqrt{\mathrm{Im}[f_c(\vec{v},\vec{g})]^2 + \mathrm{Re}[f_c(\vec{v},\vec{g})]^2},$$

where the b-value $b > 0$ is fixed and $S_{b,c}^2(\vec{v},\vec{g})$ equals, for the receiver coil $c \in \{1,...,L\}$, the absolute value of the inverse Fourier transformed signal f_c. Generally, the SoS formula does not require any assumptions on the magnetic fields of the different receiver coils. With a homogeneous variance σ^2 for all receiver coils, it yields

$$S_b(\vec{v},\vec{g})/\sigma \sim \chi_{2L}(\theta) \quad \text{with} \quad \theta = \sqrt{\mathbb{E}\left[S_b^2(\vec{v},\vec{g})\right]/\sigma^2 - 2L}. \tag{6.6}$$

Here, the non-centrality parameter follows from the second moment $\mathbb{E}\left[S_b^2(\vec{v},\vec{g})\right]$ of the signal $S_b(\vec{v},\vec{g})$. In our method (ms)POAS, we consider instead the expected value $\mathbb{E}[S_b(\vec{v},\vec{g})]$, which relates to the parameter θ via the function

$$\mu(\theta) := \sqrt{\frac{\pi}{2}}\,\frac{\Gamma(L+\frac{1}{2})}{\Gamma(\frac{3}{2})\Gamma(L)}\,{}_1F_1\left(-\frac{1}{2}, L, -\frac{\theta^2}{2}\right) = \sqrt{\frac{\pi}{2}}\,\mathbb{L}_{1/2}^{(L-1)}\left(-\frac{\theta^2}{2}\right) = \mathbb{E}\left[S_b(\vec{v},\vec{g})\right]/\sigma, \quad (6.7)$$

where $\mathbb{L}_{1/2}^{(L-1)}$ denotes the generalized Laguerre polynomial, Γ is the Gamma function, and ${}_1F_1$ the confluent hypergeometric function. For their relation, which we used in Equation (6.7), see for example El-Sayed [2000, Eq. 11] and Olver et al. [2010, Eq. 13.3.4].

However, the SoS reconstruction is inefficient since it does not reflect the location dependent sensitivity of the different MR coils. Instead, we may combine location dependent subsets of images, using a weighted SoS with weights equal to zero or one. Then, the non-central chi-distribution remains a valid model, with lower number of degrees of freedom. More sophisticated reconstruction algorithms like SENSE and GRAPPA lead to correlated noise with heteroscedastic variance and consequently slight changes in the distribution [Thunberg and Zetterberg, 2007]. Dietrich et al. [2008] demonstrated that the non-central chi-distribution is also approximately valid for these reconstruction methods.

For the sake of simplicity, we assume a fixed number of effectively utilized MR coils $L' \leq L$. Additionally, we assume the variance σ^2 to be known or an appropriate approximation $\hat{\sigma}^2$ to be achievable. Variance estimation for dMRI data is a challenging problem which is beyond the scope of this study. A survey of estimation procedures can be found in [Aja-Fernández et al., 2009]. Most of them are performed on the background of the MR image. Unfortunately, this cannot be defined for every data set. The approach in [Becker et al., 2012, App. C] is based on the Propagation-Separation Approach. It has the advantage to be also feasible for low signal-to-noise ratio and if no background can be defined in the image.

6.1.5 In a nutshell

Summing up, we recall the crucial steps. For a graphical abstract, we refer the reader to Figure 6.4.

The data acquisition takes advantage of the fact that coherently precessing proton spins provide a measurable signal, which can be localized via spin dephasing, using magnetic field gradients. Then, the PGSE sequence provides information on the diffusion characteristics, allowing the examination of the environmental structure at a micron level as diffusion is anisotropic within fibers. For each b-value, the measured signals can be described by a real valued function on the measurement space $\mathbb{R}^3 \times \mathbb{S}^2$. The data suffer from random noise, leading (approximately) to non-central chi-distributed observations. However, the respective degree of freedom may vary locally as it depends on the effective number of receiver coils. Usually, the variance is unknown and needs to be estimated.

The presented diffusion models provide information about the diffusion characteristics and principal fiber directions of the examined tissue, where the derived quantities should be invariant with respect to the rotation of the sample in the scanner. However, no model is generally satisfactory as all of them have some drawbacks. Most of the methods developed to overcome the limitations of diffusion tensor imaging (DTI) require a larger number of diffusion-weighting gradients and hence a longer acquisition time. This can be reduced by our smoothing method (ms)POAS as we will demonstrate in Section 6.5.

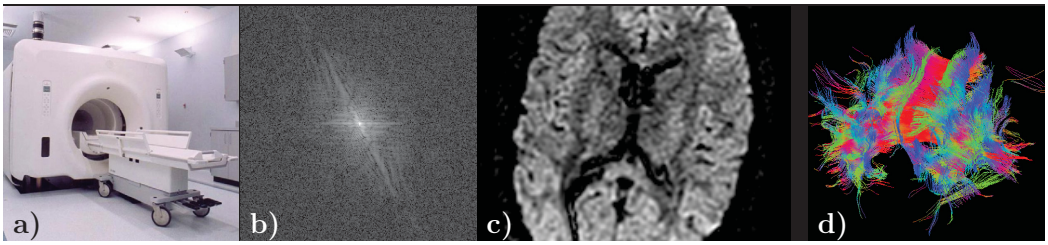

Figure 6.4: Graphical abstract of Section 6.1: a) Data acquisition (§ 6.1.1), using the PGSE sequence (§ 6.1.2) in a magnetic resonance tomograph; b) data in k-space as described in Eq. (6.1); for better visibility, we show the logarithmic absolute value of the complex signals; c) diffusion-weighted data reconstructed from k-space, see also § 6.1.4 for more details about possible artifacts and random noise; d) modeling and analysis of the data (§ 6.1.3), here represented by a fiber track.

6.2 Adaptive smoothing of diffusion-weighted MRI

Now we present the methods *position-orientation adaptive smoothing (POAS)* and *multi-shell POAS (msPOAS)* which are both based on the Propagation-Separation Approach. Additionally, we will give further details concerning the parameter choices. A flowchart of the msPOAS algorithm is shown in Figure 6.8 at the end of this section (page 101).

The Propagation-Separation Approach is a general approach for nonparametric estimation. For a specific application, some important preparations have to be made. First, an embedding of the data into the setting of Notation 2.1 is required. In (ms)POAS, we take advantage of the whole information of the diffusion-weighted MRI data, in position and orientation. Therefore, in Equation (6.2) (page 86), we presented a specific perspective on the data. The following notation summarizes the description of the dMRI data as introduced in § 6.1.2.

Notation 6.1 (Setting)**.** *Let $B \subseteq (0,\infty)$ be the set of applied b-values $b > 0$ and $\mathfrak{B} := |B| \in \mathbb{N}$ its cardinality. Additionally, let $V \subseteq \mathbb{R}^3$ denote the voxel space and $G_b \subseteq \mathbb{S}^2$ the set of applied diffusion-weighting gradient directions at b-value b, called the q-shell of b. For $b = 0$, we set $G_0 := \{\vec{0}\}$ and $S_0(\vec{v},\vec{0}) := S_0(\vec{v})$, where S_0 is the mean image of all acquired non-diffusion-weighted images. Then, the whole data set is given by the $\mathfrak{B}+1$ functions*

$$S_b : V \times G_b \ni m \mapsto S_b(m) \in \mathbb{R}, \;\; b \in B_0 := B \cup \{0\}.$$

We call measurements with only one value $b > 0$ single-shell *and data with more b-values* multi-shell*.*

POAS is based on this interpretation of the data for a single-shell, $\mathfrak{B} = 1$, while msPOAS considers the $\mathfrak{B}+1$ q-shells simultaneously, including the non-diffusion-weighted S_0-image. Pursuant to § 6.1.4, we assume the signals to follow a non-central chi-distribution.

Assumption A5. *In the setting of Notation 6.1, we assume, for every $m \in V \times G_b$ and all $b \in B_0$, that*

$$S_b(m)/\sigma \sim \chi_{2L'}(\theta_{b,m}) \quad \textit{with} \quad \theta_{b,m} = \sqrt{\mathbb{E}\left[S_b^2(m)\right]/\sigma^2 - 2L'},$$

where $\sigma^2 > 0$ denotes the variance of the observations, and $L' \in \mathbb{N}$ is the effective number of MR receiver coils.

Remark 6.2. Obviously, this violates Assumption A1 (page 10) since the non-central chi-distributions do not form an exponential family. Additionally, the expected value $\mathbb{E}[S_b(m)]$ does not equal the non-centrality parameter $\theta_{b,m}$, see Equation (6.7) (page 90). As a consequence, the theoretical results in Chapters 2, 3, and 4 do not apply to dMRI data, and a generalization of the proofs is impeded by the missing explicit formula of the Kullback-Leibler divergence in the case of non-central chi-distributions. However, the univariate examples in Section 5.3.5 give sufficient evidence, that the general behavior of the algorithm remains unchanged. We refer the reader to § 4.4.2 for further details concerning the application of the Propagation-Separation Approach in the case of a violated exponential family model.

As explained in Section 2.2, the Propagation-Separation Approach focuses on a suitable local definition of adaptive weights at a sequence of spatial scales. There, we need two distance functions, one on the design space and another one on the observation space. The resulting weights are then used in a weighted mean of the observations in order to estimate a parameter function which relates to the locally varying expected value. In order to accelerate and stabilize the estimation procedure, we again concentrate on the simplest version of the algorithm, which assumes a local constant model. For dMRI data, to some extent, this can be justified by visual inspection of the diffusion-weighted images shown in Figures 6.2 (page 86), where we observe similar values in the neighborhood of any voxel separated by sharp discontinuities, for example, at tissue borders. However, in gradient space just as for some spatial regions, this is an approximation of the more appropriate locally polynomial model only. Possible consequences of a violated structural assumption are therefore considered in § 6.5.3 and § 6.6.1.

6.2.1 The multi-shell POAS procedure

It follows from the measurement process that the diffusion characteristics will be comparable on all q-shells. More precisely, in each voxel, the spherical directions of observed signal extrema will coincide for all shells, while the signal size decreases with increasing b-values. Therefore, for each diffusion gradient direction, we will construct a vector of observations with varying b-values, including the non-diffusion-weighted image with $b = 0$. As the gradient schemes $G_b \subseteq \mathbb{S}^2$ do not necessarily coincide for all b-values $b > 0$, we will fill in missing data values in this vector description, using spherical interpolation. This approach allows the use of the information of all q-shells for adaptation and consequently improves efficiency as illustrated in Section 6.5. After presenting the spherical interpolation, we will adjust the statistical penalty of the Propagation-Separation Approach, where the impact of the previously applied interpolation will be taken into account. Then, we will introduce the algorithm of the msPOAS procedure, and we will briefly justify the proposed estimator.

Description of the data and spherical interpolation

In the case of identical gradient schemes $G_b = G_{b'}$, for all applied b-values $b, b' \in B$, the measured data can be easily arranged in a vector

$$S(\vec{v},\vec{g}) := (S_0(\vec{v}), S_{b_1}(\vec{v},\vec{g}), ..., S_{b_{\mathfrak{B}}}(\vec{v},\vec{g}))^T \in \mathbb{R}^{\mathfrak{B}+1} \quad \text{for all } (\vec{v},\vec{g}) \in V \times G_b. \tag{6.8}$$

This vector will not be complete if there are b-values $b, b' \in B$ with $G_b \neq G_{b'}$. Then, for every $\vec{g} \in G_{b'} \setminus G_b$, we use spherical interpolation to fill the missing value $S_b(\vec{v},\vec{g})$. The resulting data structure is visualized in Figure 6.5.

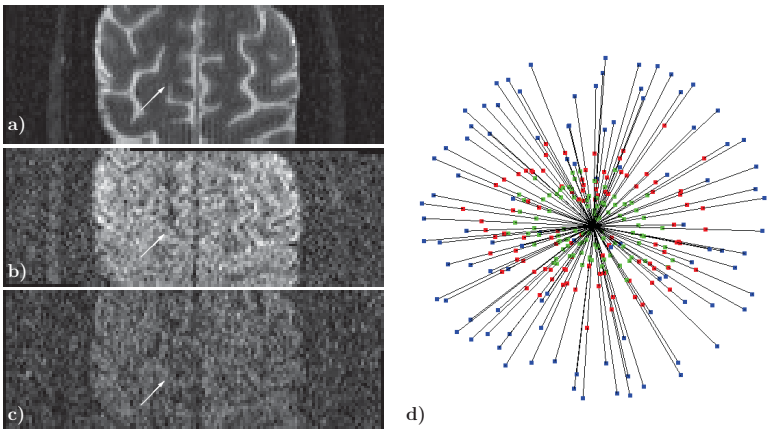

Figure 6.5: Visualization of the multi-shell data described in § 6.5.1. a) An arbitrarily selected slice of the non-diffusion-weighted data S_0. (b+c) Same slice taken with some arbitrarily selected diffusion-weighting gradient $\vec{g} \in \mathbb{S}^2$ at $b = 800\,\mathrm{s/mm^2}$ and $b = 2000\,\mathrm{s/mm^2}$, respectively. The intensity of (b+c) has been up-weighted to make (a),(b), and (c) visually feasible at once. d) Data within a single voxel $\vec{v} \in \mathbb{R}^3$ (see arrow) as a three-dimensional plot for all measured diffusion gradients at $b = 800\,\mathrm{s/mm^2}$ (red) and $b = 2000\,\mathrm{s/mm^2}$ (green). Additionally, the non-diffusion-weighted value $S_0(\vec{v})$ (blue) is repeatedly shown on each gradient, leading to the vector description $\{S(\vec{v},\vec{g})\}_{\vec{g}\in G} \subseteq \mathbb{R}^3$ in Equation (6.8). The distance of the points to the center of the sphere equals the corresponding signal value.

Let $b \in B$ be fixed and $\vec{g} \in G_{b'} \setminus G_b$ for some $b' \in B$. Then, we search for a triple of measured gradients $\{\mathfrak{g}^{(l)}_{(b,\vec{g})}\}_{l=1}^3 \subseteq G_b$ which span a spherical triangle which contains $\vec{g}$ with minimal total angular distance, that is $\vec{g} \in \triangle(\mathfrak{g}^{(1)}_{(b,\vec{g})}, \mathfrak{g}^{(2)}_{(b,\vec{g})}, \mathfrak{g}^{(3)}_{(b,\vec{g})})$ and

$$\sum_{l=1}^{3} \arccos(\langle \vec{g}, \mathfrak{g}^{(l)}_{(b,\vec{g})} \rangle) = \min \left\{ \sum_{l=1}^{3} \arccos(\langle \vec{g}, \vec{g}_l \rangle) : \vec{g}_l \in G_b \text{ for } l = 1,2,3 \right\},$$

where $\langle \cdot, \cdot \rangle$ denotes the Euclidean scalar product. The spherical Bary-coordinates $\{c^{(l)}_{(b,\vec{g})}\}_{l=1}^3$ of $\vec{g} \in \triangle(\mathfrak{g}^{(1)}_{(b,\vec{g})}, \mathfrak{g}^{(2)}_{(b,\vec{g})}, \mathfrak{g}^{(3)}_{(b,\vec{g})})$ are given as

$$c^{(l)}_{(b,\vec{g})} := \frac{\mathrm{area}(\triangle(\vec{g}, \mathfrak{g}^{(l_1)}_{(b,\vec{g})}, \mathfrak{g}^{(l_2)}_{(b,\vec{g})}))}{\mathrm{area}(\triangle(\mathfrak{g}^{(1)}_{(b,\vec{g})}, \mathfrak{g}^{(2)}_{(b,\vec{g})}, \mathfrak{g}^{(3)}_{(b,\vec{g})}))}, \quad \text{where } l_1, l_2 \in \{1,2,3\} \setminus \{l\}, l_1 \neq l_2, \tag{6.9}$$

see Carfora [2007, 3.1(c)]. The missing value $S_b(\vec{v},\vec{g})$ can then be generated by the linear spherical interpolation

$$\mathcal{S}_b(\vec{v},\vec{g}) := \sum_{l=1}^{3} c^{(l)}_{(b,\vec{g})} S_b(\vec{v}, \mathfrak{g}^{(l)}_{(b,\vec{g})}).$$

For a visualization of the Bary coordinates and the corresponding spherical triangles, we refer the reader to Figure 6.6.

For smoothing of the non-diffusion-weighted image S_0, we create a vector $S \in \mathbb{R}^{\mathfrak{B}+1}$, see Equation (6.8), with gradient $\vec{g} := \vec{0}$. Recall that $G_0 = \{\vec{0}\}$, but $\vec{0} \notin G_b$ for all $b > 0$. We fill the missing values with the mean value of the signals on each shell

$$\mathcal{S}_b(\vec{v},\vec{0}) := |G_b|^{-1} \sum_{\vec{g}\in G_b} S_b(\vec{v},\vec{g}), \qquad b \in B,$$

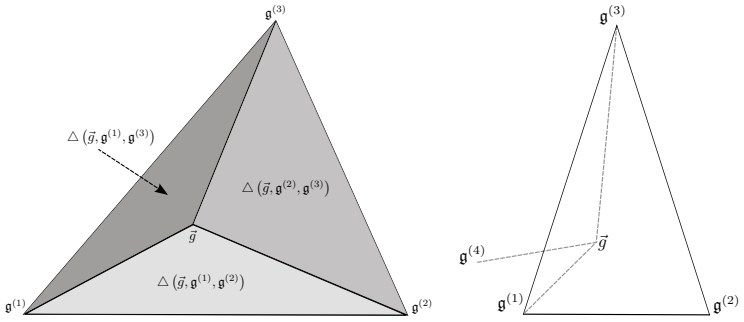

Figure 6.6: Visualization of the spherical interpolation: For each $b \in B$ and every gradient $\vec{g} \in G_{b'} \setminus G_b$, $b' \in B$, there is a spherical triangle formed by the gradients $\mathfrak{g}^{(1)}_{(b,\vec{g})}, \mathfrak{g}^{(2)}_{(b,\vec{g})}, \mathfrak{g}^{(3)}_{(b,\vec{g})} \in G_b$. The interpolation weights $c^{(l)}_{(b,\vec{g})}$, $l = 1,2,3$, in Equation (6.9) are determined by the respective proportion of the surface area of the partial triangles $\triangle(\vec{g}, \mathfrak{g}^{(l_1)}_{(b,\vec{g})}, \mathfrak{g}^{(l_2)}_{(b,\vec{g})})$ with $l_1, l_2 \neq l$. On the right, we illustrate the case where the triangle with minimal total angular distance does not contain the gradient $\vec{g}$. For simplicity, the figure shows planar triangles instead of spherical triangles.

where $|G_b| \in \mathbb{N}$ is the number of gradients measured on the shell with b-value b. By averaging all signals of the respective shell, the impact of the corresponding diffusion-weighting gradients is removed.

Finally, we consider the union $G := \bigcup_{b \in B_0} G_b \subseteq \mathbb{S}^2 \cup \{\vec{0}\}$ of all gradient directions measured for any b-value $b \in B_0$, and we define the desired vector function

$$\mathcal{S} : V \times G \ni (\vec{v}, \vec{g}) \mapsto (S_0(\vec{v}), \mathcal{S}_{b_1}(\vec{v}, \vec{g}), ..., \mathcal{S}_{b_{\mathfrak{B}}}(\vec{v}, \vec{g}))^T \in \mathbb{R}^{\mathfrak{B}+1}, \tag{6.10}$$

where, for every $b > 0$, we set

$$\mathcal{S}_b(\vec{v}, \vec{g}) = \begin{cases} S_b(\vec{v}, \vec{g}) & \text{if } \vec{g} \in G_b, \\ \sum_{l=1}^{3} c^{(l)}_{(b,\vec{g})} S_b(\vec{v}, \mathfrak{g}^{(l)}_{(b,\vec{g})}) & \text{if } \vec{g} \in G_{b'} \setminus G_b, b' > 0, \\ |G_b|^{-1} \sum_{\vec{g}_* \in G_b} S_b(\vec{v}, \vec{g}_*) & \text{if } \vec{g} \in G_0. \end{cases} \tag{6.11}$$

Hence, we distinguish, for b-values $b > 0$, between measured values $\mathcal{S}_b(\vec{v}, \vec{g})$ with $\vec{g} \in G_b$, interpolated values with $\vec{g} \in G \setminus G_b$, and the mean signal for $\vec{g} \in G_0$.

In the following, we do no longer distinguish between the original signals S and the interpolated values $\mathcal{S}$, denoting both by S. Additionally, we apply the interpolation in Equation (6.11) to the adaptive estimates $\tilde{S}^{(k)}_b$, defined in Equation (6.18) (page 97) below, replacing S_b by $\tilde{S}^{(k)}_b$ in the respective formulas. Once again, we denote the resulting values $\tilde{\mathcal{S}}^{(k)}_b$ by $\tilde{S}^{(k)}_b$.

Extension of the statistical penalty

For the adaptive estimator of our method, we take advantage of the whole information contained in the data vector (6.10). This requires an appropriate extension of the statistical penalty in Algorithm 2 (page 18).

In Chapter 2, the statistical penalty was based on the Kullback-Leibler divergence due to its relation to the fitted log-likelihood in the case of a local exponential family model. Although this model is violated for dMRI data, we continue to use the Kullback-Leibler divergence. This

is motivated by the following reasons. First, this divergence is a widely-used measure for the difference between two probability distributions. Second, the approximation at the end of this section provides an explicit expression which accelerates the computation considerably. Third, under statistical independence, the Kullback-Leibler divergence can be easily extended to multivariate densities as we will see in Equation (6.12). In Chapter 3, the Kullback-Leibler divergence provided for several probability distributions the very useful invariance of the propagation condition with respect to the fixed parameter. Therefore, we hope to get at most a weak variability for non-central chi-distributions. This would allow a choice of the adaptation bandwidth independent of the data at hand. Finally, in Section 5.3.5, the univariate examples of the Propagation-Separation Approach for non-central chi-distributed observations behaved as expected, using the Kullback-Leibler divergence for the definition of the statistical penalty. Nevertheless, it could be a nice piece of future research to evaluate the impact of the Kullback-Leibler divergence on the (ms)POAS method and possible consequences of a replacement by another f-divergence or by a possibly asymmetric distance function on the observation space. We introduce an abbreviatory notion for the Kullback-Leibler divergence between two non-central chi-distributions. This compensates the bias between the expected value $\mathbb{E}[S_b(m)]$, which we estimate by msPOAS, and the target parameter $\theta_{b,m}$, which determines the corresponding probability distribution, see Assumption A5 (page 91).

Notation 6.3. *For every pair of expected values $\eta_1, \eta_2 > 0$, we set*

$$\mathcal{KL}(\eta_1, \eta_2) := \mathcal{KL}\left(\mathbb{P}_{(\mu^{-1}(\eta_1))}, \mathbb{P}_{(\mu^{-1}(\eta_2))}\right),$$

where $\mathbb{P}_\theta = \chi_{2L'}(\theta)$ and the function μ^{-1} is defined via Equation (6.7) *(page 90) for all $\eta > \sqrt{2}\Gamma(L'+1/2)/\Gamma(L')$ and $\mu^{-1}(\eta) = 0$, else.*

Below, we will replace, with a slight abuse of notation, the unknown standardized expected value $\eta := \mathbb{E}S_b(m)/\sigma$ by its estimate $\tilde{\eta} := \tilde{S}_b^{(k)}(m)/\hat{\sigma}$, using the adaptive estimator $\tilde{S}_b^{(k)}(m)$ in Equation (6.18) below and the estimated standard deviation of the observations $\hat{\sigma} > 0$. Here, the notion $\tilde{S}_b^{(k)}(m)$ refers to the realization $\tilde{S}_b^{(k)}(m) \in \mathbb{R}$, but not to the random variable $\tilde{S}_b^{(k)}(m)$.

Next we observe that the measurement errors for different b-values are statistically independent since each q-shell is measured on its own. Consequently, the joint probability density $f_{S/\sigma}$ of the standardized random vector S in Equation (6.10) equals the product of the corresponding marginal densities $f_{S_b/\sigma}$, $b \in B_0$, where $S := \mathcal{S}$ is as in Equation (6.10). Hence, it holds

$$f_{S/\sigma}(u) = \prod_{l=0}^{\mathfrak{B}} f_{S_{b_l}/\sigma}(u_l), \quad u = (u_0, ..., u_{\mathfrak{B}})^T \in \mathbb{R}^{\mathfrak{B}+1}, b_0 := 0.$$

For the Kullback-Leibler divergence between $f_{\tilde{S}^{(k-1)}(m)/\hat{\sigma}}$ and $f_{\tilde{S}^{(k-1)}(n)/\hat{\sigma}}$, this yields

$$\mathcal{KL}\left(\frac{\tilde{S}^{(k-1)}(m)}{\hat{\sigma}}, \frac{\tilde{S}^{(k-1)}(n)}{\hat{\sigma}}\right) = \sum_{b \in B_0} \mathcal{KL}\left(\frac{\tilde{S}_b^{(k-1)}(m)}{\hat{\sigma}}, \frac{\tilde{S}_b^{(k-1)}(n)}{\hat{\sigma}}\right), \tag{6.12}$$

where we use Notation 6.3.

Therefore, we redefine the statistical penalty as

$$s_{mn}^{(k)} := \sum_{b \in B_0} \tilde{N}_{m,b}^{(k-1)} \mathcal{KL}\left(\frac{\tilde{S}_b^{(k-1)}(m)}{\hat{\sigma}}, \frac{\tilde{S}_b^{(k-1)}(n)}{\hat{\sigma}}\right). \tag{6.13}$$

Here, $\tilde{N}_{m,b}^{(k-1)}$ relates, as before, to the achieved variance reduction, using the adaptive weights $\{\tilde{w}_{mn}^{(k)}\}_{n\in V\times G_b}$ in Equation (6.16) (page 97). In order to compensate the impact of the spherical interpolation, we distinguish the same cases as for the interpolation formula (6.11). Additionally, for $b=0$, we down-weight the influence of the S_0-images. Otherwise, the S_0-images would dominate the adaptation in an undesirable manner due to the already achieved variance reduction by taking the mean of all non-diffusion-weighted images. Hence, for $m=(\vec{v}_m,\vec{g}_m)$, we set

$$\tilde{N}_{m,b}^{(k)} = \begin{cases} \max\limits_{k'\le k}\left(\sum\limits_{n\in V\times G_b}\tilde{w}_{mn}^{(k')}\right) & \text{if } b>0\wedge\vec{g}_m\in G_b,\\ \max\limits_{k'\le k}\left(\sum_{l=1}^{3} c_{b,\vec{g}_m}^{(l)}/\tilde{N}_{(\vec{v}_m,\mathfrak{g}_{b,\vec{g}_m}^{(l)}),b}^{(k')}\right)^{-1} & \text{if } b>0,\vec{g}_m\in G_{b'},0<b'\neq b,\\ \max\limits_{k'\le k}\left(\sum_{\vec{g}\in G_b} 1/\tilde{N}_{(\vec{v}_m,\vec{g}),b}^{(k')}\right)^{-1}\cdot|G_b| & \text{if } b>0\wedge\vec{g}_m\in G_0,\\ \max\limits_{k'\le k}\left(\sum\limits_{n\in V\times G_0}\tilde{w}_{mn}^{(k')}\right)\cdot|S_0|^{-1} & \text{if } b=0, \end{cases} \tag{6.14}$$

where $|S_0|$ is the number of acquired non-diffusion-weighted images which form the mean image S_0. Note that we consider the maximal variance reduction $\tilde{N}_{m,b}^{(k)} := \max\limits_{k'\le k}(\cdot)$ until step k in order to preserve the already achieved adaptation quality as proposed in § 5.4.1.

The algorithm

Finally, we summarize the algorithm for multi-shell position-orientation adaptive smoothing (msPOAS). It differs from the simplified Propagation-Separation Approach in Algorithm 2 (page 18) by the definition of the statistical penalty, which requires the interpolation of missing values as described above. Due to the interpolation, the resulting data $\{S_b(\vec{v},\vec{g})\}_{\vec{g}\in G}$ are statistically dependent. Therefore, in the estimator of msPOAS, we only include the measured signals $\{S_b(\vec{v},\vec{g})\}_{\vec{g}\in G_b}$. The interpolated values $\{S_b(\vec{v},\vec{g})\}_{\vec{g}\in G\setminus G_b}$ are solely used to determine the adaptive weights.

The kernel functions K_{loc}, K_{ad}, the adaptation bandwidth $\lambda>0$, and the sequence of location bandwidths $\{h^{(k)}\}_{k=0}^{k^*}$ can be chosen as in Section 2.2. For further details concerning the parameter choices for dMRI data, we refer the reader to § 6.2.2. There, we will introduce an additional dependence of $\{h^{(k)}\}_{k=0}^{k^*}$ on the respective gradient $\vec{g}\in G_b\subseteq\mathbb{S}^2$ to compensate possible inhomogeneities of the gradient schemes.

As before, we use for initialization of the algorithm a non-adaptive estimator. Basically, the non-adaptive weights could be determined by any distance function δ on $\mathbb{R}^3\times\mathbb{S}^2$. Here, we just assume that an appropriate distance has been fixed, two specific examples will be given in Section 6.4. In dMRI, the voxel and the gradient space are not intrinsically related by the measurement process. More precisely, the physical measurement units differ, and there is no natural relation between the spatial distance and the applied magnetic field gradient directions. Therefore, we introduce an additional parameter κ which allows a balance between the spatial and the spherical part. Hence, we consider a family of distance functions $\delta:=\{\delta_\kappa\}_{\kappa>0}$, where the choice of κ will be specified in § 6.2.2. For $b=0$, the following notation reduces any distance δ_κ on $\mathbb{R}^3\times\mathbb{S}^2$ to the corresponding distance on the voxel space. This enables the application of δ_κ on the S_0-image.

Notation 6.4. *Let K_{loc} and $h^{(k)}$ be as in Notation 2.14 and let $\delta_\kappa:(\mathbb{R}^3\times\mathbb{S}^2)\times(\mathbb{R}^3\times\mathbb{S}^2)\to\mathbb{R}$ denote a distance function with balancing parameter $\kappa>0$. For every $m,n\in\mathbb{R}^3\times\mathbb{S}^2$ we*

consider the non-adaptive weights $\overline{w}_{mn}^{(k)} := K_{\mathrm{loc}}(\delta_\kappa(m,n)/h^{(k)})$*. Then, for* $m,n \in V \times G_0$*, we set* $\overline{w}_{mn}^{(k)} := \overline{w}_{m_{\vec{g}} n_{\vec{g}}}^{(k)}$*, where* $m = (\vec{v}_m, \vec{0})$ *implies* $m_{\vec{g}} = (\vec{v}_m, \vec{g})$ *and* $n = (\vec{v}_n, \vec{0})$ *yields* $n_{\vec{g}} = (\vec{v}_n, \vec{g})$ *for some arbitrary but fixed gradient* $\vec{g} \in G$*.*

Algorithm 3 (Multi-shell position-orientation adaptive smoothing)**.**

1. ***Input parameters:*** *Sequence of location bandwidths* $\{h^{(k)}\}_{k=0}^{k^*}$*, balancing parameter* κ*, adaptation bandwidth* λ*.*
2. ***Initialization:*** *Set* $k := 0$*, and apply, for all* $m,n \in V \times G_b$ *and every* $b \in B_0$*, Equations* (6.16)*,* (6.17)*, and* (6.18) *with* $s_{mn}^{(0)} := 0$*. Then, increase* k *by* 1*.*
3. ***Iteration:*** *For each* $b \in B_0$ *and* $m := (\vec{v}_m, \vec{g}_m) \in V \times G_b$*, do the following. Interpolate the missing values of* $\tilde{S}_{b'}^{(k-1)}(n)$ *and* $\tilde{N}_{n,b'}^{(k)}$*,* $b' \in B \setminus \{b\}$ *and* $n \in V \times G_b$*, according to Equations* (6.11) *and* (6.14)*. Then, calculate the statistical penalty*

$$s_{mn}^{(k)} := \sum_{b \in B_0} \tilde{N}_{m,b}^{(k-1)} \mathcal{KL}\left(\frac{\tilde{S}_b^{(k-1)}(m)}{\hat{\sigma}}, \frac{\tilde{S}_b^{(k-1)}(n)}{\hat{\sigma}} \right), \quad n \in V \times G_b, \tag{6.15}$$

the adaptive weights

$$\tilde{w}_{mn}^{(k)} := K_{\mathrm{loc}}\left(\delta_\kappa(m,n)/h^{(k)}\right) \cdot K_{\mathrm{ad}}\left(s_{mn}^{(k)}/\lambda\right), \quad n \in V \times G_b, \tag{6.16}$$

the sum of the adaptive weights

$$\tilde{N}_{m,b}^{(k)} := \max_{k' \le k} \left(\sum_{n \in V \times G_b} \tilde{w}_{mn}^{(k')} \right), \tag{6.17}$$

and the adaptive estimator

$$\tilde{S}_b^{(k)}(m) := \sum_{n \in V \times G_b} \tilde{w}_{mn}^{(k)} S_b(n) / \tilde{N}_{m,b}^{(k)}. \tag{6.18}$$

4. ***Stopping:*** *Stop if* $k = k^*$*, and return* $\tilde{S}_b^{(k^*)}(m)$ *for each* $b \in B_0$ *and all* $m \in V \times G_b$*, else set* $k := k+1$*.*

Definition 6.5. *The non-adaptive estimator* $\overline{S}_b^{(k)}(m)$ *corresponding to msPOAS is defined by Equation* (6.18)*, setting* $s_{mn}^{(k)} := 0$ *in Equation* (6.16)*.*

Justification of the estimator

In Assumption A5 (page 91), we described the probability distribution of the standardized dMRI data by a non-central chi-distribution with $2L'$ degrees of freedom and varying non-centrality parameter θ, that is $S_b(m)/\sigma \sim \chi_{2L'}(\theta_{b,m})$. Obviously, this violates Assumption A1 (page 10) and hence the setting where the Propagation-Separation Approach was motivated. The assumed probability distribution of the data effects two parts of the algorithm, the definition of the estimator and the Kullback-Leibler divergence in the statistical penalty.

For the estimator, we should ensure that it remains in the same distribution class as the observations. Basically, for non-central chi-distributed observations, this requires the application of a quadratic mean in place of the arithmetic mean in Equation (6.18). Instead, we introduce a Gaussian approximation of the non-central chi-distribution in Assumption A5. This justifies the

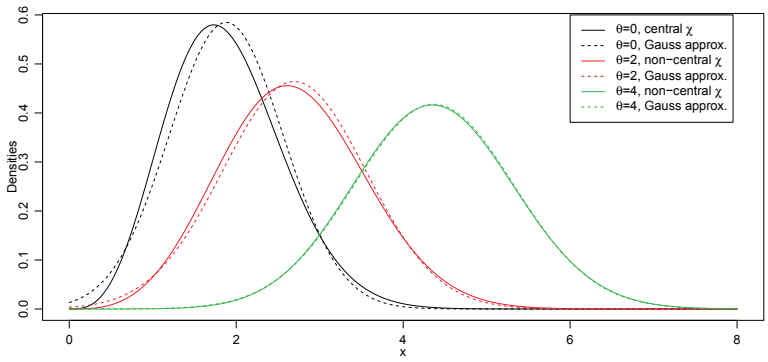

Figure 6.7: Densities of non-central chi-distributions with 4 degrees of freedom and the corresponding approximating Gaussian densities with coinciding first and second moments.

usage of an arithmetic mean to some extent since the weighted mean of Gaussian distributed observations is Gaussian.

We approximate the non-central chi-distribution $\chi_{2L'}(\theta_{b,m})$ by a Gaussian distribution $\mathcal{N}(\eta, v)$, where the choices

$$\eta(\theta, 2L') := \sqrt{\frac{\pi}{2}}\, \mathbb{L}_{1/2}^{(L'-1)}\left(-\frac{\theta^2}{2}\right) \quad \text{and} \quad v(\theta, 2L') := 2L' + \theta^2 - \eta^2(\theta, 2L')$$

ensure that the first and second moments of the original and the approximating distribution coincide, see Equation (6.7) (page 90). Then, the associated densities show a very similar behavior as illustrated in Figure 6.7 for $L' = 2$, where we compared various non-central chi-distributions with their corresponding Gaussian approximation. Other values of L' yield comparable results. We emphasize that this approximation still differs from the Gaussian regression model in Example 2.2 since the expectation and the variance of the approximating Gaussian distribution depend both on the target parameter θ.

Additionally, this approach allows an approximation of the Kullback-Leibler divergence between two non-central chi-distributions by

$$\widetilde{\mathcal{KL}}(\chi_{2L'}(\theta_1), \chi_{2L'}(\theta_2)) = \frac{[\eta(\theta_1, 2L') - \eta(\theta_2, 2L')]^2}{[v(\theta_1, 2L') + v(\theta_2, 2L'))]}, \tag{6.19}$$

which can be interpreted as the symmetrized Kullback-Leibler divergence between the approximating Gaussian distributions. This approximation improves and accelerates the results in [Becker et al., 2012, App. B], where a numerical approximation was used. In [Becker et al., 2013, Fig. 5], the relative and absolute error of the Gaussian approximation are given, where the exact Kullback-Leibler divergence is calculated via a numerical integration. This would be prohibitive in msPOAS due to the computational workload.

6.2.2 Parameter choices

In § 2.2.2, we discussed the impact and choices of the parameters in the Propagation-Separation Algorithm. As it turned out, most of them can be chosen in reasonable ranges without causing trouble if varied slightly. We now adapt this choices to the msPOAS procedure. Here, we concentrate on the adaptation bandwidth λ and the sequence of location bandwidths $\{h^{(k)}\}_{k=0}^{k^*}$.

Moreover, we discuss the additional parameter κ, which appears in the distance δ_κ on $\mathbb{R}^3 \times \mathbb{S}^2$. The maximal number of iteration k^* can be chosen as before, see § 2.2.2 and Section 5.4. We recall that the precise choice of the kernel functions K_{loc} and K_{ad} have minor impact only. In the implementation of (ms)POAS in the **R**-package **dti** by Tabelow and Polzehl [2013], the kernels in Equation (2.10) (page 19) were used.

The sequence $\{h^{(k)}\}_{k=0}^{k^*}$ should be increasing with $h^{(0)} > 0$. In the implementation of the method msPOAS, the following choice ensures a constant variance reduction of the non-adaptive estimate by 25%, up to boundary effects in the voxel space, see § 2.2.2 for comparison. This introduces a dependence of the b-value $b \in B_0$ and the gradient $\vec{g} \in G_b$ under consideration due to possible inhomogeneities within the respective gradient scheme G_b on the q-shell with b-value b. The bandwidths are measured in units of voxel counts with respect to the smallest of the three voxel extensions. Then, we fix some voxel $\vec{v}_m$ close to the center of the voxel space V, and we choose $h^{(0)} = 1$, which equals one unit of the shortest edge of the voxel. This restricts smoothing to the sphere and hence avoids blurring at structural borders in the voxel space in the initial iteration step $k = 0$. Subsequently, for every $b \in B_0$ and all $\vec{g}_m \in G_b$, we calculate the sequence of location bandwidths $\{h^{(k)}(b, \vec{g}_m)\}_{k=1}^{k^*}$ by numerically solving the equation

$$\sum_{n \in V \times G_b} \frac{(\overline{w}_{mn}^{(k-1)})^2}{(\overline{N}_{m,b}^{(k-1)})^2} = 1.25 \sum_{n \in V \times G_b} \frac{(\overline{w}_{mn}^{(k)})^2}{(\overline{N}_{m,b}^{(k)})^2}, \tag{6.20}$$

where $m = (\vec{v}_m, \vec{g_m})$. The sequences $\{h^{(k)}(b, \vec{g}_m)\}_{k=0}^{k^*}$ with $b \in B_0$ and $\vec{g}_m \in G_b$ are then used for all voxels $\vec{v} \in V$. We emphasize that the resulting non-adaptive weights $\overline{w}_{mn}^{(k)}$ depend on the respective b-value $b \in B_0$ via the applied location bandwidth $h^{(k)}(b, \vec{g}_m)$.

We assumed the distance $\delta_\kappa : (\mathbb{R}^3 \times \mathbb{S}^2) \times (\mathbb{R}^3 \times \mathbb{S}^2) \to \mathbb{R}$ to depend on an additional parameter $\kappa > 0$, which we introduced in order to balance between spatial and spherical smoothing in $V \subseteq \mathbb{R}^3$ and $G_b \subseteq \mathbb{S}^2$, $b \in B$. Smoothing on the sphere increases the maximally achievable variance reduction, and it ensures a stabilization of the estimates for the first steps of the method, although it introduces a bias. Therefore, we recommend to choose κ such that the amount of smoothing on the sphere is constant in the considered voxel $\vec{v}_m$ for all $k \in \{0, ..., k^*\}$. This choice compensates the spherical bias during iteration since the statistical penalty becomes stricter while the location bandwidths $\{h^{(k)}\}_{k=0}^{k^*}$ are increasing. This leads to a decreasing amount of spherical smoothing in comparison with the amount of spatial smoothing in the voxel space. The precise choice of κ should balance between the amount of required stabilization in the first steps and the possibility of bias correction. The latter depends via the statistical penalty on the unknown sample size of the homogeneous regions in the voxel space. The former requires κ to be the larger the lower the signal-to-noise ratio and the smaller the larger the mean number of applied gradient directions per shell, given as $N_g := \sum_{b \in B} |G_b| / \mathfrak{B}$.

For our specific distances in Notation 6.63 and 6.64, the above arguments lead to the choice $\kappa(h^{(k)}) := \kappa_0 / h^{(k)}$ with $\kappa_0 > 0$ fixed. Then, for $m = (\vec{v}_m, \vec{g}_m)$ and $n = (\vec{v}_n, \vec{g}_n)$ with $\vec{v}_m = \vec{v}_n$, we get $\overline{w}_{mn}^{(k)} = K_{\mathrm{loc}}\left(d(\vec{g}_m, \vec{g}_n)/\kappa_0\right)$ for all $k \in \{0, ..., k^*\}$, where the definition of the spherical distance $d(\vec{g}_m, \vec{g}_n)$ follows from Notation 6.63 or 6.64, respectively. For δ_κ as in Notations 6.64, the quantity $N_g(1 - \cos(\kappa_0))$ determines the mean number of gradient directions with positive weights for $h^{(0)} = 1$. We suggest to select κ_0 such that this number is between 5 and 10.

Finally, the adaptation bandwidth λ determines the amount of adaptation. Its choice follows from the propagation condition in § 3.1.1. We aim to ensure, for the adaptive estimator of non-central chi-distributed observations, the same rate of convergence as shown for the non-adaptive

estimator in the case of an exponential family, see Theorem 2.13. Hence, for single-shell data with some fixed b-value $b \in B_0$, we simulate data

$$\{m, S_b(m)\}_{m \in V \times G_b} \subseteq \mathbb{R}, \quad S_b(m) \overset{\text{iid}}{\sim} \mathbb{P}_{\theta_b},$$

for some least favorable $\theta_b \in \Theta$, where $b \in B_0$ is fixed. In particular, by considering a worst case scenario, we avoid the dependence on the variance of the data, setting $\sigma^2 := 1$, as the variance only influences the range of the non-centrality parameters. Then, we choose the minimal value λ_b which satisfies the propagation condition in Definition 3.2 at the desired propagation level $\epsilon > 0$ for this fixed parameter θ_b, considering the transformed adaptive estimator

$$\tilde{\theta}_b^{(k)}(m, \lambda) := \begin{cases} \mu^{-1}\left(\tilde{S}_{m,b}^{(k)}\right) & \text{if } \tilde{S}_{m,b}^{(k)} > \sqrt{2}\Gamma(L' + 1/2)/\Gamma(L'), \\ 0, & \text{else,} \end{cases}$$

where the function μ is as in Equation (6.7) (page 90).

For multi-shell data, the same strategy was implemented with $\mathfrak{Z}_\lambda$ depending on the sum

$$\sum_{b \in B} \overline{N}_{m,b}^{(k)} \mathcal{KL}(\tilde{\theta}_b^{(k)}(m, \lambda), \theta_b) \quad \text{in place of} \quad \overline{N}_{m,b}^{(k)} \mathcal{KL}(\tilde{\theta}_b^{(k)}(m, \lambda), \theta_b), \tag{6.21}$$

where the missing values of $\tilde{\theta}_b^{(k)}(m, \lambda)$ and $\overline{N}_{m,b}^{(k)}$ are again constructed by the interpolation in Equations (6.11) and (6.14) (page 94 and 96). As the natural design spaces of the S_0-image and the diffusion-weighted images differ, we consider the corresponding propagation levels ϵ_0 and ϵ_b separately, while the algorithm still uses a coupling of all q-shells, including $b = 0$. Equation (6.21) is motivated by the equivalent modification of the statistical penalty presented in Equation (6.13) (page 95). Selection of λ is then done, using the gradient schemes and b-values at hand and specifying appropriate values of θ_b for the different shells. For instance, one may fix a non-centrality parameter θ_0 which is applied to the S_0-image. Then, the corresponding parameters θ_b for varying b-values $b > 0$ can be calculated by modeling a standard diffusion situation as usually observed in the human brain, leading to exponentially decreasing values. Here again, the adaptation bandwidth λ is chosen as the minimal value which satisfies the propagation condition at certain propagation levels ϵ_0 and ϵ_b on the S_0-image and the coupled diffusion-weighted images, respectively.

The implementation of this propagation condition in the **R**-package **dti** by Tabelow and Polzehl [2013] is based on Algorithm 3 (page 97) and the above choices of $\{h^{(k)}\}_{k=0}^{k^*}$ and κ. In Section 6.5, we will show by numerical simulations that the propagation level is quite robust against the fixed parameter θ_0, the number of diffusion-weighting gradients per q-shell, the effective number of receiver coils L', and the realization of the sample seed. Additionally, we will give precise choices of λ for the presented data sets. The mentioned invariances of the propagation level and the mutual effects of the adaptation bandwidth λ and the noise variance σ in the algorithm will be discussed in § 6.6.2 in more detail.

6.2.3 Special case: The single-shell POAS procedure

Before the development of the msPOAS procedure, we proposed a method for position-orientation adaptive smoothing (POAS) which is applied to each q-shell separately. Basically, POAS can be considered as a special case of the msPOAS algorithm. However, for msPOAS, we introduced several changes in order to improve and accelerate the procedure in comparison

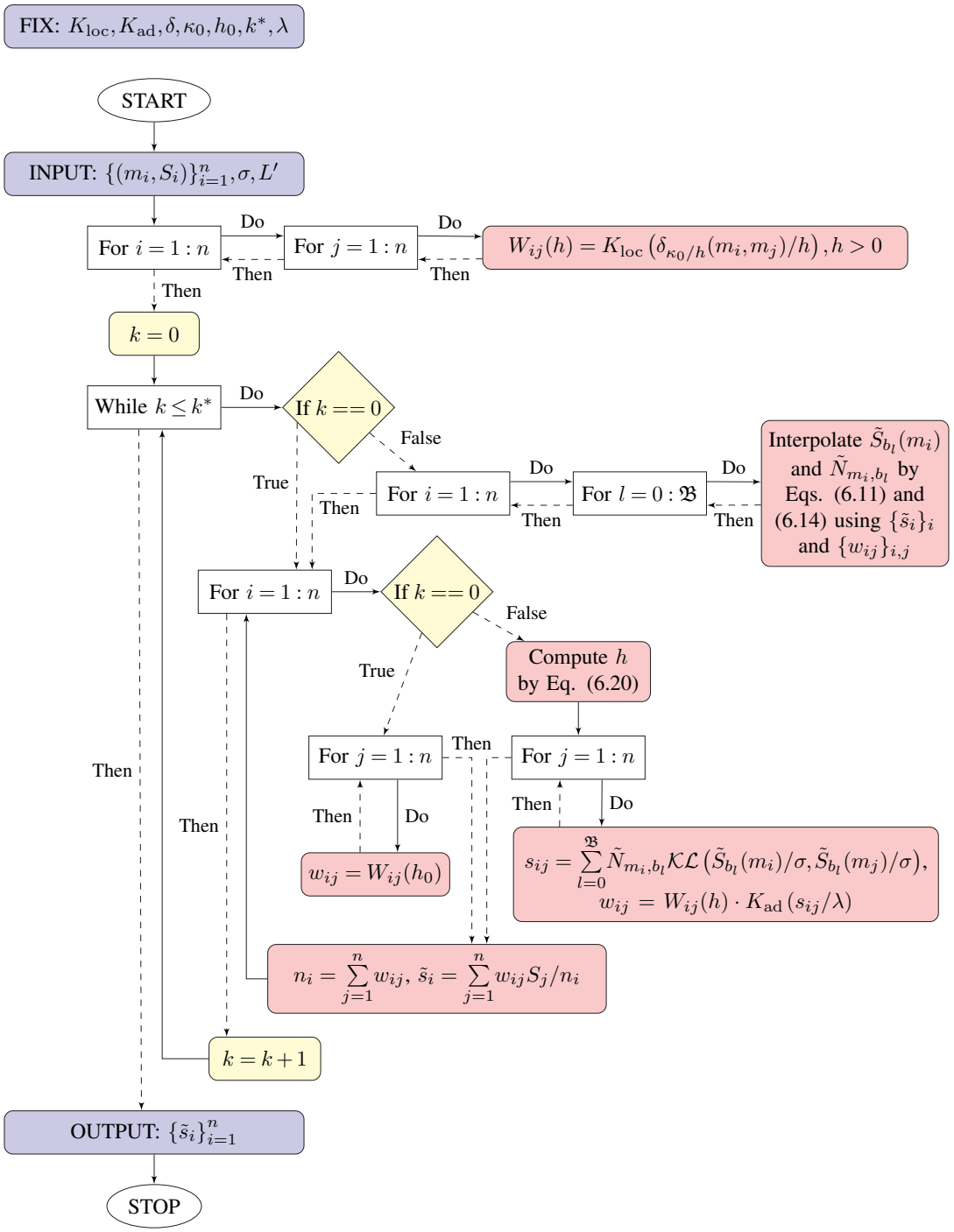

Figure 6.8: Flowchart of the msPOAS procedure as described in Algorithm 3 (page 97), using the parameter choices in § 6.2.2 and a simplified notation. Here, the dMRI data are given by the sequence $\{(m_i, S_i)\}_{i=1}^n$, where $n := \sum_{b \in B_0} |V \times G_b|$.

with the previously implemented POAS method. We refer the reader to § 6.6.3 for a detailed comparison of both methods.

In general, POAS is applied $\mathfrak{B}+1$ times, first to the $\mathfrak{B}=|B|$ diffusion-weighted images $\{S_b(m)\}_{m\in V\times G_b}$ with $b\in B$ fixed, and then to the non-diffusion-weighted image $\{S_0(\vec{v})\}_{\vec{v}\in V}$. The procedure follows Algorithm 3 (page 97), up to the following modifications.

1. In order to consider a single b-value $b\geq 0$, we set $B_0:=\{b\}$.
2. The statistical penalty in Equation (6.15) is replaced by
$$s_{mn}^{(k)}:=\tilde{N}_{m,b}^{(k-1)}\mathcal{KL}\left(\mathbb{P}(m),\mathbb{P}(n)\right),\qquad n\in V\times G_b, \tag{6.22}$$
with $\mathbb{P}(i):=\chi^2_{2L'}(\max[0,(\tilde{S}_b^{(k-1)}(i))^2/\hat{\sigma}^2-2L'])$, $i=m,n$.
3. The adaptive estimator in Equation (6.18) is replaced by a weighted quadratic mean,
$$\tilde{S}_b^{(k)}(m):=\sqrt{\sum_{n\in V\times G_b}\tilde{w}_{mn}^{(k)}S_b^2(n)/\tilde{N}_{m,b}^{(k)}}. \tag{6.23}$$

Note that $\tilde{S}_b^{(k)}(m)$ estimates the expectation $\mathbb{E}\left[S_b(m)^2\right]^{1/2}$ instead of the target parameter $\theta_{b,m}$ in Assumption A5 (page 91). This is compensated by a slight abuse of notation in the above statistical penalty, see Equation (6.22) and the subsequent comment. The quadratic mean in Equation (6.23) ensures that the estimator remains in the same distribution class as the signals, see the discussion in § 6.2.1.

6.3 Theoretical properties

In Notation 6.1 (page 91), we described diffusion-weighted magnetic resonance data as real valued functions on the measurement space $\mathbb{R}^3\times\mathbb{S}^2$. Such functions are known as orientation scores, and an embedding of $\mathbb{R}^3\times\mathbb{S}^2$ into the special Euclidean motion group $\mathrm{SE}(3)$ allows the application of well-known concepts from differential geometry and Lie group theory, see Duits and Franken [2011], Duits et al. [2011], and Franken [2008]. This approach can be considered as a natural lifting of an orientation score on $\mathbb{R}^3\times\mathbb{S}^2$ to an orientation score on $\mathrm{SE}(3)$. In order to ensure invertibility, this lifting requires a certain invariance property on $\mathrm{SE}(3)$. Then, the specific geometry of the measurement space can be examined, providing a general approach for the processing of oriented structures. This approach is a nice example of an abstract mathematical concept which provides very descriptive and useful tools for a real application, here medical imaging. In [Duits and Franken, 2011; Duits et al., 2011; Franken, 2008], the authors applied anisotropic diffusion on orientation scores. Here, we seek for a theoretical justification of the (ms)POAS method in Section 6.2. Additionally, the orientation score approach provides a natural distance on the measurement space, which we will introduce and analyse in Section 6.4.

We start with a brief summary about the special Euclidean motion group and motivate its definition as a semidirect product of two groups. Then, we will consider the parametrizations of $\mathbb{R}^3\times\mathbb{S}^2$ and $\mathrm{SE}(3)$, leading to the already mentioned embedding. Moreover, we will recall some basic results about three-dimensional orientation scores. In particular, this provides information about the type of operations which are recommendable for the processing of dMRI data without violating the observed structure. In Theorem 6.38, this will be used for the theoretical justification of (ms)POAS. The main results of this section are summarized in Figure 6.12 (page 112). We will skip the proofs of well-known differential geometric results as they are beyond the scope of this thesis. Instead, we will refer the reader to the respective literature.

6.3.1 The special Euclidean motion group

The special Euclidean motion group is a semidirect product of two groups. The following Definition and the proximate lemma can be found in [Hewitt and Ross, 1979, (2.6)].

Definition 6.6. *Let $G := N \times H = \{(n,h) : n \in N, h \in H\}$ denote the Cartesian product of the groups $(N, \cdot_N)$ and $(H, \cdot_H)$. Every homomorphism $\phi : H \to \mathrm{Aut}(N)$ defines a semidirect product $\cdot_\phi$ on G, where*

$$(n_1,h_1) \cdot_\phi (n_2,h_2) = (n_1 \cdot_N \phi(h_1)(n_2), h_1 \cdot_H h_2).$$

We denote the semidirect product of N and H by $N \rtimes H$.

Lemma 6.7. *For groups N and H, the Cartesian product $G := N \times H$ together with the semidirect product $\cdot_\phi$ forms the group $(N \rtimes H, \cdot_\phi)$. The identity is (e_N, e_H), where e_N and e_H denote the identities of N and H, respectively. The inverse of $(n,h) \in N \rtimes H$ is given as $(\phi(h^{-1})(n^{-1}), h^{-1})$.*

In contrast to the direct product, the two groups H and N have different roles in the semidirect product, as indicated by the notation $N \rtimes H$. Here, H acts via ϕ on N. Now we specify the abstract concept of a semidirect product for the well-known special Euclidean motion group.

Definition 6.8. *Let $(\mathbb{R}^d, +)$ denote the translation group and $(\mathrm{SO}(d), \cdot_{\mathrm{SO}(d)})$ the special orthogonal group, where $\cdot_{\mathrm{SO}(d)}$ denotes the matrix product and*

$$\mathrm{SO}(d) := \left\{ \boldsymbol{R} \in \mathrm{GL}(d,\mathbb{R}) : \boldsymbol{R}^T = \boldsymbol{R}^{-1}, \det(\boldsymbol{R}) = 1 \right\}$$

is a subgroup of the general linear group $GL(d,\mathbb{R}) := \{M \subseteq \mathbb{R}^{d\times d} : \det(M) \neq 0\}$. The special Euclidean motion group $\mathrm{SE}(d) := \mathbb{R}^d \rtimes \mathrm{SO}(d)$, $d \in \mathbb{N}$, is defined as a semidirect product of $(\mathbb{R}^d, +)$ and $(\mathrm{SO}(d), \cdot_{\mathrm{SO}(d)})$, induced by the homomorphism $\phi(\boldsymbol{R})(\vec{v}) := \boldsymbol{R}\vec{v}$.

Corollary 6.9. *For $g_1, g_2 \in \mathrm{SE}(d)$ with $g_i := (\vec{v}_i, \boldsymbol{R}_i)$, $i = 1,2$, the group operation $\cdot_{\mathrm{SE}(d)}$ of $\mathrm{SE}(d)$ is given by*

$$g_1 \cdot_{\mathrm{SE}(d)} g_2 := (\vec{v}_1 + \boldsymbol{R}_1\vec{v}_2, \boldsymbol{R}_1\boldsymbol{R}_2), \tag{6.24}$$

leading to $e = (\vec{0}, \boldsymbol{1})$ and $g^{-1} = (-\boldsymbol{R}^{-1}\vec{v}, \boldsymbol{R}^{-1})$ for $g = (\vec{v}, \boldsymbol{R})$.

Example 6.10. The special Euclidean motion group describes Euclidean motions, that is rotations and translations. First, we concentrate on the two-dimensional special Euclidean motion group $\mathrm{SE}(2)$. The special orthogonal group $\mathrm{SO}(2)$ describes rotations of two-dimensional vectors. The counter-clockwise rotation of $\vec{v} := (x,y)^T \in \mathbb{R}^2$ with rotation angle $\theta \in (-\pi, \pi]$ is defined by

$$\mathbf{R}_\theta \vec{v} = \begin{pmatrix} x\cos\theta - y\sin\theta \\ x\sin\theta + y\cos\theta \end{pmatrix}, \quad \text{where } \mathbf{R}_\theta := \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \in \mathrm{SO}(2), \tag{6.25}$$

see for instance Rossmann [2002, §2.1 Ex. 5]. An illustration is shown in Figure 6.9. The translation operator T is given by $T_{\vec{v}_2}(\vec{v}_1) := \vec{v}_1 + \vec{v}_2$ for all $\vec{v}_1, \vec{v}_2 \in \mathbb{R}^2$. Then, for $\vec{v}_1, \vec{v}_2 \in \mathbb{R}^2$ and $\mathbf{R}_\theta \in \mathrm{SO}(2)$, the rotation of $\vec{v}_1$ by θ with subsequent translation by $\vec{v}_2$ equals

$$T_{\vec{v}_2}(\mathbf{R}_\theta \vec{v}_1) = \mathbf{R}_\theta \vec{v}_1 + \vec{v}_2 = \begin{pmatrix} x_1\cos\theta - y_1\sin\theta + x_2 \\ x_1\sin\theta + y_1\cos\theta + y_2 \end{pmatrix},$$

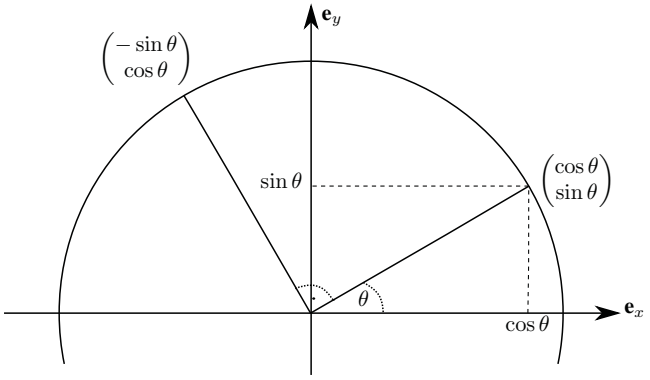

Figure 6.9: 2-dimensional rotation

where $\vec{v}_i = (x_i, y_i)^T$, $i = 1, 2$. The same operation can be described via the matrix representation of $g = (\vec{v}_2, \mathbf{R}_\theta) \in \mathrm{SE}(2)$ setting

$$g \equiv \mathbf{M}_g := \begin{pmatrix} \mathbf{R}_\theta & \vec{v}_2 \\ \vec{0}^T & 1 \end{pmatrix} \quad \text{with the result that} \quad \mathbf{M}_g \cdot \begin{pmatrix} \vec{v}_1 \\ 1 \end{pmatrix} = \begin{pmatrix} \mathbf{R}_\theta \vec{v}_1 + \vec{v}_2 \\ 1 \end{pmatrix}.$$

Then, for all $g_1, g_2 \in \mathrm{SE}(2)$ with $g_i = (\vec{v}_i, \mathbf{R}_{\theta_i}) \in \mathrm{SE}(2)$, $i = 1, 2$, matrix multiplication yields

$$\mathbf{M}_{g_2} \cdot \mathbf{M}_{g_1} = \begin{pmatrix} \mathbf{R}_{\theta_2}\mathbf{R}_{\theta_1} & \mathbf{R}_{\theta_2}\vec{v}_1 + \vec{v}_2 \\ \vec{0}^T & 1 \end{pmatrix} = \mathbf{M}_{(g_2 \cdot_{\mathrm{SE}(2)} g_1)},$$

and the group product of two elements $g_1, g_2 \in \mathrm{SE}(2)$ equals the subsequent application of the corresponding Euclidean motions. Generalization to $\mathrm{SE}(d)$ with $d \in \mathbb{N}$ is straight forward with

$$\mathrm{SE}(d) \ni (\vec{v}, \mathbf{R}) =: g \equiv \mathbf{M}_g := \begin{pmatrix} \mathbf{R} & \vec{v} \\ \vec{0} & 1 \end{pmatrix} \in \mathrm{GL}(d+1, \mathbb{R}) \tag{6.26}$$

and $\mathbf{M}_{(g_2 \cdot_{\mathrm{SE}(d)} g_1)} = \mathbf{M}_{g_2} \cdot \mathbf{M}_{g_1}$ for all $g_1, g_2 \in \mathrm{SE}(d)$.

6.3.2 Embedding of $\mathbb{R}^3 \times \mathbb{S}^2$ into $\mathrm{SE}(3)$

The embedding of $\mathbb{R}^3 \times \mathbb{S}^2$ into $\mathrm{SE}(3)$ follows from the corresponding parametrizations. Let us recall some related definitions, which can be found, for example, in [Rossmann, 2002, §4.1].

Definition 6.11. *A d-dimensional analytic manifold is a Hausdorff space M together with an atlas $\mathcal{A} := \{(M_i, \varphi_i)\}_i$ which satisfy the following conditions:*

1. *For every point $g \in M$, there is some chart $(M_i, \varphi_i) \in \mathcal{A}$ with $g \in M_i$ and $M = \bigcup_i M_i$.*
2. *Each chart $(M_i, \varphi_i) \in \mathcal{A}$ defines a bijective map $\varphi_i : M_i \to \varphi_i(M_i) \subseteq \mathbb{R}^d$, where $\varphi_i(M_i)$ is open in $\mathbb{R}^d$.*
3. *For all charts $(M_i, \varphi_i), (M_j, \varphi_j) \in \mathcal{A}$ with $M_i \cap M_j \neq \emptyset$, it holds that $\varphi_j(M_i \cap M_j) \subseteq \mathbb{R}^d$ is open, and the composition $\varphi_j \circ \varphi_i^{-1} \colon \varphi_i(M_i \cap M_j) \to \varphi_j(M_i \cap M_j)$ is analytic.*

The inverse functions $\{(\varphi_i^{-1})\}_i$ are known as parametrizations of M.

The definition of a d-dimensional manifold is illustrated in Figure 6.10.

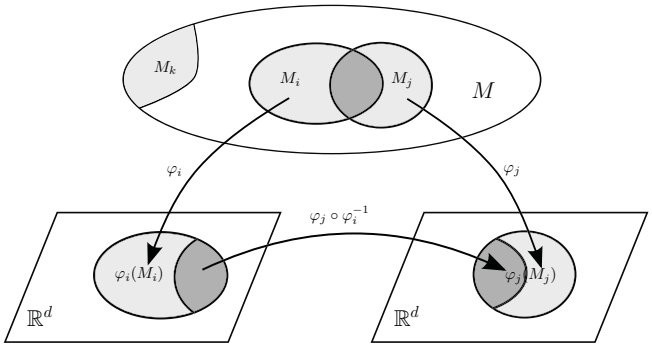

Figure 6.10: Schematic of a d-dimensional manifold

Example 6.12.

1. The 2-sphere $\mathbb{S}^2$ forms a 2-dimensional manifold [Sagle and Walde, 1973, §2.1 Ex. (5)].
2. The special Euclidean motion group $\mathrm{SE}(3)$ forms a 6-dimensional manifold. This follows directly from the results in [Rossmann, 2002, §4.1].

In the following, we do not consider the complete atlas of the respective manifold, but single parametrizations in a generalized sense without the restriction to open sets in the conditions (2) and (3) of Definition 6.11. We give the parametrizations of $\mathrm{SO}(3)$ and $\mathbb{S}^2$ as stated by Duits and Franken [2011].

Example 6.13. The special orthogonal group $\mathrm{SO}(3)$ describes rotations in $\mathbb{R}^3$. Hence, any element of $\mathrm{SO}(3)$ can be described by subsequent counter-clockwise rotations around the z-, y- and x-axis via

$$\mathbf{R}_\theta^{\vec{e}_x} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{pmatrix}, \mathbf{R}_\theta^{\vec{e}_y} = \begin{pmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{pmatrix}, \mathbf{R}_\theta^{\vec{e}_z} = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

where $\theta \in (-\pi, \pi]$. This leads to the parametrization

$$\mathbf{R}_{(\alpha,\beta,\gamma)} = \mathbf{R}_\gamma^{\vec{e}_x} \mathbf{R}_\beta^{\vec{e}_y} \mathbf{R}_\alpha^{\vec{e}_z} \quad \text{for } \beta \neq \pm\pi/2, \tag{6.27}$$

where

$$\mathbf{R}_{(\alpha,\beta,\gamma)} = \begin{pmatrix} \cos\alpha\cos\beta & -\sin\alpha\cos\beta & \sin\beta \\ \sin\alpha\cos\gamma + \cos\alpha\sin\beta\sin\gamma & \cos\alpha\cos\gamma - \sin\alpha\sin\beta\sin\gamma & -\cos\beta\sin\gamma \\ \sin\alpha\sin\gamma - \cos\alpha\sin\beta\cos\gamma & \sin\alpha\sin\beta\cos\gamma + \cos\alpha\sin\gamma & \cos\beta\cos\gamma \end{pmatrix}.$$

Another parametrization is given by the Euler angles. Here, counter-clockwise rotations around the z-, y-, and again the z-axis lead to

$$\check{\mathbf{R}}_{(\check{\alpha},\check{\beta},\check{\gamma})} = \mathbf{R}_{\check{\gamma}}^{\vec{e}_z} \mathbf{R}_{\check{\beta}}^{\vec{e}_y} \mathbf{R}_{\check{\alpha}}^{\vec{e}_z} \quad \text{for } \beta \neq 0, \pi.$$

A nice illustration of both parametrizations can be found in [Duits and Franken, 2011, Fig. 4]. The corresponding conversion formula can be deduced from

$$\mathbf{R}_\gamma^{\vec{e}_x} \mathbf{R}_\beta^{\vec{e}_y} \mathbf{R}_\alpha^{\vec{e}_z} \overset{!}{=} \check{\mathbf{R}}_{\check{\gamma}}^{\vec{e}_z} \check{\mathbf{R}}_{\check{\beta}}^{\vec{e}_y} \check{\mathbf{R}}_{\check{\alpha}}^{\vec{e}_z}, \qquad \beta, \check{\beta} \in (-\pi, \pi) \setminus \{-\pi/2, 0, \pi/2\}.$$

In the same manner, we parametrize the 2-sphere $\mathbb{S}^2$, which describes directions in $\mathbb{R}^3$. Every direction in $\mathbb{R}^3$ can be constructed via a 3D-rotation of a vector, for instance, the z-axis $\vec{e}_z = (0,0,1)^T$. Therefore, we may deduce the following parametrizations of $\mathbb{S}^2$, setting

$$\vec{u}(\beta,\gamma) := \mathbf{R}_{(\alpha,\beta,\gamma)}\vec{e}_z = \mathbf{R}_\gamma^{\vec{e}_x}\mathbf{R}_\beta^{\vec{e}_y}\vec{e}_z = \begin{pmatrix} \sin\beta \\ -\cos\beta\sin\gamma \\ \cos\beta\cos\gamma \end{pmatrix} \quad \text{for } \beta \neq \pm\pi/2$$

and

$$\check{\vec{u}}(\check{\beta},\check{\gamma}) := \check{\mathbf{R}}_{(\check{\alpha},\check{\beta},\check{\gamma})}\vec{e}_z = \check{\mathbf{R}}_{\check{\gamma}}^{\vec{e}_z}\check{\mathbf{R}}_{\check{\beta}}^{\vec{e}_y}\vec{e}_z = \begin{pmatrix} \sin\check{\beta}\cos\check{\gamma} \\ \sin\check{\beta}\sin\check{\gamma} \\ \cos\check{\beta} \end{pmatrix} \quad \text{for } \check{\beta} \neq 0,\pi.$$

In this study, we concentrate on the parametrization of $\mathrm{SO}(3)$ in Equation (6.27) since this is well-defined at the identity.

For the given parametrizations, all elements of the 2-sphere $\mathbb{S}^2$ are invariant with respect to the rotation angle α of the corresponding $\mathrm{SO}(3)$-parametrizations. This provides an embedding of $\mathbb{S}^2$ into $\mathrm{SO}(3)$ and hence of $\mathbb{R}^3 \times \mathbb{S}^2$ into $\mathrm{SE}(3) = \mathbb{R}^3 \rtimes \mathrm{SO}(3)$. For this purpose, we introduce left coset spaces.

Definition 6.14. *Let $(G,\cdot)$ be a group and $(H,\cdot)$ a subgroup of G. The left coset or orbit gH is defined to be the set*

$$gH := [g]_H := \{g \cdot h : h \in H\}, \qquad g \in G,$$

leading to the left coset space $G/H := \{gH : g \in G\}$.

This definition and the following lemma have been stated, for instance, in [Franken, 2008, §7.3.2.] and in a more general setting in [Rossmann, 2002, §4.2].

Lemma 6.15. *The left coset space G/H defines a partition of G via the equivalence relation $\sim$, where $g_1 \sim g_2$ if and only if $g_1 H = g_2 H$.*

Let us consider the same example as Franken [2008, §7.3.2.].

Notation 6.16. *Let $\mathrm{stab}(\vec{e}_z) \subseteq \mathrm{SO}(3)$ denote the subgroup of rotations around the z-axis,*

$$\mathrm{stab}(\vec{e}_z) := \{\boldsymbol{R}_\alpha^{\vec{e}_z} : \alpha \in (-\pi,\pi]\}.$$

Example 6.17. The left coset space $\mathrm{SO}(3)/\mathrm{stab}(\vec{e}_z)$ defines a partition of $\mathrm{SO}(3)$ via

$$(\alpha_1,\beta_1,\gamma_1) \sim (\alpha_2,\beta_2,\gamma_2) \quad \text{ if and only if } \quad \beta_1 = \beta_2 \text{ and } \gamma_1 = \gamma_2.$$

Then, we have

$$\left[\mathbf{R}_{(\alpha,\beta,\gamma)}\right]_{\mathrm{stab}(\vec{e}_z)} = \{\mathbf{R}_{(\alpha',\beta,\gamma)} : \alpha' \in (-\pi,\pi]\}.$$

Now we state the main result of this section. The proof is given in Section 6.7.

Proposition 6.18. *For the subgroup $(\{\vec{0}\} \rtimes \mathrm{stab}(\vec{e}_z)) \subseteq \mathrm{SE}(3)$ with translation $\vec{0}$, the space $\mathbb{R}^3 \times \mathbb{S}^2$ is isomorphic to the left coset space $\mathrm{SE}(3)/(\{\vec{0}\} \rtimes \mathrm{stab}(\vec{e}_z))$, and their respective elements can be identified via*

$$\mathbb{R}^3 \times \mathbb{S}^2 \ni (\vec{v},\vec{u}) \equiv [(\vec{v},\boldsymbol{R})]_{(\{\vec{0}\} \rtimes \mathrm{stab}(\vec{e}_z))} \in \mathrm{SE}(3)/(\{\vec{0}\} \rtimes \mathrm{stab}(\vec{e}_z)), \text{ where } \boldsymbol{R}\vec{e}_z = \vec{u}.$$

Notation 6.19. *We set* $\mathbb{R}^3 \rtimes \mathbb{S}^2 := \mathrm{SE}(3)/(\{\vec{0}\} \rtimes \mathrm{stab}(\vec{e}_z))$.

We emphasize that the left coset space does not form a group. In particular, the group operation of $\mathrm{SE}(3)$ is not well-defined on $\mathbb{R}^3 \rtimes \mathbb{S}^2$. Here, we will avoid its application by instead using the left-regular action of $\mathrm{SE}(3)$ on $\mathbb{R}^3 \times \mathbb{S}^2$ and $\mathbb{R}^3 \rtimes \mathbb{S}^2$, respectively, see § 6.3.4 and Corollary 6.62.

Lemma 6.20. *The group operation of* $\mathrm{SE}(3)$ *is not well-defined on the corresponding equivalence classes of the left coset space* $\mathbb{R}^3 \rtimes \mathbb{S}^2$.

6.3.3 Orientation scores

In diffusion-weighted magnetic resonance imaging, we examine fibrous structure. Hence, we should preserve the orientational information when processing such data. In this section, we will deduce from the embedding of $\mathbb{R}^3 \times \mathbb{S}^2$ into $\mathrm{SE}(3)$ a criterion for the appropriateness of operations on dMRI data. For this purpose, we will follow Franken and Duits, see Duits and Franken [2011]; Duits et al. [2011]; Franken [2008], by introducing formal definitions of images, orientation scores, and group representations.

Definition 6.21. *A* d*-dimensional image,* $d \in \mathbb{N}$*, is a square-integrable function* $f \in \mathbb{L}_2(\mathbb{R}^d, \mathbb{R})$ *with compact support in the image domain* $\Omega := \bigotimes_{i=1}^{d} [0, X_i]$*, where* $X_i > 0$.

Operations on images should be invariant with respect to translations and rotations. Since the positions of patients in a tomograph may vary, a standardization of the acquired images is important. However, it should not matter what we apply first, the processing operator or a transformation by translations and rotations. Formally, this is described as follows.

Definition 6.22. *An operator* $\Gamma : \mathbb{L}_2(\mathbb{R}^d, \mathbb{R}) \to \mathbb{L}_2(\mathbb{R}^d, \mathbb{R})$ *is Euclidean invariant if it is translation and rotation invariant, that is if*

$$\mathcal{E}_{(\vec{v},\boldsymbol{R})} \circ \Gamma = \Gamma \circ \mathcal{E}_{(\vec{v},\boldsymbol{R})}$$

for all $\vec{v} \in \mathbb{R}^d$ *and every* $\boldsymbol{R} \in \mathrm{SO}(d)$*, where* $\mathcal{E}_{(\vec{v},\boldsymbol{R})} = \mathcal{R}_{\boldsymbol{R}} \circ T_{\vec{v}}$ *with* $(\mathcal{R}_{\boldsymbol{R}} f)(\vec{v}_0) = f(\boldsymbol{R}^{-1}\vec{v}_0)$ *and* $(T_{\vec{v}} f)(\vec{v}_0) = f(\vec{v}_0 - \vec{v})$.

Adding the orientation as an additional dimension to the domain of an image allows to distinguish structures with different local orientations within the image. This corresponds to the biological visual system, which separates objects with different orientations. In Figure 6.11, we show two examples.

Definition 6.23. *Let* $d \in \mathbb{N}$ *be fixed.*

- *Any function* $\mathfrak{U} \in \mathbb{L}_2(\mathbb{R}^d \times \mathbb{S}^{d-1}, \mathbb{R})$ *is called* standard d*-dimensional orientation score.*
- *Any function* $U \in \mathbb{L}_2(\mathrm{SE}(d), \mathbb{R})$ *is called* general d*-dimensional orientation score.*

Example 6.24.

- For every 2-dimensional image f, a 2-dimensional orientation score U_f can be constructed. Appropriate construction and reconstruction formulas are given by Franken [2008, Eq. (2.6) & (2.9)], where the author proposed to correlate the image f with the complex conjugated rotation of an anisotropic kernel $\psi \in \mathbb{L}_2(\mathbb{R}^2, \mathbb{R})$, that is

$$U_f(\vec{v}, \mathbf{R}_\theta) := \int_{\mathbb{R}^2} \overline{\mathcal{R}_{\mathbf{R}_\theta}[\psi](\vec{v}_0 - \vec{v})} f(\vec{v}_0) d\vec{v}_0, \quad (\vec{v}, \mathbf{R}_\theta) \in \mathrm{SE}(2),$$

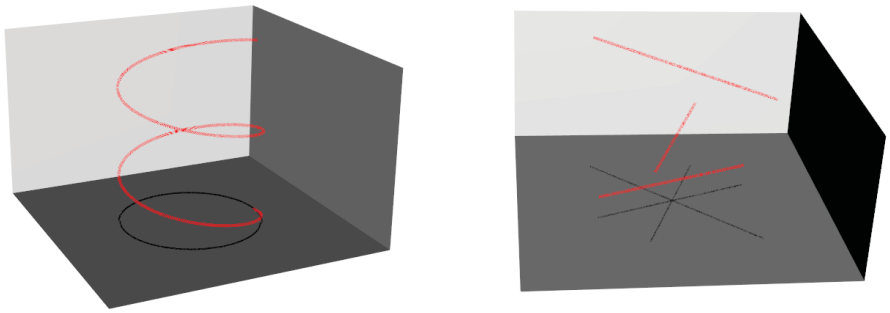

Figure 6.11: Two examples of 2-dimensional orientation scores and their corresponding images as projection on the plane.

where $\mathcal{R}_{\mathbf{R}_\theta}$ is as in Definition 6.22 with $\mathbf{R}_\theta$ from Equation (6.25) (page 103). Any invertible construction formula allows to define the image operation

$$\Gamma : \mathbb{L}_2(\mathbb{R}^2, \mathbb{R}) \to \mathbb{L}_2(\mathbb{R}^2, \mathbb{R}) \quad \text{as} \quad \Gamma[f] := (\mathcal{W}_0^* \circ \Phi \circ \mathcal{W}_0)[f],$$

where $\mathcal{W}_0$ denotes the construction formula of the orientation score U_f from f, Φ the orientation score operation, and $\mathcal{W}_0^*$ the inverse of $\mathcal{W}_0$ which reconstructs f from U_f.

- Diffusion-weighted magnetic resonance imaging yields data in form of a standard three-dimensional orientation score $\mathfrak{U}$, see Equation (6.2) (page 86) and Notation 6.1 (page 91).

For $d = 2$, we may identify the two types of orientation scores since $\mathbb{S}^1$ and $\mathrm{SO}(2)$ are isomorphic. For higher dimensions, general orientation scores benefit from the group structure of $\mathrm{SE}(d)$. By the following property, 3-dimensional standard orientation scores can be identified with general orientation scores. This can be considered as a lifting, where a certain invariance ensures the invertibility. Then, most of the following concepts hold for both, standard and general orientation scores. However, for standard orientation scores $\mathfrak{U}$, we always need to ensure that an applied transformation $\mathfrak{U} \mapsto \Phi(\mathfrak{U})$ is well-defined on the left coset space $\mathbb{R}^3 \rtimes \mathbb{S}^2$.

Definition 6.25. *The function $F : \mathrm{SE}(3) \to \mathbb{R}$ is α-right invariant if*

$$F(\vec{v}, \boldsymbol{R}_{(\alpha'+\alpha,\beta,\gamma)}) = F(\vec{v}, \boldsymbol{R}_{(\alpha,\beta,\gamma)}) \quad \textit{for all } \alpha' \in (-\pi, \pi].$$

Corollary 6.26. *For every standard orientation score $\mathfrak{U} : \mathbb{R}^3 \times S^2 \to \mathbb{R}$, there is a unique α-right invariant general orientation score $U : \mathrm{SE}(3) \to \mathbb{R}$ given by*

$$U(\vec{v}, \boldsymbol{R}) := \mathfrak{U}(\vec{v}, \vec{u}) \qquad \textit{if } \boldsymbol{R}\vec{e}_z = \vec{u}.$$

Vice versa, for every α-right invariant general orientation score U, there is a unique standard orientation score $\mathfrak{U}$ with

$$\mathfrak{U}(\vec{v}, \vec{u}(\beta, \gamma)) := U(\vec{v}, \boldsymbol{R}_{(\alpha,\beta,\gamma)}) \qquad \textit{for all } \alpha \in (-\pi, \pi].$$

For data which are measured in form of an orientation score, the orientation marginal yields an associated image.

Definition 6.27. *The orientation marginal of a general d-dimensional orientation score U is defined by*

$$\mathcal{W}^*[U](\vec{v}) := \int_{\mathrm{SO}(d)} U(\vec{v}, \boldsymbol{R}) d\mu(\boldsymbol{R}), \qquad U \in \mathbb{L}_2(\mathrm{SE}(d), \mathbb{R}), \vec{v} \in \mathbb{R}^d,$$

where μ denotes a Haar measure on $\mathrm{SO}(d)$.

Example 6.28. The data vector $S(\vec{v}, \vec{g})$ with $\vec{g} \in G_0$ is constructed in Equation (6.11) (page 94) by the orientation marginals of the corresponding orientation scores or diffusion-weighted signals $S_b : \vec{v} \times G_b \to \mathbb{R}$ with $b \in B_0$.

Next we aim to answer the following question. How to choose the orientation score operation Φ in order to ensure the Euclidean invariance of the image (or orientation marginal) operation Γ? For this purpose, we consider representations of groups, see Dungey et al. [2003, §2.2 p. 13] or Rossmann [2002, §6.1],

Definition 6.29. *Let G denote a group, H a finite dimensional vector space, and $\mathcal{B}(H)$ the space of all linear, bounded, and invertible operators $H \to H$. The map $\mathcal{A} : G \to \mathcal{B}(H)$ with $g \mapsto \mathcal{A}_g$ is called representation of G on $\mathcal{B}(H)$ if $\mathcal{A}$ is a group homomorphism, that is if $e \mapsto I$, $g_1 \cdot g_2 \mapsto \mathcal{A}_{g_1} \circ \mathcal{A}_{g_2}$, and $g^{-1} \mapsto (\mathcal{A}_g)^{-1}$.*

Example 6.30. Let G denote a locally compact group and μ a Haar measure on G. The left-regular representation $\mathcal{L}_G$ of G on the square-integrable functions $\mathbb{L}_2(G, \mu)$ is defined by

$$\left(\mathcal{L}_G(g_2)\phi\right)(g_1) := (\phi \circ L_{g_2}^{-1})(g_1) := \phi(g_2^{-1} g_1)$$

for all $g_2 \in G$, $\phi \in \mathbb{L}_2(G, \mu)$ and μ-almost every $g_1 \in G$. The right-regular representation $\mathcal{Q}_G$ of G on $\mathbb{L}_2(G, \mu)$ equals

$$\left(\mathcal{Q}_G(g_2)\phi\right)(g_1) := (\phi \circ Q_{g_2})(g_1) := \phi(g_1 g_2).$$

The operator $g \mapsto \mathcal{E}_g$, $g \in \mathrm{SE}(d)$, in Definition 6.22 equals the left-regular representation of $\mathrm{SE}(d)$ on $\mathbb{L}_2(\mathbb{R}^d, \mathbb{R})$.

Notation 6.31. *An orientation score operator $\Phi : \mathbb{L}_2(\mathrm{SE}(d), \mathbb{R}) \to \mathbb{L}_2(\mathrm{SE}(d), \mathbb{R})$ is called left-invariant (right-invariant) if it is invariant with respect to the left-regular (right-regular)* $\mathrm{SE}(d)$*-representation, that is if*

$$\mathcal{L}_g \circ \Phi = \Phi \circ \mathcal{L}_g \quad \text{or} \quad \mathcal{Q}_g \circ \Phi = \Phi \circ \mathcal{Q}_g, \quad \text{respectively,}$$

where $\mathcal{L}_g := \mathcal{L}_{\mathrm{SE}(d)}(g)$ and $\mathcal{Q}_g := \mathcal{Q}_{\mathrm{SE}(d)}(g)$ are as in Example 6.30. Moreover, we call a distance $\delta : \mathrm{SE}(d) \times \mathrm{SE}(d) \to [0, \infty)$ left-invariant (right-invariant) if

$$\delta(n_1, n_2) = \delta(L_m^{-1}[n_1], L_m^{-1}[n_2]) \quad \text{or} \quad \delta(n_1, n_2) = \delta(Q_m[n_1], Q_m[n_2]), \quad \text{respectively,}$$

for all $n_1, n_2, m \in \mathrm{SE}(3)$ with L_m and Q_m as in Example 6.30.

We study the relationships between the mentioned group representations of $\mathrm{SE}(d)$ on images and orientation scores.

Lemma 6.32. *Let $\mathcal{W}^*$ be as in Definition 6.27, $\mathcal{L}_g = \mathcal{L}_{\mathrm{SE}(d)}(g)$, $\mathcal{Q}_g = \mathcal{Q}_{\mathrm{SE}(d)}(g)$, and $\mathcal{E}_g$ as in Example 6.30. Then, for all $g \in \mathrm{SE}(d)$ and every definition of $\tilde{\mathcal{F}}_g$, it holds*

$$\mathcal{W}^* \circ \mathcal{L}_g \circ U = \mathcal{E}_g \circ \mathcal{W}^* \circ U \quad \text{and} \quad \mathcal{W}^* \circ \mathcal{Q}_g \circ U \neq \tilde{\mathcal{F}}_g \circ \mathcal{W}^* \circ U. \tag{6.28}$$

Additionally, for $d = 2$, let $\Gamma = \mathcal{W}_0^* \circ \Phi \circ \mathcal{W}_0$ denote an image operator, where $\mathcal{W}_0$ and $\mathcal{W}_0^*$ are as in [Franken, 2008, Eqs. (2.6) & (2.9)]. Then, Γ is Euclidean invariant if and only if the corresponding orientation score operator Φ is left-invariant.

Equation (6.28) follows in the same manner as in [Franken, 2008, §7.4.3], where the same result is considered supposing $d = 3$. The last assertion of Lemma 6.32 was proven by Franken [2008, §2.7.1]. This lemma establishes together with Proposition 6.33, below, that orientation score operations Φ should be left-invariant. For operators on dMRI data, this ensures the Euclidean invariance in the voxel space $\mathbb{R}^3$ and the rotation invariance in the gradient space $\mathbb{S}^2$.

Proposition 6.33. *Let $\Phi : \mathbb{L}_2(\mathrm{SE}(d),\mathbb{R}) \to \mathbb{L}_2(\mathrm{SE}(d),\mathbb{R})$ be an orientation score operator, and consider the corresponding orientation marginal operator $\Gamma : \mathbb{L}_2(\mathbb{R}^d,\mathbb{R}) \to \mathbb{L}_2(\mathbb{R}^d,\mathbb{R})$ with $\Gamma[\mathcal{W}^*(U)] := \mathcal{W}^*[\Phi(U)]$, where $\mathcal{W}^*$ is as in Definition 6.27. Then, Γ is Euclidean invariant if Φ is left-invariant, that is*

$$[\mathcal{L}_g \circ \Phi = \Phi \circ \mathcal{L}_g \text{ for all } g \in \mathrm{SE}(d)] \quad \textit{implies} \quad [\mathcal{E}_g \circ \Gamma = \Gamma \circ \mathcal{E}_g \text{ for all } g \in \mathrm{SE}(d)],$$

where $\mathcal{L}_g = \mathcal{L}_{\mathrm{SE}(d)}(g)$ and $\mathcal{E}_g$ are given in Example 6.30.

There is a nice relation between left-invariant and convolution operators on $\mathbb{L}_2(\mathrm{SE}(d),\mathbb{R})$. The proof is based on the important Dunford-Pettis Theorem.

Proposition 6.34. *Every linear, bounded, and left-invariant operator*

$$\Phi : \mathbb{L}_2(\mathrm{SE}(d),\mathbb{R}) \to \mathbb{L}_\infty(\mathrm{SE}(d),\mathbb{R})$$

can be represented as an $\mathrm{SE}(d)$-convolution with some kernel K, that is

$$[\Phi U](g) = (K *_{\mathrm{SE}(d)} U)(g) = \int_{\mathrm{SE}(d)} K(h^{-1}g)U(h)d\mu(h),$$

where μ denotes a Haar measure on $\mathrm{SE}(d)$. Moreover, every convolution operator Φ_K with $\Phi_K U := K *_{\mathrm{SE}(3)} U$, $U \in \mathbb{L}_2(\mathrm{SE}(d),\mathbb{R})$, is left-invariant.

For standard orientation scores, we need to ensure that $\Phi\mathfrak{U}$ is α-right invariant. The following result was proven for $d = 2$ by Franken [2008, §7.4.5]. The extension to larger dimensions $d > 2$ is straight forward.

Lemma 6.35. *Let the general orientation score U be α-right invariant. Then, $K *_{\mathrm{SE}(3)} U$ is α-right invariant if the convolution kernel K is α-right and α-left invariant, that is if*

$$K = \mathcal{L}_{(\vec{0},\boldsymbol{R}^{\vec{e}_z}_{-\alpha''})} \circ \mathcal{Q}_{(\vec{0},\boldsymbol{R}^{\vec{e}_z}_{\alpha'})} \circ K \quad \textit{for all } \alpha',\alpha'' \in (-\pi,\pi].$$

6.3.4 Left-invariance of (ms)POAS

Finally, we justify that POAS and msPOAS are left-invariant as required in § 6.3.3. For this purpose, we adapt Notation 6.31 to operators on dMRI data. For the sake of notational simplicity, we avoid the embedding of $\mathbb{R}^3 \times \mathbb{S}^2$ into $\mathrm{SE}(3)$, using instead the left-regular action of $\mathrm{SE}(3)$ on $\mathbb{R}^3 \times \mathbb{S}^2$. We emphasize that the following notation could be formulated in an analogous manner with the left coset space $\mathbb{R}^3 \rtimes \mathbb{S}^2$ in place of $\mathbb{R}^3 \times \mathbb{S}^2$.

Notation 6.36.

- *The left-regular action L of $\mathrm{SE}(3)$ on $\mathbb{R}^3 \times \mathbb{S}^2$ is given by*

$$L_m^{-1}[n] := (\boldsymbol{R}^{-1}(\vec{v}-\vec{w}),\boldsymbol{R}^{-1}\vec{g}) \in \mathbb{R}^3 \times \mathbb{S}^2, \quad m := (\vec{w},\boldsymbol{R}) \in \mathrm{SE}(3), n := (\vec{v},\vec{g}) \in \mathbb{R}^3 \times \mathbb{S}^2.$$

- *We define the left-regular action $\mathcal{L}$ of* $\mathrm{SE}(3)$ *on* $\mathbb{L}_2(\mathbb{R}^3 \times \mathbb{S}^2, \mathbb{R})$ *by*

$$\mathcal{L}_n[U](m) := U(L_n^{-1}(m)) = U(\boldsymbol{R}^{-1}(\vec{v} - \vec{w}), \boldsymbol{R}^{-1}\vec{g}),$$

  where $U \in \mathbb{L}_2(\mathbb{R}^3 \times \mathbb{S}^2, \mathbb{R})$, $m = (\vec{v}, \vec{g}) \in \mathbb{R}^3 \times \mathbb{S}^2$, *and* $n = (\vec{w}, \boldsymbol{R}) \in \mathrm{SE}(3)$.
- *Then, we call the operator* $\Phi : \mathbb{L}_2(\mathbb{R}^3 \times \mathbb{S}^2, \mathbb{R}) \to \mathbb{L}_2(\mathbb{R}^3 \times \mathbb{S}^2, \mathbb{R})$ *left-invariant if, for all* $n \in \mathrm{SE}(3)$, *it holds* $\mathcal{L}_n \circ \Phi = \Phi \circ \mathcal{L}_n$.
- *A distance* $\delta : (\mathbb{R}^3 \times \mathbb{S}^2) \times (\mathbb{R}^3 \times \mathbb{S}^2) \to [0, \infty)$ *is called left-invariant if*

$$\delta(n_1, n_2) = \delta(L_m^{-1}[n_1], L_m^{-1}[n_2])$$

  for all $n_1, n_2 \in \mathbb{R}^3 \times \mathbb{S}^2$ *and every* $m \in \mathrm{SE}(3)$.

Additionally, we introduce the following operators.

Notation 6.37. *Let the POAS procedure be defined as in § 6.2.3 and the msPOAS method as in Algorithm 3 (page 97), using some distance* $\delta_\kappa : (\mathbb{R}^3 \times \mathbb{S}^2) \times (\mathbb{R}^3 \times \mathbb{S}^2) \to [0, \infty)$. *Additionally, assume* $\lambda > 0$ *to be fixed and the sequence of location bandwidths* $\{h^{(k)}\}_{k=0}^{k^*}$ *and the balancing parameter* κ *to be chosen as described in § 6.2.2. For* $k \in \{0, ..., k^*\}$, *we define two operators*

$$\phi_k, \varphi_k : \mathbb{L}_2(\mathbb{R}^3 \times \mathbb{S}^2, \mathbb{R}) \to \mathbb{L}_2(\mathbb{R}^3 \times \mathbb{S}^2, \mathbb{R}),$$

where ϕ_k *describes the POAS procedure, and* φ_k *equals the msPOAS method. More precisely, the operator* ϕ_k *is defined by*

$$\phi_k[S_b](m) := \tilde{S}_b^{(k)}(m), \qquad S_b \in \mathbb{L}_2(V \times G_b, \mathbb{R}), b \in B_0, m \in V \times G_b,$$

where $\tilde{S}_b^{(k)}(m)$ *is as in Equation* (6.23) *(page 102). Similarly, the operator* φ_k *is defined by*

$$\varphi_k[S_b](m) := \tilde{S}_b^{(k)}(m), \qquad S_b \in \mathbb{L}_2(V \times G_b, \mathbb{R}), b \in B_0, m \in V \times G_b,$$

where $\tilde{S}_b^{(k)}(m)$ *is as in Equation* (6.18) *(page 97).*

Theorem 6.38. *We consider the setting in Notation 6.1 (page 91) and suppose the gradient schemes* G_b, $b \in B$, *of all* q*-shells to be identical. The operators* ϕ_k *and* φ_k *in Notation 6.37 are left-invariant for every* $k \in \{0, ..., k^*\}$ *if and only if the applied distance* δ_κ *is left-invariant. In the case that the distance* δ_κ *is based on the left coset space* $\mathbb{R}^3 \rtimes \mathbb{S}^2$, *it holds that the operators* ϕ_k *and* φ_k *are well-defined with respect to the embedding of* $\mathbb{R}^3 \times \mathbb{S}^2$ *into* $\mathrm{SE}(3)$ *for every* $k \in \{0, ..., k^*\}$ *if and only if* δ_κ *is well-defined.*

The proof is given in Section 6.7.

6.4 Distance functions on the measurement space

For application of the (ms)POAS procedure, we seek for a distance on $\mathbb{R}^3 \times \mathbb{S}^2$. Basically, there are two canonical approaches. Motivated by the description of dMRI data as an α-right invariant orientation score on the special Euclidean motion group $\mathrm{SE}(3)$, we will first consider the natural distance on a Lie group G, see § 6.4.2. Then, we will compare in § 6.4.3 the resulting distance with the natural metric on $\mathbb{R}^3 \times \mathbb{S}^2$.

For our first approach, we will recall in § 6.4.1 some basic concepts of differential geometry, such as Lie groups and Lie algebras, different kinds of coordinates and left-invariant vector

- Left-invariant operations on dMRI data ensure Euclidean invariance in the voxel space $\mathbb{R}^3$ and rotation invariance in the gradient space $\mathbb{S}^2$.
- The isomorphism

$$\mathbb{R}^3 \times \mathbb{S}^2 \ni (\vec{v}, \vec{u}) \mapsto [(\vec{v}, \mathbf{R}_{\vec{u}})]_{(\{\vec{0}\} \rtimes \mathrm{stab}(\vec{e}_z))} \in \mathrm{SE}(3)/(\{\vec{0}\} \rtimes \mathrm{stab}(\vec{e}_z))$$

  embeds $\mathbb{R}^3 \times \mathbb{S}^2$ into $\mathrm{SE}(3)$, where $\mathbf{R}_{\vec{u}}\vec{e}_z = \vec{u}$.
- For every standard orientation score $\mathfrak{U} : \mathbb{R}^3 \times S^2 \to \mathbb{R}$, there is an α-right invariant general orientation score $U : \mathrm{SE}(3) \to \mathbb{R}$ and vice versa, where

$$\mathfrak{U}(\vec{v}, \vec{u}(\beta, \gamma)) = U(\vec{v}, \mathbf{R}_{(\alpha,\beta,\gamma)}) \qquad \text{for all } \alpha \in (-\pi, \pi].$$

- Let, in Notation 6.1 (page 91), the gradient schemes of all q-shells be identical. The method (ms)POAS is left-invariant and well-defined with respect to the embedding of $\mathbb{R}^3 \times \mathbb{S}^2$ into $\mathrm{SE}(3)$ if and only if the distance δ_κ in the location kernel is left-invariant and well-defined.

Figure 6.12: Section 6.3 in a nutshell

fields. Using a criterion by Duits et al. [2011, App. E], we will prove in § 6.4.2 the Carnot-Carathéodory distance in Definition 6.56 to be left-invariant and well-defined on the left coset space $\mathbb{R}^3 \rtimes \mathbb{S}^2$. Its approximation in Notation 6.63 has been used in the implementation of the POAS method, see § 6.2.3, Becker et al. [2012], and Tabelow and Polzehl [2013]. However, this is a rough approximation which violates the required rotation invariance in the gradient space. Therefore, in the implementation of the msPOAS procedure, see § 6.2.1, Becker et al. [2013], and Tabelow and Polzehl [2013], we followed another approach.

By definition of the group product on $\mathrm{SE}(3)$, the rotation group $\mathrm{SO}(3)$ acts on the translation group $\mathbb{R}^3$, not vice versa. Hence, $\mathbb{R}^3$ and $\mathrm{SO}(3)$ play different roles. These roles carry over to the left coset space $\mathbb{R}^3 \rtimes \mathbb{S}^2$ by the embedding, see Corollary 6.62. Nevertheless, the approximated distance Δ_κ on $\mathbb{R}^3 \times \mathbb{S}^2$, see Notation 6.63, depends on two well separated parts, the squared Euclidean metric in Equation (6.32) (page 121) and a term which solely depends on the two gradients $\vec{u}_1, \vec{u}_2 \in \mathbb{S}^2$. Against the comprehensible belief that both parts would interact, this implies that the Euclidean and the spherical distances of voxels and gradients, respectively, can be considered on their own. This will be done in our second proposal, see Notation 6.64, where we will introduce a left-invariant pseudometric on $\mathbb{R}^3 \times \mathbb{S}^2$, which does not rely on the embedding into $\mathrm{SE}(3)$.

6.4.1 A differential geometric approach

The embedding of $\mathbb{R}^3 \times \mathbb{S}^2$ into $\mathrm{SE}(3)$ enables the application of useful tools from Lie group theory, providing several distance functions on $\mathrm{SE}(3)$. On Lie groups a distance can be defined, using the algebraic basis of the corresponding Lie algebra. For simplicity, we will restrict to the present case of linear Lie groups as considered by Rossmann [2002]. We will summarize some important definitions and properties stated in [Rossmann, 2002] on the pages 2, 12, 14, 23, 30, 44–46, and 132–137. The following results will be used for the definition and justification of our first distance in Notation 6.63. The second proposal in Notation 6.64 does not rely on the embedding into $\mathrm{SE}(3)$.

Definition 6.39.

1. *A group $(G, \cdot_G)$ which forms an analytic manifold is called Lie group if the group operations*

$$G \times G \ni (g_1, g_2) \mapsto g_1 \cdot_G g_2 \in G \quad \text{and} \quad G \ni g \mapsto g^{-1} \in G$$

   are analytic.
2. *A group G is linear if all of its elements can be represented by invertible matrices and if the group operation is given by the matrix multiplication.*
3. *The tangent space $T_g(G)$ of a linear Lie group G at the point $g \in G$ consists of all matrices X for which a continuously differentiable curve $a : I \to G$ exists, such that $I \subset \mathbb{R}$ is an interval, $0 \in I$, $a(0) = g$, and $a'(0) = X$.*
4. *The space $\mathfrak{g} := (T_e(G), [.,.])$ is called Lie algebra of G with Lie bracket or Lie commutator $[.,.]$ given by*

$$[X, Y] := XY - YX \text{ for all } X, Y \in T_e(G).$$

Example 6.40. We know from Example 6.12 that the special Euclidean motion group $\mathrm{SE}(d)$ is an analytic manifold. Additionally, its group operations are analytic. Hence, for $\mathrm{SE}(d)$, all conditions to be a linear Lie group are satisfied. More generally, the group of affine transformation $\mathbb{R}^d \rtimes GL(d, \mathbb{R})$ forms a linear Lie group [Rossmann, 2002, §4.1 Ex. 2].

Proposition 6.41. *We use the notation of Definition 6.39.*

1. *The binary operator $[.,.] : T_e(G) \times T_e(G) \to T_e(G)$ satisfies, for all $X, Y, Z \in T_e(G)$ and $a, b \in \mathbb{R}$, the following properties:*
   - *Bilinearity:* $[aX + bY, Z] = a[X, Z] + b[Y, Z]$, $[Z, aX + bY] = a[Z, X] + b[Z, Y]$,
   - *Anticommutativity:* $[X, Y] = -[Y, X]$,
   - *Jacobi identity:* $[X, [Y, Z]] + [Y, [Z, X]] + [Z, [X, Y]] = 0$.
2. *The tangent space $T_e(G)$ is a real vector space,*

$$X, Y \in T_e(G) \text{ implies } aX + bY \in T_e(G) \text{ for all } a, b \in \mathbb{R}.$$

   Moreover, $T_e(G)$ is closed under the Lie bracket operation,

$$X, Y \in T_e(G) \text{ implies } [X, Y] \in T_e(G).$$

3. *The exponential function $\exp : \mathfrak{g} \to G$ describes the relationship between the Lie algebra $\mathfrak{g} = (T_e(G), [.,.])$ and the Lie group G. It is defined by*

$$\exp(X) := \boldsymbol{1}_{d \times d} + \sum_{n=1}^{\infty} \frac{X^n}{n!} \in G, \quad X \in \mathfrak{g},$$

   and it carries a neighborhood $\mathfrak{g}_0 \subseteq \mathfrak{g}$ of $\boldsymbol{0} \in \mathfrak{g}$ one-to-one onto a neighborhood $G_1 \subseteq G$ of $\boldsymbol{1} \in G$. Then, the local inverse $\log : G_1 \to \mathfrak{g}_0$ of $\exp|_{\mathfrak{g}_0}$ is given by

$$\log(M) := \sum_{n=1}^{\infty} (-\boldsymbol{1})^{k-1} (M - \boldsymbol{1})^k / k \in \mathfrak{g}_0, \quad M \in G_1.$$

   If $XY = YX$, then it holds $\exp(X)\exp(Y) = \exp(Z)$ with $Z := X + Y$. Else, there is a unique solution $Z = C(X, Y)$, given as a series of repeated brackets of X and Y, which converges if $X, Y, Z \in \mathfrak{g}$ are sufficiently close to zero. This is known as Campbell-Hausdorff formula.

We have seen that every Lie algebra is a vector space. Next we define an algebraic basis of a Lie algebra as a subset of its vector space basis, see Dungey et al. [2003, §II.2 p. 14]. In § 6.4.2, we will use a weighted algebraic basis of $\mathrm{SE}(3)$ in order to approximate the Carnot-Carathéodory distance.

Definition 6.42. *Let G be a linear Lie group and $\mathfrak{g}$ the Lie algebra of G with vector space basis $\mathcal{B} := \{X_i \in T_e(G) : i = 1, ..., d\}$. The subset $\mathcal{B}' := \{X_1, ..., X_{d'}\}$, $d' \leq d$, forms an algebraic basis with rank $r \in \mathbb{N}$ if*

$$\begin{aligned} \mathfrak{g}_1 &:= \mathit{span}\{X_1, ..., X_{d'}\}, & \mathfrak{g}_2 &:= \mathit{span}\{\mathfrak{g}_1 \cup [\mathfrak{g}_1, \mathfrak{g}_1]\}, \\ &... & \mathfrak{g}_r &:= \mathit{span}\{\mathfrak{g}_{r-1} \cup [\mathfrak{g}_{r-1}, \mathfrak{g}_{r-1}]\} = \mathfrak{g}, \end{aligned}$$

where $[\mathfrak{g}_l, \mathfrak{g}_l] := \{[X, Y] : X, Y \in \mathfrak{g}_l\}$ for every $l \in \mathbb{N}$. The algebraic basis B' is weighted if a weight $w_i \in [1, \infty)$ is given for every $X_i \in \mathcal{B}'$.

The following proposition [Rossmann, 2002, §2.4 Prop. 3] allows to distinguish two kinds of coordinates as considered, for instance, by Rossmann [2002, §2.3 & 4.1 Probl. 11] and Dungey et al. [2003, II.4.17]. Here, we use the equivalence between connectedness and path connectedness on linear groups. We emphasize that the exponential map of an arbitrary Lie group G does not need to be surjective.

Proposition 6.43. *Let G be a linear Lie Group with Lie algebra $\mathfrak{g}$. The connected component G_0 of the identity $e \in G$ is given as the set of finite products of elements of the subset $\exp(\mathfrak{g}) \subseteq G$ and its inverses,*

$$G_0 := \left\{ \prod_{i=1}^{k} g_i : g_i = \exp(X_i) \text{ or } g_i^{-1} = \exp(X_i), X_i \in \mathfrak{g} \text{ for all } i = 1, ..., k \text{ and } k \in \mathbb{N} \right\}.$$

It is the unique open connected subgroup of G. For every $g \in G$, the left coset gG_0 equals the connected component of g, that is the set of all elements $g_1 \in G$ which can be joined to g by a continuous path $\phi : [0, 1] \to G$ with $\phi(0) = g$ and $\phi(1) = g_1$. The corresponding left coset space G/G_0 forms a non-linear group.

Definition 6.44. *Let G be a linear Lie Group and $\mathfrak{g}$ the Lie algebra of G with vector space basis $\mathcal{B} := \{X_i \in T_e(G) : i = 1, ..., d\}$. For every element $g \in G_0$, with G_0 as in Proposition 6.43, we define two types of coordinates (if existing),*

- $\{\rho_i\}_{i=1,...,d}$ *with* $g = \exp(\sum_{i=1}^{d} \rho_i X_i)$ *(first kind),*
- $\{\zeta_i\}_{i=1,...,d}$ *with* $g = \prod_{i=1}^{d} \exp(\zeta_i X_i)$ *(second kind).*

The coordinates of the first kind are also known as the exponential or canonical coordinates of G. If G_0 does not cover the whole Lie group, that is if $G \setminus G_0 \neq \emptyset$, these coordinates can be defined for $g \in G \setminus G_0$ with respect to the connected component gG_0. Note that the coordinates do not need to be unique. Additionally, the coordinates do not need to exist if the Campbell-Hausdorff formula diverges, see Proposition 6.41 (3) and Rossmann [2002, §1.3]. We apply the introduced concepts to the special Euclidean motion group $\mathrm{SE}(3)$, where both kinds of coordinates exist for every $g \in \mathrm{SE}(3)$.

Example 6.45. We consider the three-dimensional special Euclidean motion group $\mathrm{SE}(3)$. Its matrix representation is given in Equation (6.26) (page 104). The corresponding matrix Lie algebra $\mathfrak{g} = (T_e(\mathrm{SE}(3)), [.,.])$ is spanned by the basis [Duits and Franken, 2011, Eq. (52)]

$$X_1 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad X_2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad X_3 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$
$$X_4 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad X_5 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad X_6 = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

This leads to the following table of Lie brackets [Duits and Franken, 2011, Eq. (53)]

$$[X_i, X_j]_{i,j} = \begin{pmatrix} 0 & 0 & 0 & 0 & X_3 & -X_2 \\ 0 & 0 & 0 & -X_3 & 0 & X_1 \\ 0 & 0 & 0 & X_2 & -X_1 & 0 \\ 0 & X_3 & -X_2 & 0 & X_6 & -X_5 \\ -X_3 & 0 & X_1 & -X_6 & 0 & X_4 \\ X_2 & -X_1 & 0 & X_5 & -X_4 & 0 \end{pmatrix}.$$

We observe that the subsets $\{X_3, X_4, X_5\}$ and $\{X_1, X_4, X_6\}$ are examples for an algebraic basis of $\mathfrak{g}$.

For $g = ((x, y, z)^T, \mathbf{R}_{(\alpha,\beta,\gamma)}) \in \mathrm{SE}(3)$, the coordinates of the second kind with respect to the above basis are given as

$$(\zeta_1, \zeta_2, \zeta_3, \zeta_4, \zeta_5, \zeta_6) := (x, y, z, \gamma, \beta, \alpha).$$

This follows from the series expansion of sine and cosine since

$$\exp(\zeta_4 X_4) = \begin{pmatrix} \mathbf{R}_{\zeta_4}^{\vec{e}_x} & \vec{0} \\ \vec{0}^T & 1 \end{pmatrix}, \qquad \exp(\zeta_5 X_5) = \begin{pmatrix} \mathbf{R}_{\zeta_5}^{\vec{e}_y} & \vec{0} \\ \vec{0}^T & 1 \end{pmatrix},$$
$$\exp(\zeta_6 X_6) = \begin{pmatrix} \mathbf{R}_{\zeta_6}^{\vec{e}_z} & \vec{0} \\ \vec{0}^T & 1 \end{pmatrix}, \qquad \prod_{i=1}^{3} \exp(\zeta_i X) = \begin{pmatrix} 1 & 0 & 0 & \zeta_1 \\ 0 & 1 & 0 & \zeta_2 \\ 0 & 0 & 1 & \zeta_3 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

By Equation (6.27) (page 105), we get

$$\prod_{i=1}^{6} \exp(\zeta_i X_i) = \begin{pmatrix} & & & \zeta_1 \\ & \mathbf{R}_{(\zeta_6,\zeta_5,\zeta_4)} & & \zeta_2 \\ & & & \zeta_3 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \mathbf{M}_g \equiv g \in \mathrm{SE}(3),$$

which is in accordance with the matrix representation in Equation (6.26) (page 104). The coordinates of the first kind will not be used in this study. Therefore, we omit their explicit representations, which are quite extensive to compute. There existence is ensured since the map $\exp : \mathfrak{g} \to \mathrm{SE}(3)$ is bijective [Rossmann, 2002, §2.1 p. 43 & 2.5 Thm. 1].

For the denoising of dMRI data, we are especially interested in left-invariant operations on orientation scores, see Proposition 6.33 and Figure 6.12 (page 112). For this purpose, we define

left-invariant vector fields on a linear Lie group G. We refer the reader to Franken [2008, §2.8.1 & 7.5.1] for the special cases $\mathrm{SE}(d)$, $d = 2, 3$. Sagle and Walde [1973] consider a more general setting, where the following definitions and properties can be found on the pages 47, 68, 74, 78–79, 82, 115–116.

Definition 6.46. *Let G be a linear Lie group. The set $T(G) := \bigcup_{g \in G} T_g(G)$ is called the tangent bundle of G. A (tangential) vector field $X : G' \to T(G')$ on a subset $G' \subseteq G$ is a map which assigns a tangential vector $X_g \in T_g(G')$ to each point $g \in G'$.*

There is a close relation between vector fields and derivations.

Definition 6.47. *Let G be a linear Lie group and $F(G)$ the set of all real valued analytic functions on G. The mapping $X : F(G) \to F(G)$ is a derivation of the algebra $F(G)$ into $F(G)$ if, for all $f_1, f_2 \in F(G)$ with intersecting open domains $U_l \subseteq G$, $l = 1, 2$, and every $a, b \in \mathbb{R}$, it holds*

- $X(af_1 + bf_2) = aX(f_1) + bX(f_2)$ *with* $(af_1 + bf_2)(h) := af_1(h) + bf_2(h)$, $h \in U_1 \cap U_2$,
- $X(f_1 f_2) = X(f_1) f_2 + f_1 X(f_2)$ *with* $(f_1 f_2)(h) := f_1(h) f_2(h)$, $h \in U_1 \cap U_2$.

Proposition 6.48. *Let G and $F(G)$ be as in Definition 6.47. The map $X : G \to T(G)$ is a vector field on G if and only if X is a derivation of $F(G)$ into $F(G)$.*

Then, we may define the following.

Definition 6.49. *Let G be a linear Lie group and $D(G)$ the set of all vector fields on G.*

- *The differential of an analytic map $f : G \to G$ at $g \in G$ is a map $df(g) : T_g(G) \to T_{f(g)}(G)$ which is defined by $[df(g)(X)](\phi) = X(\phi \circ f)$, where $\phi \in F(G)$ and $X \in T_g(G)$.*
- *A vector field $X \in D(G)$ is called invariant if $[dL_g(e)][X(e)] = X(g)$ for all $g \in G$, where $L_g(h) := g \cdot_G h$ is as in Example 6.30.*
- *A vector field $X \in D(G)$ is left-invariant if $[dL_g(h)][X(h)] = X(gh)$ for all $g, h \in G$.*

Proposition 6.50. *Every left-invariant vector field $X \in D(G)$ is invariant. Moreover, for every tangent vector $X_e \in T_e(G)$, there is a unique left-invariant vector field $\tilde{X} \in D(G)$ which is defined, for every $f \in F(G)$, via*

$$(\tilde{X} f)(g) = X_e(f \circ L_g) \quad \text{for all } g \in G.$$

Then, it holds $\tilde{X}(e) = X_e$, and the space of all left-invariant vector fields on G is isomorphic to the Lie algebra $T_e(G)$ of G.

For an illustrative description of left-invariant vector fields, we follow Franken [2008, §2.8.1]. Let G be a linear Lie group with Lie algebra basis $\{X_i \in T_e(G) : i = 1, ..., d\}$, and let $\phi : \mathbb{R} \to G$ be a curve on G with tangential vector $X_e = \sum_i c_i X_i \in T_e(G)$ at the identity $e \in G$. Then, by left-multiplication with $g \in G$, we get the curve $g\phi$ with tangential vector $X_g = dL_g(X_e) \in T_g(G)$ at g. In other words, the map $dL_g : T_e(G) \to T_g(G)$, known as the push-forward of the left-multiplication, transports X_e in a left-invariant manner to X_g, that is $X_g = \sum_i c_i dL_g X_i$, where the constants c_i remain unchanged in comparison with X_e, while the basis $\{X_i\}_{i=1}^d$ is transformed appropriately.

Example 6.51. We consider the linear Lie group $\mathrm{SE}(2)$ whose Lie algebra equals the three-dimensional Euclidean space $\mathbb{R}^3$. Therefore, we identify the vector space $T_e(\mathrm{SE}(2))$ with $\mathbb{R}^3 = \mathrm{span}\{\vec{e}_x, \vec{e}_y, \vec{e}_\theta\}$, where $\vec{e}_x := (1, 0, 0)^T$, $\vec{e}_y := (0, 1, 0)^T$, and $\vec{e}_\theta := (0, 0, 1)^T$. Left-multiplication of a curve $\phi : \mathbb{R} \longrightarrow \mathrm{SE}(2)$ with $g := (\vec{v}, \mathbf{R}) \in \mathrm{SE}(2)$ leads to a translated and rotated

curve $g\phi = \vec{v} + \mathbf{R}\phi$, where $\mathbf{R}$ is as in Equation (6.25) (page 103). Hence, the push-forward of the left-multiplication rotates the Lie algebra basis, and we have

$$\{(L_g)_*\vec{e}_x, (L_g)_*\vec{e}_y, (L_g)_*\vec{e}_\theta\} = \{\cos\theta\vec{e}_x + \sin\theta\vec{e}_y, -\sin\theta\vec{e}_x + \cos\theta\vec{e}_y, \vec{e}_\theta\}\,.$$

The left-invariant basis has two advantages. First, the basis vectors have a clear interpretation as $(L_g)_*\vec{e}_x$ is tangential to the orientation θ, and $(L_g)_*\vec{e}_y$ is orthogonal to $(L_g)_*\vec{e}_x$. Second, all linear and non-linear combinations of the tangential vectors $(L_g)_*\vec{e}_x, (L_g)_*\vec{e}_y, \vec{e}_\theta$, which are independent of the coordinates, are left-invariant.

For the three-dimensional special Euclidean motion group $\mathrm{SE}(3)$, we get the following [Duits and Franken, 2011, Eq. (25)].

Example 6.52. Let $\{X_i\}_{i=1}^6$ be as in Example 6.45. Then, the left-invariant vector fields with respect to the parametrization of $\mathrm{SE}(3)$ in Equation (6.27) (page 105) are given as

$$\begin{aligned}
A_1 &= \cos\alpha\cos\beta\, X_1 + (\cos\gamma\sin\alpha + \cos\alpha\sin\beta\sin\gamma)\, X_2 \\
&\quad + (\sin\alpha\sin\gamma - \cos\alpha\cos\gamma\sin\beta)\, X_3, \\
A_2 &= -\sin\alpha\cos\beta\, X_1 + (\cos\alpha\cos\gamma - \sin\alpha\sin\beta\sin\gamma)\, X_2 \\
&\quad + (\sin\alpha\sin\beta\cos\gamma + \cos\alpha\sin\gamma)\, X_3, \\
A_3 &= \sin\beta\, X_1 - \cos\beta\sin\gamma\, X_2 + \cos\beta\cos\gamma\, X_3, \\
A_4 &= -\cos\alpha\tan\beta\, X_6 + \sin\alpha\, X_5 + (\cos\alpha/\cos\beta)\, X_4, \\
A_5 &= \sin\alpha\tan\beta\, X_6 + \cos\alpha\, X_5 - (\sin\alpha/\cos\beta)\, X_4, \\
A_6 &= X_6.
\end{aligned}$$

6.4.2 First proposal for a distance on the measurement space of dMRI

For the application on $\mathbb{R}^3 \times \mathbb{S}^2$, a distance on $\mathrm{SE}(3)$ should be well-defined on the left coset space $\mathbb{R}^3 \rtimes \mathbb{S}^2$ and left-invariant as motivated in § 6.3.3. Duits et al. [2011, App. E] provided a sufficient criterion for a Riemannian metric to satisfy both requirements. We recall some related results that can be found in [Jost, 2011, Def. 1.4.1 & Lem. 2.1.1] under a more general setting.

Definition 6.53. *A Riemannian metric on a linear Lie group G is given by a map*

$$\boldsymbol{T} : G \to (T_gG \times T_gG \to \mathbb{R}),$$

which maps each point $g \in G$ to a scalar product $\boldsymbol{T}_g(.,.)$ on T_gG which smoothly depends on g.

Lemma 6.54. *A Riemannian metric on a linear Lie group G with Lie algebra basis $\{X_i\}_{i=1}^d$ can be defined by a two times covariant, symmetric, and positive definite tensor on G, where*

$$\boldsymbol{T}_g = \sum_{i,j=1}^{d} p_{ij}(g)\, dX^i \otimes dX^j$$

for every $g \in G$ with $p_{ij} : G \to \mathbb{R}$ smooth. Here, $\otimes$ denotes the tensor product, and $dX^i \in T_g^*(G)$ is an element of the dual space $T_g^*(G)$ of $T_g(G)$, that is $\langle dX^i, X_j\rangle := \delta_{ij}$.

Then, we state the mentioned criterion by Duits et al. [2011, Eq. (7) & App. E].

Proposition 6.55. *Let $\{A_i\}_{i=1}^6$ denote the left-invariant basis of $\mathrm{SE}(3)$ given in Example 6.52. A metric tensor $\boldsymbol{T}$ on $\mathrm{SE}(3)$ is left-invariant and well-defined on the left coset space $\mathbb{R}^3 \rtimes \mathbb{S}^2$ if and only if*

$$\boldsymbol{T}_{(\vec{v},\boldsymbol{R})} = \sum_{i,j=1}^{6} p_{ij}(\vec{v},\boldsymbol{R})\, dA^i|_{(\vec{v},\boldsymbol{R})} \otimes dA^j|_{(\vec{v},\boldsymbol{R})}, \qquad (\vec{v},\boldsymbol{R}) \in \mathrm{SE}(3),$$

with constants

$$\{p_{ij}(\vec{v},\boldsymbol{R})\}_{ij} \equiv \{p_{ij}\}_{ij} = \mathrm{diag}\{p_{11}, p_{11}, p_{33}, p_{44}, p_{44}, 0\} \tag{6.29}$$

for all $(\vec{v},\boldsymbol{R}) \in \mathrm{SE}(3)$.

Here, p_{6j} and p_{j6} are set to zero for $j \in \{1,...,6\}$ in order to avoid possible short cuts via the direction of the left-invariant vector field A_6 associated to the rotation angle α.

Next we present a distance function on a linear Lie group G which resembles the well-known $\mathbb{L}_2$-norm. As it turns out, this distance is in accordance with the above criterion.

Definition 6.56. *Let G be a connected linear Lie group with algebraic basis $\mathcal{B}' := \{A_i\}_{i=1}^{d'}$ of left-invariant vector fields. For $\tau > 0$, we set*

$$C(\tau) := \{\phi : [0,1] \to G \text{ absolutely continuous with } \dot{\phi}(s) = \sum_{i=1}^{d'} \phi_i(s) A_i|_{\phi(s)}$$
$$\text{almost everywhere and } \int_0^1 \left(\sum_{i=1}^{d'} \phi_i(s)^2 \right)^{1/2} ds \leq \tau\}.$$

Then, for $g_1, g_2 \in G$, we define the distance

$$\Delta(g_1,g_2) := \inf\{\tau > 0 : \text{ there is a curve } \phi \in C(\tau) \text{ with } \phi(0) = g_1 \text{ and } \phi(1) = g_2\}$$

and the corresponding norm

$$|.| : G \to [0,\infty) \quad with \quad |g| := \Delta(g,e).$$

For further details concerning this norm, we refer the reader to Dungey et al. [2003, §II.4]. In [Jost, 2011, Lem. 1.4.1] and [Dungey et al., 2003, §II.4], Δ is introduced as the natural definition of a distance on any connected Lie group G or, more general, on any manifold G. It is called Carnot-Carathéodory distance [Varopoulos et al., 1992, §III.4], and the corresponding norm $|.|$ is known as Riemannian 2-norm [Duits, 2005]. The proof of the following lemma is given in Section 6.7.

Lemma 6.57. *Recall the notations of Definition 6.56 with $G = \mathrm{SE}(3)$, $d' = 5$, and $\{A_i\}_{i=1}^5$ as in Example 6.52. The distance $\Delta : \mathrm{SE}(3) \times \mathrm{SE}(3) \to [0,\infty)$ in Definition 6.56 is symmetric, satisfies the triangle inequality, and locally dominates the Euclidean distance induced by a chart. Additionally, it is well-defined on the left-coset space $\mathbb{R}^3 \rtimes \mathbb{S}^2$ and left-invariant. For Δ and $|.|$ as in Definition 6.56, we get*

$$\Delta(g_1,g_2) = |g_2^{-1} \cdot_{\mathrm{SE}(3)} g_1|, \quad g_1, g_2 \in \mathrm{SE}(3).$$

The distance Δ is defined via integral curves. For practical application, a local approximation is more feasible. Following Dungey et al. [2003, II.4.17], we introduce some norms, using an algebraic basis of the corresponding Lie algebra $\mathfrak{g}$. Here, we concentrate on the coordinates of the second kind, which lead to simpler formulas than the coordinates of the first kind.

Definition 6.58. *Let G be a connected linear Lie group with Lie algebra $\mathfrak{g}$ and exponential function $\exp : \mathfrak{g} \to G$. We consider the basis $\mathcal{B} := \{A_i\}_{i=1}^d$ of left-invariant vector fields with weights*

$$w_i := j \qquad \text{if } A_i \in \mathfrak{g}_j \setminus \mathfrak{g}_{j-1},$$

where $\mathfrak{g}_0 := \emptyset$, $\mathfrak{g}_1 := \{A_i\}_{i=1}^{d'}$ denotes some algebraic basis of $\mathfrak{g}$ and $\mathfrak{g}_l := \mathfrak{g}_{l-1} \cup [\mathfrak{g}_{l-1}, \mathfrak{g}_{l-1}]$ for all $l \in \{2, ..., r\}$ with $\mathfrak{g}_r = \mathfrak{g}$. Then, we define the norms $\|.\|_k : G \to [0, \infty)$ by

$$\left\| \prod_{i=1}^{d} \exp(\zeta_i A_i) \right\|_k := \left(\sum_{i=1}^{d} |\zeta_i|^{k/w_i} \right)^{1/k}, \quad \text{where } k \in \{1\} \cup 2\mathbb{N}.$$

For the last result of the following proposition, we assume the Lie group G to be nilpotent [Dungey et al., 2003, Eq. (II.1)].

Definition 6.59. *Let G be a connected linear Lie group with Lie algebra $\mathfrak{g}$. The lower central series of $\mathfrak{g}$ is given as*

$$\mathfrak{g}_1 = \mathfrak{g} \text{ and } \mathfrak{g}_{j+1} = [\mathfrak{g}, \mathfrak{g}_j].$$

Then, the Lie algebra $\mathfrak{g}$ and its Lie group G are called nilpotent if $\mathfrak{g}_r = \{0\}$ for some $r \in \mathbb{N}$.

Proposition 6.60. *Using the notation from Definitions 6.56 and 6.58, we get the following properties.*

1. *For all $k, k' \in \{1\} \cup 2\mathbb{N}$, there is some constant $c > 0$ which satisfies*

$$c^{-1} \|X\|_k \leq \|X\|_{k'} \leq c \|X\|_k \quad \text{for all } X \in \mathfrak{g}.$$

2. *Let $\mathcal{B}_1$ and $\mathcal{B}_2$ denote distinct bases of $\mathfrak{g}$. Then, the associated norms $|.|^{(1)}$ and $|.|^{(2)}$ are equivalent. In other words, there is some constant $c > 0$ with*

$$c^{-1}|g|^{(1)} \leq |g|^{(2)} \leq c|g|^{(1)} \quad \text{for all } g \in G.$$

3. *Suppose the linear Lie group G to be simply connected and nilpotent. Then, for all $k \in \{1\} \cup 2\mathbb{N}$ there is some constant $c > 0$ with*

$$c^{-1} \|g\|_k \leq |g| \leq c \|g\|_k \quad \text{for } g := \prod_{i=1}^{d} \exp(\zeta_i A_i),$$

where the first inequality is restricted to the case that $|g| \geq 1$.

For the proofs of assertions (2) and (3) with $k = 1$, we refer the reader to Dungey et al. [2003, II.4.1 and II.4.17]. The generalization of the latter for $k \in 2\mathbb{N}$ and assertion (1) hold since $\frac{1}{2}(|a| + |b|)^2 \leq a^2 + b^2 \leq (|a| + |b|)^2$ for all $a, b \in \mathbb{R}$. This leads, for every $k \in \mathbb{N}$, to the existence of some constant $c > 0$ which satisfies

$$c^{-1} \left(\sum_{i=1}^{d} |\zeta_i|^{2k/w_i} \right)^{1/(2k)} \leq \sum_{i=1}^{d} |\zeta_i|^{1/w_i} \leq c \left(\sum_{i=1}^{d} |\zeta_i|^{2k/w_i} \right)^{1/(2k)}.$$

For dMRI data, a slight modification of the distance Δ in Definition 6.56 is recommendable as discussed in § 6.2.1. There, we introduced an additional parameter κ in order to balance between spatial and spherical smoothing.

Definition 6.61. *Let $\{A_i\}_{i=1}^6$ denote the left-invariant basis of $\mathrm{SE}(3)$ given in Example 6.52. For $\tau, \kappa > 0$, we set*

$$C_\kappa(\tau) := \{\phi : [0,1] \to \mathrm{SE}(3) \text{ absolutely continuous with } \dot{\phi}(s) = \sum_{i=1}^{5} \phi_i(s) A_i|_{\phi(s)} \\ \text{almost everywhere and } \int_0^1 \left(\sum_{i=1}^{3} |\phi_i(s)|^2 + \kappa^{-2} \sum_{i=4}^{5} |\phi_i(s)|^2 \right)^{1/2} ds \leq \tau\}.$$

Then, for $g_1, g_2 \in \mathrm{SE}(3)$, we define

$$\Delta_{\mathrm{SE}(3),\kappa}(g_1, g_2) := \inf\{\tau > 0 : \text{ there is a curve } \phi \in C_\kappa(\tau) \text{ with } \phi(0) = \boldsymbol{M}_{g_1}, \phi(1) = \boldsymbol{M}_{g_2}\}$$

and $|g_1|_{\mathrm{SE}(3),\kappa} := \Delta_{\mathrm{SE}(3),\kappa}(g_1, e)$, where we used the matrix representation $\boldsymbol{M}_g$ of $g \in \mathrm{SE}(3)$ in Equation (6.26) *(page 104).*

Corollary 6.62. *Using the notation in Proposition 6.55, the distance $\Delta_{\mathrm{SE}(3),\kappa}$ in Definition 6.61 can be described by a metric tensor with constants $\{p_{ij}\}_{i,j} := \mathrm{diag}\{1,1,1,\kappa^{-2},\kappa^{-2},0\}$. The associated norm $|.|_{\mathrm{SE}(3),\kappa}$ induces, via the embedding of $\mathbb{R}^3 \times \mathbb{S}^2$ into $\mathrm{SE}(3)$, a left-invariant distance $\Delta_{(\mathbb{R}^3\times\mathbb{S}^2),\kappa} : (\mathbb{R}^3 \times \mathbb{S}^2) \times (\mathbb{R}^3 \times \mathbb{S}^2) \to [0, \infty)$. For all $g_i = (\vec{v}_i, \vec{u}_i) \in \mathbb{R}^3 \times \mathbb{S}^2$, $i = 1, 2$, this distance is given by*

$$\Delta_{(\mathbb{R}^3\times\mathbb{S}^2),\kappa}(g_1, g_2) := |\left(\boldsymbol{R}_{\vec{u}_2}^{-1}(\vec{v}_1 - \vec{v}_2), \boldsymbol{R}_{\vec{u}_2}^{-1}\boldsymbol{R}_{\vec{u}_1}\right)|_{\mathrm{SE}(3),\kappa},$$

where $\boldsymbol{R}_{\vec{u}_i} \in \mathrm{SO}(3)$ is any rotation with $\boldsymbol{R}_{\vec{u}_i}\vec{e}_z = \vec{u}_i$ as introduced in Proposition 6.18.

The distance $\Delta_{(\mathbb{R}^3\times\mathbb{S}^2),\kappa}$ is defined on $\mathbb{R}^3 \times \mathbb{S}^2$, and it satisfies all required properties. For practical application, it will be approximated by the norm $\|.\|_2$ given in Definition 6.58. Here, the additional constant κ can be interpreted as a change of coordinates, setting

$$\hat{A}_i := \begin{cases} A_i & \text{if } i = 1,2,3, \\ \kappa A_i & \text{if } i = 4,5, \end{cases}$$

and

$$\hat{A}_6 := \left[\hat{A}_4, \hat{A}_5\right] = \kappa^2 \left[A_4, A_5\right] = \kappa^2 A_6.$$

Then, $\{A_i\}_{i=1}^6$ and $\{\hat{A}_i\}_{i=1}^6$ span isomorphic Lie algebras, and it holds

$$\sum_{i=1}^{6} \hat{\rho}_i \hat{A}_i = \sum_{i=1}^{6} \rho_i A_i \quad \text{and} \quad \prod_{i=1}^{6} \exp\left(\hat{\zeta}_i \hat{A}_i\right) = \prod_{i=1}^{6} \exp\left(\zeta_i A_i\right)$$

with

$$\hat{\rho}_i := \begin{cases} \rho_i & \text{if } i = 1,2,3, \\ \kappa^{-1}\rho_i & \text{if } i = 4,5, \\ \kappa^{-2}\rho_i & \text{if } i = 6, \end{cases} \quad \text{and} \quad \hat{\zeta}_i := \begin{cases} \zeta_i & \text{if } i = 1,2,3, \\ \kappa^{-1}\zeta_i & \text{if } i = 4,5, \\ \kappa^{-2}\zeta_i & \text{if } i = 6. \end{cases} \tag{6.30}$$

We introduce the following notation.

Notation 6.63. *Let $\boldsymbol{M}_g$, $g \in \mathrm{SE}(3)$, be the matrix representation of $\mathrm{SE}(3)$ in Equation* (6.26) *(page 104), and let $\{A_i\}_{i=1}^6$ denote the left-invariant basis matrices given in Example 6.52 (page 117). For $g_1, g_2 \in \mathbb{R}^3 \times \mathbb{S}^2$ with $g_i = (\vec{v}_i, \vec{u}_i)$, $i = 1, 2$, we define the distance*

$$\Delta_\kappa(g_1, g_2) := \inf\left\{ \left(\sum_{i=1}^{3} \zeta_i^2 + \kappa^{-2}(\zeta_4^2 + \zeta_5^2 + |\zeta_6|) \right)^{1/2} : \prod_{i=1}^{6} \exp(\zeta_i A_i|_{\hat{g}}) = \boldsymbol{M}_{\hat{g}} \right\}, \tag{6.31}$$

where $\hat{g} := \left(\boldsymbol{R}_{\vec{u}_2}^{-1}(\vec{v}_1 - \vec{v}_2), \boldsymbol{R}_{\vec{u}_2}^{-1}\boldsymbol{R}_{\vec{u}_1}\right) \in \mathrm{SE}(3)$ with $\boldsymbol{R}_{\vec{u}}\vec{e}_z = \vec{u}$.

This can be considered as an ad-hoc approximation of the modified Carnot-Carathéodory distance $\Delta_{(\mathbb{R}^3\times\mathbb{S}^2),\kappa}$ in Corollary 6.62, using the change of coordinates in Equation (6.30). Recall that this distance has been proven to be left-invariant and well-defined on the left coset space $\mathbb{R}^3 \rtimes \mathbb{S}^2$. The following considerations justify this rough approximation to some extent. In Proposition 6.60 (3), we stated, for nilpotent Lie groups, the equivalence of the Carnot-Carathéodory distance Δ and the norms $\|.\|_k$, $k \in \{1\} \cup 2\mathbb{N}$, in Definition 6.58. These norms remind of the well-known $\mathbb{L}_p$-norms, where $p = 2$ is often preferred since $\mathbb{L}_2$ forms a Hilbert space. To some extent, the missing nilpotence of $\mathrm{SE}(3)$ can be compensated by a homogeneous contraction, providing a nilpotent group $(\mathrm{SE}(3))_0$ which locally approximates $\mathrm{SE}(3)$, see Dungey et al. [2003, §II.6, § IV], Duits and van Almsick [2008], Duits et al. [2011], and ter Elst and Robinson [1998]. However, a detailed proof of the relation between Δ and $\|.\|_2$ is much beyond the scope of this study as it requires sophisticated concepts of differential geometry. Below, we will see that Equation (6.31) is closely related to the geodesics on $\mathbb{R}^3$ and $\mathbb{S}^2$. This follows by a slight modification, which will be motivated by the following computations.

We determine the coordinates $\{\zeta_i\}_{i=1}^6$ with

$$\prod_{i=1}^{6} \exp(\zeta_i A_i|_{\hat{g}}) = \begin{pmatrix} \mathbf{R}_{(\hat{\alpha},\hat{\beta},\hat{\gamma})} & \hat{v}^T \\ 0\ 0\ 0 & 1 \end{pmatrix} = \mathbf{M}_{\hat{g}} \equiv \hat{g} = \left(\mathbf{R}_{\vec{u}_2}^{-1}(\vec{v}_1 - \vec{v}_2), \mathbf{R}_{\vec{u}_2}^{-1}\mathbf{R}_{\vec{u}_1}\right).$$

In Lemma 6.57, we proved the Carnot-Carathéodory distance to be well-defined on the left coset space $\mathbb{R}^3 \rtimes \mathbb{S}^2$. Supposing that the approximation Δ_κ inherits (locally) this property, we can freely choose $\alpha_i \in (-\pi, \pi]$. Hence, we set $\alpha_1 = \alpha_2 = \hat{\alpha} = 0$, yielding, for the left-invariant basis $\{A_i|_g\}_{i=1}^6$, an easier form than given in Example 6.52 (page 117). Then, it holds

$$\begin{aligned} \hat{v} &= \mathbf{R}^{-1}_{(0,\beta_2,\gamma_2)}(\vec{v}_1 - \vec{v}_2), \\ \hat{\beta} &= \arcsin[\sin\beta_1 \cos\beta_2 - \cos\beta_1 \sin\beta_2 \cos(\gamma_1 - \gamma_2)], \\ \hat{\gamma} &= \arcsin[\cos\beta_1 \sin(\gamma_1 - \gamma_2)(\cos\hat{\beta})^{-1}]. \end{aligned}$$

Furthermore, we get

$$\prod_{i=1}^{3} \exp(\zeta_i A_i|_{\hat{g}}) = \begin{pmatrix} \mathbf{1} & \mathbf{R}_{(0,\hat{\beta},\hat{\gamma})} \cdot (\zeta_1, \zeta_2, \zeta_3)^T \\ \vec{0}^T & 1 \end{pmatrix}$$

and

$$\prod_{i=4}^{6} \exp(\zeta_i A_i|_{\hat{g}}) = \begin{pmatrix} \mathbf{N}_{(\zeta_4,\zeta_5,\zeta_6)} & \vec{0} \\ 0\ 0\ 0 & 1 \end{pmatrix},$$

where $\mathbf{N}_{(\zeta_4,\zeta_5,\zeta_6)}$ denotes an appropriate matrix which only depends on

$$A_4|_{\hat{g}} = (\cos\hat{\beta})^{-1}\mathbf{X}_4 - \tan\hat{\beta}\,\mathbf{X}_6, \qquad A_5|_{\hat{g}} = \mathbf{X}_5, \qquad A_6|_{\hat{g}} = \mathbf{X}_6.$$

It follows

$$(\zeta_1, \zeta_2, \zeta_3)^T = \mathbf{R}^{-1}_{(0,\hat{\beta},\hat{\gamma})}\hat{v} = \mathbf{R}^{-1}_{(0,\beta_1,\gamma_1)}(\vec{v}_1 - \vec{v}_2)$$

and consequently

$$\sum_{i=1}^{3} \zeta_i^2 = \|\mathbf{R}^{-1}_{(0,\beta_1,\gamma_1)}(\vec{v}_1 - \vec{v}_2)\|^2 = \|\vec{v}_1 - \vec{v}_2\|^2 \tag{6.32}$$

since $\mathbf{R}^T = \mathbf{R}^{-1}$ for all $\mathbf{R} \in \mathrm{SO}(3)$. The equation $\mathbf{N}_{(\zeta_4,\zeta_5,\zeta_6)} = \mathbf{R}_{(0,\hat{\beta},\hat{\gamma})}$ is solved numerically, yielding the coordinates ζ_4, ζ_5 and ζ_6 by minimizing $\zeta_4^2 + \zeta_5^2 + |\zeta_6|$ over the set of solutions.

6.4.3 Second proposal for a distance on the measurement space of dMRI

The results in the last subsection motivate the definition of a simplified distance on $\mathbb{R}^3 \times \mathbb{S}^2$. This accelerates the algorithm, and it avoids artifacts which may appear for our first proposal Δ_κ in Notation 6.63 as this is a rough approximation of the left-invariant and well-defined Carnot-Carathéodory distance only.

The Euclidean metric in Equation (6.32) as a distance in the voxel space complies with our intuition, but the term $\zeta_4^2 + \zeta_5^2 + |\zeta_6|$ violates the required rotation invariance in the gradient space, see Figure 6.13. Hence, we replace this term by the squared great circle distance, which equals the spherical geodesics [Jost, 2011, §1.4 pp. 25–27]. This distance is given by the radian, that is the arc length between the vectors $\vec{u}_1, \vec{u}_2 \in \mathbb{S}^2$ on the corresponding unit circle

$$d_{\mathbb{S}^2}(\vec{u}_1, \vec{u}_2) := \arccos\langle \vec{u}_1, \vec{u}_2 \rangle \in [0, \pi),$$

where $\langle .., . \rangle$ denotes the Euclidean scalar product, and the range of $\arccos$ is restricted to $[0, \pi)$, considering the real axis modulus π. In general, for dMRI data, one does not distinguish between one direction $\vec{u}$ and its opposite $-\vec{u}$ as justified by the (approximate) symmetry of the diffusion process. Therefore, we take the absolute value of the scalar product in $d_{\mathbb{S}^2}$, leading to the bijective function $\arccos : [0, 1) \to [0, \pi/2)$. Additionally, the distance was implemented, using the $\mathbb{L}_1$-analog instead of the $\mathbb{L}_2$-analog by just summing the corresponding metrics of $\mathbb{R}^3$ and $\mathbb{S}^2$, respectively. Other choices would be possible as well, since this choice has minor impact on the results. Finally, we introduce the following notation.

Notation 6.64. *For $g_1, g_2 \in \mathbb{R}^3 \times \mathbb{S}^2$ with $g_i := (\vec{v}_i, \vec{u}_i)$, $i = 1, 2$, we set*

$$\delta_\kappa(g_1, g_2) := \|\vec{v}_1 - \vec{v}_2\| + \kappa^{-1} \arccos|\langle \vec{u}_1, \vec{u}_2 \rangle|.$$

Disregarding the additionally introduced absolute value, the same distance function was proposed by Hagmann et al. [2006]. The distance δ_κ in Notation 6.64 does not rely on the embedding of $\mathbb{R}^3 \times \mathbb{S}^2$ into $\mathrm{SE}(3)$. Therefore, it is not necessary that it is well-defined with respect to the embedding, but we still require left-invariance. The following theoretical properties justify the appropriateness of the distance δ_κ. The associated proof is given in Section 6.7.

Proposition 6.65. *The function $\delta_\kappa : (\mathbb{R}^3 \times \mathbb{S}^2) \times (\mathbb{R}^3 \times \mathbb{S}^2) \to [0, \infty)$ in Notation 6.64 is a left-invariant pseudometric on $\mathbb{R}^3 \times \mathbb{S}^2$ in terms of Notation 6.36. In particular, it holds*

$$\delta_\kappa[g_1, g_2] = 0 \quad \textit{if and only if} \quad g_1 = g_2 \textit{ or } g_1 = (\vec{v}, \vec{u}) \textit{ and } g_2 = (\vec{v}, -\vec{u}).$$

Recall that, in contrast to a metric, a pseudometric allows the case $\delta_\kappa(g_1, g_2) = 0$ with $g_1 \neq g_2$. In Figure 6.13, we compare our two proposals, Δ_κ in Notation 6.63 and δ_κ in Notation 6.64. This shows that Δ_κ violates the required rotation invariance on the gradient space, yielding different results for pairs of gradients with a constant angle. This gets worse the larger the distance Δ_κ.

6.5 Numerical results

We applied the algorithms POAS and msPOAS to simulated and experimental dMRI data. First, we will describe the considered data sets and applied methods, before showing the smoothing results in § 6.5.3. A detailed discussion can be found in Section 6.6.

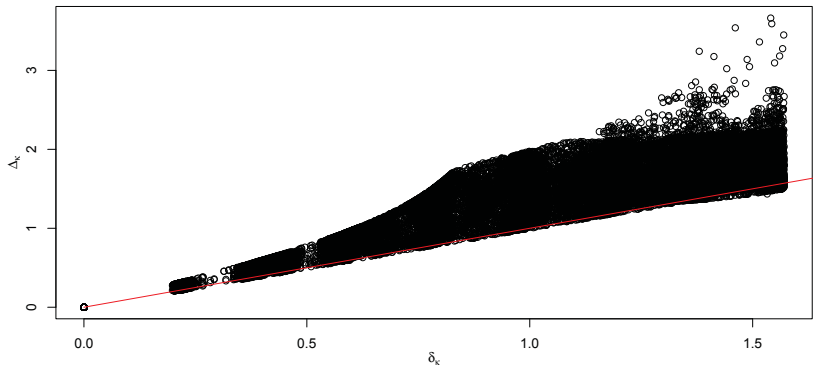

Figure 6.13: Pairwise scatterplot of the distances δ_κ and Δ_κ for $\kappa = 0.5$. We concentrate on the spherical parts of the two distances, setting $\|\vec{v}_1 - \vec{v}_2\| = 0$. The distance δ_κ equals by definition the angle between two gradients, which is rotation invariant as established in Proposition 6.65. In contrast, Δ_κ takes different values for varying pairs of gradients with constant angle. Each point corresponds to a gradient pair out of 150 gradients. The line where δ_κ equals Δ_κ is shown in red. Recall that the range of δ_κ equals $[0, \pi/2)$, while Δ_κ can take larger values.

6.5.1 Data

In this subsection, we describe the simulated and experimental data in more detail. First, we investigated two artificial diffusion-weighted data sets by J. Polzehl (WIAS Berlin), where a receiver coil system with $L' = 8$ coils is mimicked. We emphasize that the data generation by a specific diffusion model, see § 6.1.3, does not work in favor of (ms)POAS since the model does not effect the smoothing procedure. The experimental data contain one single-shell and one multi-shell data set. Both data sets have a very low signal-to-noise ratio due to their high spatial resolution and the large b-value in the outer shell of the second data set.

Artificial data 1. For our first example, we consider a voxel space with $32 \times 32 \times 32$ voxels. The 42 gradient directions were chosen to minimize symmetrized Coulomb forces on a sphere as proposed by Jones et al. [1999]. Using a tensor mixture model, see § 6.1.3 and Tabelow et al. [2012], we created a data set with fiber bundles along the x- and y-coordinate axes. Each bundle completely crosses the cube with a width of four voxels and intermediate areas between the bundles of again four voxels width. Hence, there are 22528 voxels with two fiber bundles, 8320 voxels with one fiber bundle, and 1920 voxels without fibers. Single fiber bundles are characterized by the typical diffusion tensor for white matter, choosing the eigenvalues as $(1.4, 0.35, 0.35) \cdot 10^{-3}\,\mathrm{mm}^2/\mathrm{s}$, see Alexander et al. [2001]. The SNR equals 10. In Figure 6.16 (page 129), we show the exact data (a) and the noisy data (b).

Artificial data 2. For our second data set, we created signals for $32 \times 32 \times 11$ voxels and again 42 gradient directions. In each voxel, we simulated two fiber directions by a tensor mixture model of order two, again using prolate tensors with eigenvalues $(1.4, 0.35, 0.35) \cdot 10^{-3}\,\mathrm{mm}^2/\mathrm{s}$. The data are separated into two homogeneous regions. The volume fractions of the fibers in the first region are 0.4 and 0.6, where the tensors are oriented along the x- and y-axes. In the second region, the volume fractions equal 0.5, and the tensor orientations are chosen such that both regions coincide with respect to the corresponding diffusion tensor model. Hence, the two regions contain different fiber crossings which yield the same tensor when fitting a diffusion tensor model to noiseless data. This is the worst case for smoothing methods which use the

diffusion tensor model for adaptation. Here, we simulated an SNR equal to 32. The data set is shown in Figure 6.17 (page 130) for exact data (a) and for noisy data (b).

Experimental data 1 (single-shell). Our first experimental data set was provided by A. Anwander and R.M. Heidemann (Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany). The technical details are the following. The MR experiment was performed on a 7T whole body MR scanner (MAGNETOM 7T, Siemens Healthcare, Erlangen, Germany). For signal reception, a single channel transmit with a 24-channel receive phased array head coil (Nova Medical, Wilmington, MA, USA) was used. The scans were performed on a healthy adult volunteer in one session. Written informed consent was obtained in accordance with the ethical approval from the University of Leipzig. The data were acquired, using an optimized monopolar Stejskal-Tanner sequence [Morelli et al., 2010] together with the ZOOPPA approach by Heidemann et al. [2012]. Here, 91 slices were measured with 10% overlap and $800\,\mu\mathrm{m}$ isotropic resolution at a field-of-view (FoV) of $143 \times 147\,\mathrm{mm}^2$, setting TR$= 14.1\,\mathrm{s}$, TE$= 65\,\mathrm{ms}$, BW$= 1132\,\mathrm{Hz/pixel}$, and ZOOPPA acceleration factor 4.6. For diffusion-weighting, 60 diffusion-weighting gradients were applied at a b-value of $1000\,\mathrm{s/mm}^2$. Additionally, seven interspersed S_0-images were acquired. For averaging, the scan was repeated four times, leading to a total acquisition time of 65 minutes.

Experimental data 2 (multi-shell). Our second experimental data set was acquired by N. Weiskopf and S. Mohammadi (Wellcome Trust Centre for Neuroimaging, London, UK) on a 3T MAGNETOM Trio scanner (Siemens AG, Healthcare Sector, Erlangen, Germany). One healthy adult volunteer participated in the study approved by the local ethics committee after giving written informed consent. Here, a reduced field-of-view technique by Heidemann et al. [2010] was used, considering $161 \times 58\,\mathrm{mm}$ centered about the motor cortex. Measuring 34 slices of $1.2\,\mathrm{mm}$ slice thickness with 10% gap, this resulted in an isotropic in-plane resolution of $1.2\,\mathrm{mm}$. For diffusion-weighting, two different b-values were applied, $b = 800\,\mathrm{s/mm}^2$ and $b = 2000\,\mathrm{s/mm}^2$, each with 100 gradient directions as suggested by Caruyer et al. [2011]. Additionally, 21 interspersed S_0-images were acquired. The total scan time was 22 minutes, setting TR$= 6.1\,\mathrm{s}$, TE$= 97\,\mathrm{ms}$.

6.5.2 Methods

The artificial data sets were smoothed by the POAS method. To the experimental data, we applied both methods, POAS and msPOAS. The algorithm (ms)POAS requires the effective number of coils $L' \in \mathbb{N}$ and the noise standard deviation $\sigma > 0$ to be homogeneous and known. The former is very difficult to estimate from the data. Fortunately, (ms)POAS is relatively robust against misspecification of L' as we will show in Figure 6.20 (page 133). For the first experimental data set, we mimicked an average influence of two coils, setting $L' = 2$. For the second experimental data set, we compared the results of msPOAS for $L' = 1, 4, 16$, and we applied POAS with $L' = 4$. The standard deviation σ was estimated by the method described in [Becker et al., 2012, App. C]. These estimates are rather independent of L', consistently yielding values around 75 for the first data set and values around 30 for the second data set. A comparison with the methods "Bk-M1-χ" and "Bk-M2-χ" in [Aja-Fernández et al., 2009] can be found in [Becker et al., 2013, §3.2]. On the single-shell data, POAS was applied with $\sigma = 66$ since this value resulted from a previous variance estimation in [Becker et al., 2012].

The method parameters were fixed in agreement with § 6.2.2. For all examples, we set $h^{(0)} = 1$. For the experimental data, we used $k^* = 12$, while the first artificial example was stopped at $k^* = 18$ and the second one at $k^* = 25$. The balancing parameter κ_0 was chosen in dependence

of the mean number of gradient directions per q-shell, where 0.5 for the first and 0.3 for the second data set have shown best performance of msPOAS. For POAS, we increased the first value since Δ_κ yields (on average) larger values than δ_κ, especially for large angles between the involved gradients, see Figure 6.13 (page 123). Therefore, we used for POAS $\kappa_0 = 0.6$ for the first and $\kappa_0 = 0.3$ for the second experimental data set. To the artificial data sets, we applied POAS with $\kappa_0 = 0.6$.

We will see in § 6.5.3 that the choice of the adaptation bandwidth given by the propagation condition is almost robust against the number of applied diffusion-weighting gradients and the degrees of freedom $2L'$. For msPOAS, we used, for all data sets, the bandwidth $\lambda = 20$. Recall that the implementation of the propagation condition in the **R**-package **dti** by Tabelow and Polzehl [2013] is based on the msPOAS procedure, see Algorithm 3 (page 97), using the simplified approximation of the Kullback-Leibler divergence in Equation (6.19) (page 98), the modified distance in Notation 6.64, and a weighted arithmetic mean for the estimator. For this reason, we did not choose the adaptation bandwidth of POAS by the propagation condition. Instead, we used the value of λ which provided optimal results of POAS such that a fair comparison between POAS and msPOAS could be made. More precisely, we set $\lambda = 6$ for the multi-shell data, while using $\lambda = 10$ for the first experimental and both artificial data sets.

For comparison, we additionally applied the structural adaptive smoothing method by Tabelow et al. [2008], which can be considered as a previous version of POAS. It is as well based on the Propagation-Separation Approach, using the diffusion tensor model for adaptation and the Euclidean metric in the voxel space for location. We applied this method to the second artificial and the first experimental data sets in order to show how (ms)POAS outperforms this previous approach. Here, we used the default parameters of the implementation in the **R**-package **dti**, setting $\lambda := 25$ and $\texttt{hmax} := 4$ as an upper bound of the location bandwidth $h^{(k^*)}$ for the artificial example and $\texttt{hmax} := 3$ for the experimental data. In order to avoid confusion, we refer to this method as DTI-smoothing since it is based on the diffusion tensor model.

The results are presented in various forms. For the first artificial and the first experimental data sets, we estimated the diffusion tensor model, see § 6.1.3, using a non-linear method by Polzehl and Tabelow [2009]. Here, we show the corresponding color-coded FA maps, where the fractional anisotropy (FA) is defined in Equation (6.4) (page 87), and the three colors, namely red, blue, and green, refer to the main fiber direction. The second artificial data set is illustrated via the resulting orientation distribution functions (ODF) estimated in a tensor mixture model [Tabelow et al., 2012]. Here, other models could have been chosen for visualization with similar results. For the second experimental data set, we present the diffusion-weighted images. Additionally, we calculated fiber tracks, using a streamline FACT algorithm by Mori et al. [1999], which is implemented in the package **dti**. Again, this is based on the diffusion tensor model.

Further examples can be found in [Becker et al., 2012, 2013]. There, we additionally investigated a phantom data set [Becker et al., 2012, Fig. 7] and the reduction of the mean angular deviation in a 1-stick-1-ball model [Becker et al., 2013, Fig. 10] for the multi-shell data set described above. For the single-shell data set, the original and smoothed diffusion-weighted images and corresponding fiber tracks are presented in [Becker et al., 2012, Fig. 9&10]. Finally, we refer to the univariate examples in Figures 5.12 (page 74) and 5.13, where we illustrated the general behavior of the Propagation-Separation Approach in the case of non-central chi-distributed observations. The algorithm differs from msPOAS only by the implemented design spaces and the corresponding distances.

For data preprocessing, motion and eddy-current correction of the second experimental data set, we applied SPM by Friston et al. [2006] and the ACID-toolbox by Mohammadi et al. [2010]. To the first experimental data set, we did not apply any method for pre-processing, instead we used the raw data. The computations of (ms)POAS, the DTI-smoothing method by Tabelow et al. [2008], the diffusion tensor estimates, the FA maps, the fiber tracks, and the propagation condition were performed with the **R**-package **dti** by Tabelow and Polzehl [2013], see also Polzehl and Tabelow [2011]. In this implementation, processing of the second experimental data set by msPOAS took 15 minutes on a single core of a HP SL390s compute server with an Intel Xeon, Six-Core 3467 MHz. Applying POAS to each shell separately required more than one hour. The single-shell data set is much larger. Here, msPOAS used 3 hours and 18 minutes, while POAS required 4 hours and 37 minutes computation time on a single core of the same machine. The implementation is parallelized, using OpenMP. This significantly speeds up the computation compared with the mentioned single core results.

6.5.3 Results

First, we consider the results of the propagation condition for the choice of the adaptation bandwidth λ, that we described in § 6.2.2. The **R**-package **dti** provides by the function `dwi.smooth.testprop` plots of the isolines of the probability p with respect to the iteration steps $k = 0, ..., k^*$ and the corresponding values $z = \mathfrak{Z}_\lambda(k, p, \theta_0, m)$, where $m \in \mathbb{R}^3 \times \mathbb{S}^2$ is fixed and θ_0 denotes the homogeneous non-centrality parameter on the simulated S_0-image. The isolines are plotted for all $p \in \{v \cdot 10^{-w} : v = 1, 2, 5, w = 1, 2, 3, 4, 5\}$. Therefore, the propagation levels, which we report in the following, are restricted to this set of probability values.

In Figure 6.14, we illustrate the results of our final choice $\lambda = 20$, using the same parameters (if possible) as for our experimental data sets. More precisely, we simulated single-shell data with b-value $b_1 = (1000)$ and gradn $= 60$ diffusion-weighting gradients. Additionally, we simulated multi-shell data with $b_2 = (800, 2000)$ and gradn $= 100$ gradients on each q-shell. The voxel space was created with an isotropic extension of $n = 80$, providing $n^3 = 512000$ voxels. Then, we set $\kappa_0 = 0.5$ for the single-shell and $\kappa_0 = 0.3$ for the multi-shell data. The maximal number of iterations $k^* := 20$ is larger than for our subsequent examples in order to ensure a reliable propagation level. The effective number of receiver coils was chosen as $L' = 2$ and the non-centrality parameter of the S_0-image as $\theta_0 = 16$, which yields $\theta_b = (4.36)$ for $b = (1000)$ and $\theta_b = (5.66, 1.19)$ for $b = (800, 2000)$, respectively.

For $\lambda = 20$, the resulting propagation level of the single-shell data equals $\epsilon_0 \approx 5 \cdot 10^{-5}$ for the S_0-image and $\epsilon_b \leq 10^{-5}$ for the diffusion-weighted image at b-value $b_1 = (1000)$. For the multi-shell data, we got the same values, that is $\epsilon_0 \approx 5 \cdot 10^{-5}$ for the S_0-image and $\epsilon_b \leq 10^{-5}$ for the combined diffusion-weighted image at b-values $b_2 = (800, 2000)$, see Equation (6.21) (page 100). Note, that the slight increase of the z-values can be neglected. Probably, it results from the approximation of the considered probabilities $\mathbb{P}(\sum_{b \in B_0} \overline{N}^{(k)}_{m,b} \mathcal{KL}(\tilde{\theta}^{(k)}_b(m, \lambda), \theta_b) > z)$.

Next we evaluate the variability of the propagation level with respect to the fixed parameter θ_0,the effective number of receiver coils L', the realization of the sample given by the argument `seed`, the number of diffusion-weighting gradients `gradn` per q-shell, and the number of q-shells. The corresponding results are summarized in Table 6.1. Here, we consider a varying number of q-shells, choosing $b_1 = (800)$, $b_2 = (800, 2000)$, $b_3 = (800, 1400, 2000)$, and, in the Subtables (a) and (b), $b'_1 = (2000)$. If not mentioned differently, the other parameters were fixed at $\lambda = 15$, $\theta_0 = 16$, `gradn` $= 60$, $\kappa_0 = 0.5$, $L' = 2$, and `seed` $= 1$. In the Subtables (a)

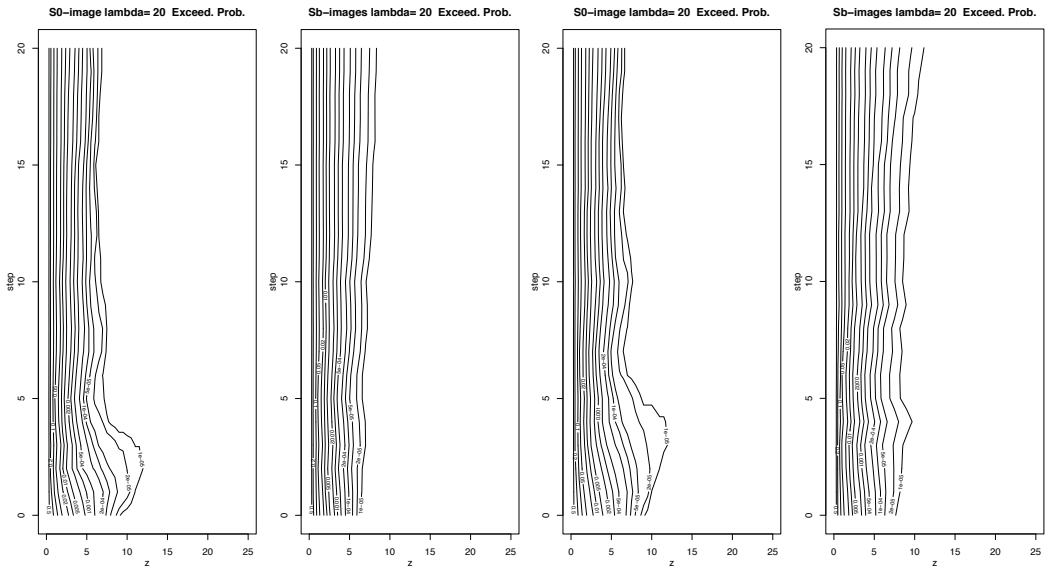

Figure 6.14: Results of the propagation condition (f.l.t.r): Single-shell data on the S_0- and the S_b-images, multi-shell data on the S_0- and the combined S_b-images. The isolines of the probability p for values between 10^{-5} and 0.5 are plotted with respect to the iteration steps $k = 0, ..., k^*$ and the corresponding values $z = \mathfrak{Z}_\lambda(k, p, \theta_0, m)$, where $\theta_0 := 16$ and $m \in \mathbb{R}^3 \times \mathbb{S}^2$ is fixed.

and (b), we report pairs of values (ϵ_b, p^*), where p^* equals the minimal probability which satisfies $\mathfrak{Z}_\lambda(k^*, p, \theta_0, m) \approx \mathfrak{Z}_\lambda(0, p, \theta_0, m)$, where the corresponding isoline does not need to be non-increasing. This generalized propagation level is usually sufficient for practical purposes. In the Subtables (a) and (b), we report both propagation levels, ϵ_b and ϵ_0. Else, we concentrate on the propagation levels ϵ_b of the (combined) diffusion-weighted images since (ms)POAS is constructed especially for smoothing on $\mathbb{R}^3 \times \mathbb{S}^2$. In the Subtables (e) and (f), we adjusted κ_0 in dependence of `gradn` in order to enable a fair comparison. Pursuant to § 6.2.2, we used the formula $\kappa_0 = \arccos(1 - u/\texttt{gradn})$, where $u := 60[1 - \cos(0.5)] \approx 7.345$ follows from our default choices with $\texttt{gradn} = 60$ and $\kappa_0 = 0.5$. Additionally, for $\texttt{gradn} = 60$, we study the robustness of the propagation level with respect to κ_0 by reporting the propagation levels for the extreme choices $\kappa_0 = 0.01$, where $u \approx 0.003$ and $\kappa_0 = 0.8$, yielding $u \approx 18.2$. Recall that very small values of κ_0 restrict (ms)POAS to smoothing on $\mathbb{R}^3$.

We observe that the propagation level is quite robust against the fixed parameter θ_0, the effective number of receiver coils L', the realization `seed` of the sample, and the number of diffusion-weighting gradients `gradn` per q-shell. Only for varying numbers of q-shells and the choice $\kappa_0 = 0.01$, we get a certain variability of the corresponding propagation level. Therefore, we recommend to choose the adaptation bandwidth on an artificial data set whose number of q-shells and corresponding b-values correspond with the data at hand. Fortunately, these values are usually known. The function $\mathfrak{Z}_\lambda(k, p, \theta_0, m)$, $p \in \{\epsilon_0, \epsilon_b, p_0^*, p_b^*\}$, provided, for all examples, values within the interval $[3, 10]$.

Before we present the smoothing results on the simulated and real data sets, we illustrate the behavior of the distances on $\mathbb{R}^3 \times \mathbb{S}^2$ which we introduced in Section 6.4. For this purpose, we show in Figure 6.15 the non-adaptive weighting schemes $\overline{w}_{ij}^{(k)} = K_{\mathrm{loc}}(\Delta_\kappa(m, n)/h^{(k)})$ of the first experimental data set, where Δ_κ is as in Notation 6.63, and $m = (\vec{v}_m, \vec{g}_m)$ is fixed. The balancing parameter $\kappa = \kappa_0/h^{(k)}$ and the location bandwidths $\{h^{(k)}\}_{k=0}^{k^*}$ are chosen in agreement with § 6.2.2. In order to demonstrate the impact of the balancing parameter and the

θ_0	8	32	160	1600
b_1	$(0.001, 2\cdot10^{-4})$	$(0.001, 2\cdot10^{-4})$	$(0.001, 2\cdot10^{-4})$	$(0.001, 2\cdot10^{-4})$
b_1'	$(2\cdot10^{-4}, 5\cdot10^{-5})$	$(5\cdot10^{-4}, 10^{-4})$	$(5\cdot10^{-4}, 10^{-4})$	$(5\cdot10^{-4}, 10^{-4})$
b_2	$(0.001, 5\cdot10^{-4})$	$(0.002, 5\cdot10^{-4})$	$(0.002, 5\cdot10^{-4})$	$(0.002, 5\cdot10^{-4})$
b_3	$(0.005, 0.001)$	$(0.01, 0.002)$	$(0.01, 0.002)$	$(0.01, 0.002)$
b_1 on S_0	$(5\cdot10^{-4}, 10^{-4})$	$(5\cdot10^{-4}, 5\cdot10^{-5})$	$(5\cdot10^{-4}, 5\cdot10^{-5})$	$(5\cdot10^{-4}, 10^{-4})$
b_1' on S_0	$(5\cdot10^{-4}, 10^{-4})$	$(5\cdot10^{-4}, 5\cdot10^{-5})$	$(5\cdot10^{-4}, 5\cdot10^{-5})$	$(5\cdot10^{-4}, 5\cdot10^{-5})$
b_2 on S_0	$(5\cdot10^{-4}, 10^{-4})$	$(5\cdot10^{-4}, 5\cdot10^{-5})$	$(5\cdot10^{-4}, 10^{-4})$	$(5\cdot10^{-4}, 10^{-4})$
b_3 on S_0	$(5\cdot10^{-4}, 10^{-4})$	$(5\cdot10^{-4}, 5\cdot10^{-5})$	$(5\cdot10^{-4}, 10^{-4})$	$(5\cdot10^{-4}, 10^{-4})$

(a) (ϵ_b, p^*) with $\lambda = 15$ and varying θ_0.

θ_0	8	32	160	1600
b_1	$(10^{-5}, 10^{-5})$	$(10^{-5}, 10^{-5})$	$(10^{-5}, 10^{-5})$	$(10^{-5}, 10^{-5})$
b_1'	$(10^{-5}, 10^{-5})$	$(10^{-5}, 10^{-5})$	$(10^{-5}, 10^{-5})$	$(10^{-5}, 10^{-5})$
b_2	$(10^{-5}, 10^{-5})$	$(10^{-5}, 10^{-5})$	$(10^{-5}, 10^{-5})$	$(10^{-5}, 10^{-5})$
b_3	$(2\cdot10^{-4}, 2\cdot10^{-5})$	$(5\cdot10^{-4}, 5\cdot10^{-5})$	$(5\cdot10^{-4}, 5\cdot10^{-5})$	$(5\cdot10^{-4}, 5\cdot10^{-5})$
S_0 (all b_l)	$(5\cdot10^{-5}, 10^{-5})$	$(5\cdot10^{-5}, 10^{-5})$	$(5\cdot10^{-5}, 10^{-5})$	$(5\cdot10^{-5}, 10^{-5})$

(b) (ϵ_b, p^*) with $\lambda = 20$ and varying θ_0.

seed	13	83	987	2407
b_1	$5\cdot10^{-4}$	$5\cdot10^{-4}$	0.001	$5\cdot10^{-4}$
b_2	0.001	0.001	0.001	0.001
b_3	0.01	0.01	0.01	0.01

(c) ϵ_b with $\lambda = 15$ and varying realizations.

L'	1	4	8	16
b_1	$5\cdot10^{-4}$	$5\cdot10^{-4}$	$5\cdot10^{-4}$	$5\cdot10^{-4}$
b_2	0.002	0.005	0.002	0.005
b_3	0.005	0.005	0.005	0.005

(d) ϵ_b with $\lambda = 15$ and varying L'.

`gradn`	40	60	80	100	60	60	min	max
κ_0	0.616	0.5	0.432	0.386	0.01	0.8		
b_1	$5\cdot10^{-4}$	$5\cdot10^{-4}$	$5\cdot10^{-4}$	$5\cdot10^{-4}$	0.001	$5\cdot10^{-4}$	$5\cdot10^{-4}$	0.001
b_2	0.001	0.001	0.001	0.001	0.002	0.001	0.001	0.002
b_3	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005

(e) ϵ_b with $\lambda = 15$ and varying `gradn`.

`gradn`	40	60	80	100	60	60	min	max
κ_0	0.616	0.5	0.432	0.386	0.01	0.8		
b_1	10^{-5}	10^{-5}	10^{-5}	10^{-5}	10^{-4}	10^{-5}	10^{-5}	10^{-4}
b_2	10^{-5}	10^{-5}	10^{-5}	10^{-5}	$2\cdot10^{-4}$	$2\cdot10^{-5}$	10^{-5}	$2\cdot10^{-4}$
b_3	$5\cdot10^{-5}$	$5\cdot10^{-5}$	$5\cdot10^{-5}$	$5\cdot10^{-5}$	$5\cdot10^{-4}$	$5\cdot10^{-5}$	$5\cdot10^{-5}$	$5\cdot10^{-4}$

(f) ϵ_b with $\lambda = 20$ and varying `gradn`.

Table 6.1: Propagation levels ϵ_b for the combined S_b-images, where the smallest reported propagation level is 10^{-5}. In the Subtables (a) and (b), we report pairs of values (ϵ_b, p^*), where p^* equals the minimal probability which satisfies $\mathfrak{Z}_\lambda(k^*, p, \theta_0, m) \approx \mathfrak{Z}_\lambda(0, p, \theta_0, m)$. Moreover, we give the corresponding results of the S_0-image.

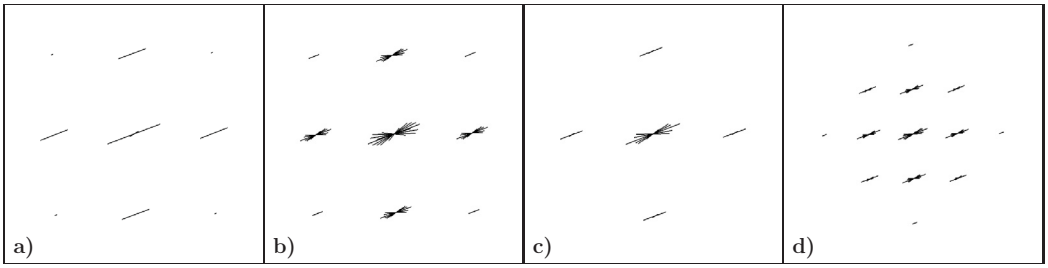

Figure 6.15: Non-adaptive weighting schemes for a fixed point $m = (\vec{v}_m, \vec{g}_m) \in \mathbb{R}^3 \times \mathbb{S}^2$, using the distance Δ_κ in Notation 6.63. The length of lines represents the weight, the direction corresponds to the respective gradient direction $\vec{g}_n$ in $n \in \mathbb{R}^3 \times \mathbb{S}^2$, and the location follows from the Euclidean distance $\|\vec{v}_m - \vec{v}_n\|$. From left to right: $k^* = 8$ with $\kappa_0 = 0.3, 0.7$ (a+b), and $\kappa_0 = 0.5$ with number of iterations $k^* = 4, 12$ (c+d).

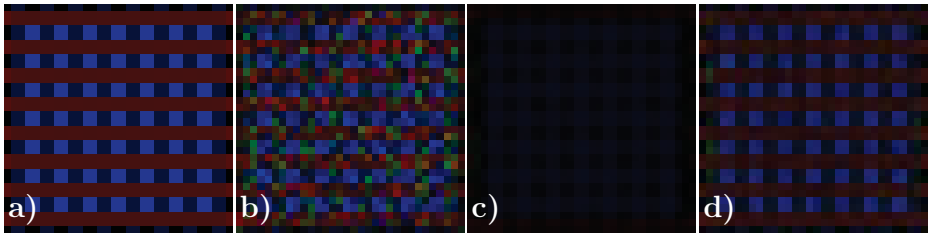

Figure 6.16: Color-coded FA maps of the first artificial example a) for exact data, b) for noisy data, c) after smoothing the noisy data with POAS, but using non-adaptive weights only, d) after smoothing the noisy data with POAS, using adaptive weights.

behavior during iteration, we show the weighting schemes for several bandwidths and varying values of κ_0. The distance Δ_κ behaves as expected, yielding for all voxels with the same spatial distance to voxel $\vec{v}_i$ the same amount of smoothing on the sphere. This amount decreases with increasing spatial distance of the involved voxels. Within each voxel, we observe monotonically decreasing weights on the sphere. This property can be violated for larger angles between the involved gradients due to the missing rotation invariance of Δ_κ, see Figure 6.13 (page 123). The same figure shows that the distances δ_κ and Δ_κ approximate each other if the angles between the involved gradients are sufficiently small. Therefore, we do not show the weighting schemes of the distance δ_κ, which look very similar to Figure 6.15.

Next we consider the results for the first artificial example. Figure 6.16 illustrates the corresponding color-coded FA maps for a central slice of the exact, the noisy and the smoothed data. For comparison, we additionally show the results of non-adaptive smoothing, where POAS was applied, setting $\lambda = 2e^{20}$. This clearly indicates a loss of information due to blurring.

In Figure 6.17, we concentrate on the difference between the POAS algorithm and the DTI-smoothing method by Tabelow et al. [2008]. Here, we observe, for the second artificial data set, that POAS removes the distortions by the noise without blurring the structural border, while DTI-smoothing lacks sensitivity at the discontinuity, leading to similar results as non-adaptive smoothing. We recall that this method uses the diffusion tensor model for adaptation. Therefore, the two regions cannot be distinguished in this special situation. The differences between POAS and the approach by Tabelow et al. [2008] can be quantified by estimating the standardized mean absolute error of the smoothed signals $\tilde{S}^{(k^*)}(m)$ adjacent to the structural border, that is $|\mathcal{M}|^{-1} \sum_{m \in \mathcal{M}} |\tilde{S}^{(k^*)}(m) - \mathbb{E}S(m)|/\sigma$, where $\mathcal{M}$ is the set of all points $m = (\vec{v}_m, \vec{g}_m) \in$

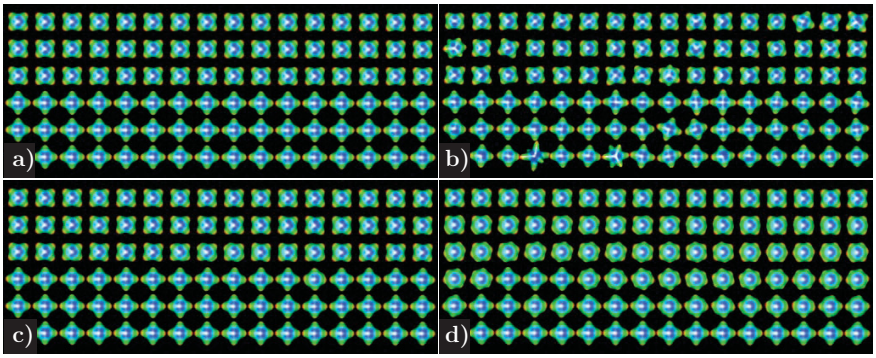

Figure 6.17: Orientation distribution functions (ODF) of the second artificial example, estimated in a tensor mixture model a) for exact data, b) for noisy data, c) after smoothing with POAS, d) after DTI-smoothing.

$k^\star$	0	4	8	12	16	20	25	30
POAS	0.902	0.359	0.277	0.233	0.186	0.156	0.133	0.122
DTI-smoothing	0.902	0.446	0.291	0.241	0.240	0.250	0.264	0.272

Table 6.2: Standardized mean absolute error of the smoothed signals $\tilde{S}^{(k^*)}(m)$ adjacent to the structural border for POAS and the DTI-smoothing method by Tabelow et al. [2008] with varying number of iteration steps k^*.

$\mathbb{R}^3 \times \mathbb{S}^2$ where $\vec{v}_m$ is located next to the structural border, and $|\mathcal{M}|$ denotes its cardinality. The results are summarized in Table 6.2. Recall that the mean absolute error provides better robustness than the mean squared error.

The first experimental data set is special due to its high spatial resolution which leads to a very low SNR. In Figure 6.18, we show the color-coded FA maps of some axial slice. As for the simulated data, we observe that POAS reduces the noise without blurring effect on borders (c). The algorithm msPOAS provides very similar results (e), while the DTI-smoothing by Tabelow et al. [2008] is not able to remove much of the noise (g). Obviously, the diffusion tensor model fails for very low SNR. For comparison, we use the averaged image from the four repeated scans as a kind of ground truth (d). By visual inspection, the smoothing results of POAS and msPOAS compare well with the mean image of the repeated data, which required a four times longer acquisition time. For further improvement of the SNR, we may smooth also the averaged data, here shown after smoothing with POAS (b) and the DTI-smoothing (h). An alternative smoothing method which we will discuss in § 6.6.5 is the Lohmann filter [Lohmann et al., 2010]. This was applied by the originators to the first experimental data set, see the result in (f).

Additionally, in Figure 6.19, we visualize the behavior during iteration, showing the results for increasing iteration steps in a region of interest in the same slice as chosen for Figure 6.18. As for the univariate examples in Section 5.3.5, we observe that the estimates are forced into the local constant model for large iteration steps.

In Figure 6.20, we provide the results for the multi-shell data set, presenting the diffusion-weighted images before and after smoothing with msPOAS and POAS. Simultaneously, we demonstrate that the result of msPOAS is rather robust against misspecification of the effective number of coils L'. The adaptive smoothing effect of msPOAS is apparent for both shells, while separate application of POAS leads to blurring on the outer shell at $b = 2000\,\mathrm{s/mm^2}$ when

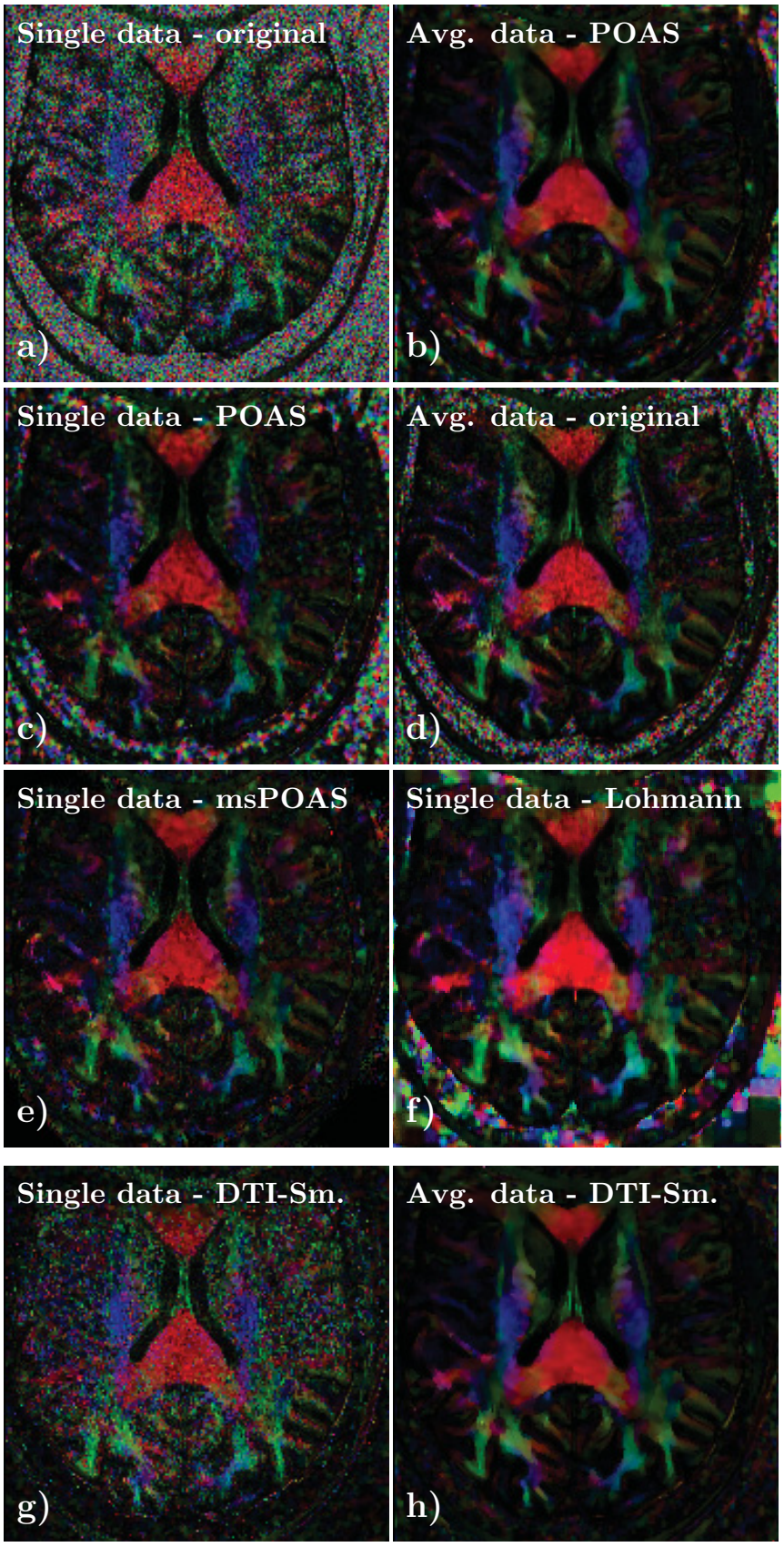

Figure 6.18: Color-coded FA maps of the single-shell data in some arbitrarily selected slice.

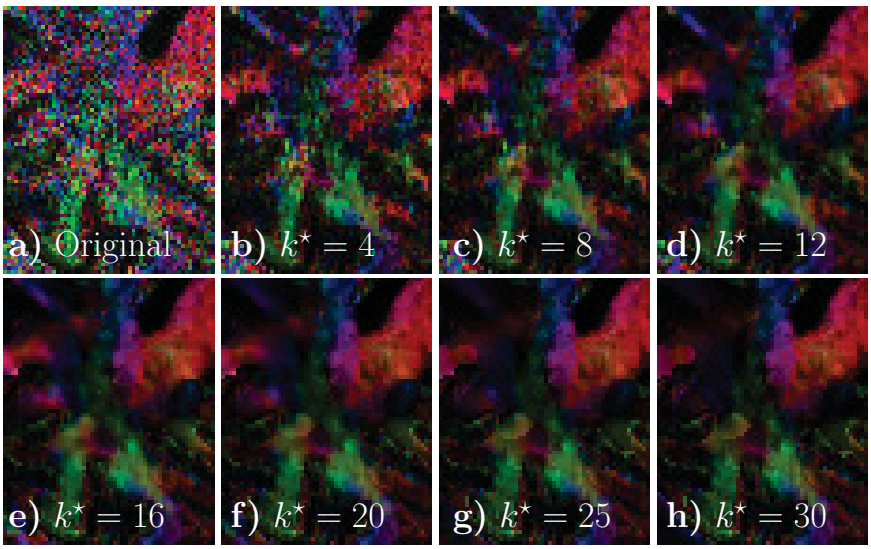

Figure 6.19: Visualization of the behavior of POAS during iteration for a selected region of interest from the slice in Figure 6.18.

choosing $\lambda = 6$ for POAS, where the inner shell at $b = 800\,\mathrm{s/mm^2}$ resembles the msPOAS results. Consequently, msPOAS indeed leads to improved results compared to the single-shell POAS approach. This is additionally confirmed by Figure 6.21, where we show the fiber track reconstructions for the multi-shell data after smoothing with msPOAS and POAS. After msPOAS, the reconstruction of the fibers even with this very simple algorithm is much richer than the one obtained from separate smoothing of each shell with POAS. The occurrence of the U-Fibers at the outer parts of the fiber tracks after msPOAS coincide with the anatomic structure in the brain.

6.6 Discussion of the application

We have developed a novel algorithm, called (ms)POAS, for noise reduction in diffusion-weighted MRI data. In this section, we will discuss its properties, alternatives, and future trends. First, we will summarize the advantages and challenges of our procedure. Then, we will analyze the impact of the parameters, and we will compare the results of the single-shell position-orientation adaptive smoothing method (POAS) with its generalization to multi-shell data, namely msPOAS. The crucial point of our approach is the description of the diffusion-weighted data as a standard orientation score, that is as a real valued function on the measurement space $\mathbb{R}^3 \times \mathbb{S}^2$ of the (voxel) positions and (gradient) directions. Therefore, in § 6.6.4, we will consider alternative perspectives on diffusion-weighted data. Moreover, we will give a brief overview of the current smoothing methods for dMRI and their differences to (ms)POAS. We will close with some proposals for future research.

6.6.1 Advantages and challenges of (ms)POAS

The power of POAS and msPOAS lies in their structural adaptivity. Both methods are based on the Propagation-Separation Approach. In an iterative procedure they accumulate information on

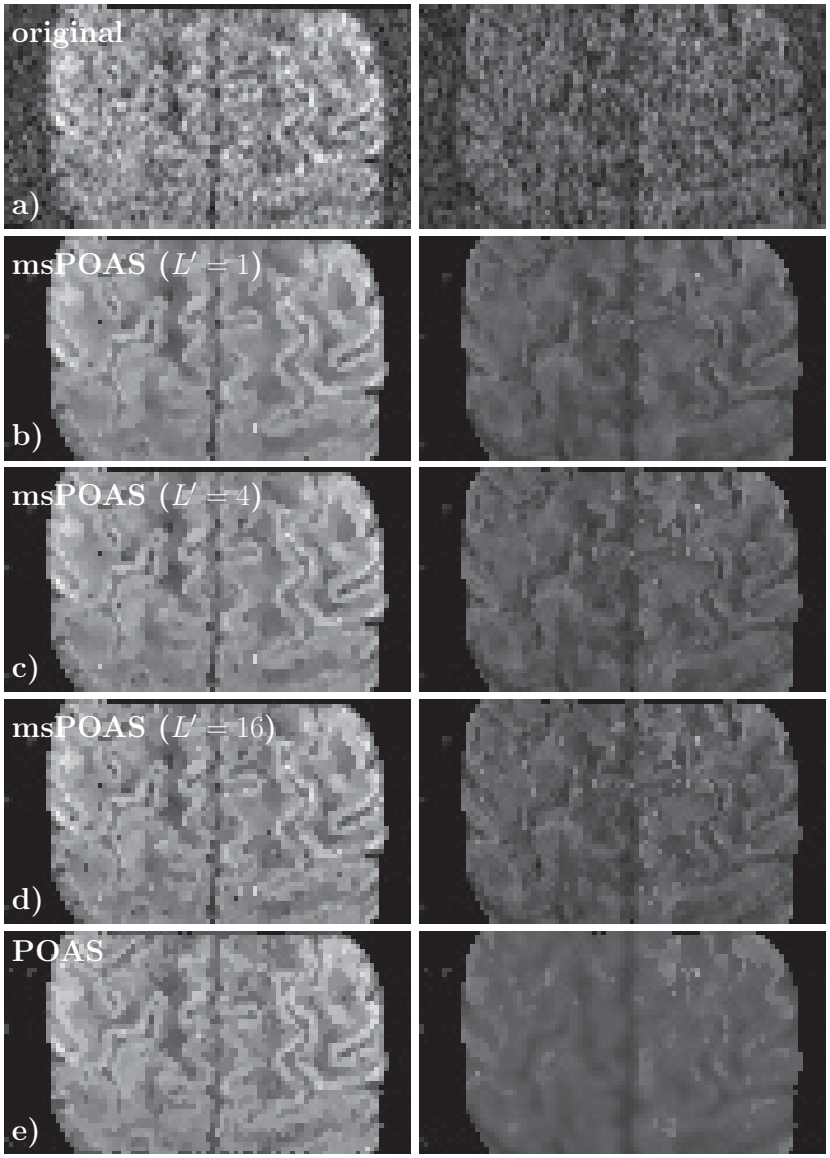

Figure 6.20: Diffusion-weighted data for some arbitrarily selected slice of the multi-shell experiment with $b = 800\,\mathrm{s/mm^2}$ (left) and $b = 2000\,\mathrm{s/mm^2}$ (right).

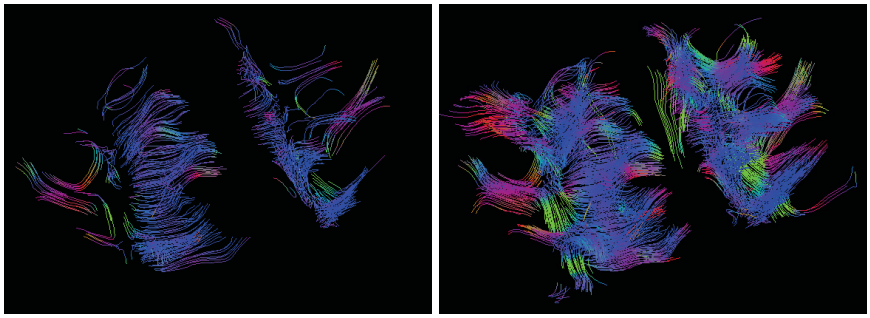

Figure 6.21: Fiber tracks of the multi-shell data after POAS (left) and msPOAS (right), using a FACT algorithm. For better visibility, we only show fibers with a minimal length of 25 segments.

the size and shape of the spatial structure at small scales and use this information to improve the pointwise estimates at coarser scales. As a consequence, the noise reduction with (ms)POAS avoids blurring by naturally adapting to the anisotropic structures observed in dMRI. This is essential as dMRI aims to examine fine fibrous structures in the human body.

Another important strength of (ms)POAS is its independence of any diffusion model, see § 6.1.3 for an overview. It is applied to the diffusion-weighted data. Consequently, the procedure does not introduce a bias towards any model, and it can be flexibly combined with the various diffusion models for subsequent data analysis. Additionally, modeling benefits from smoothed data since smoothing stabilizes modeling and avoids artifacts due to noise, see for example Parker et al. [2000] and Tabelow et al. [2008]. Finally, modeling usually results in a loss of information, which may complicate the adaptation.

(Ms)POAS forms a left-invariant operation since it is Euclidean invariant in the voxel space and rotation invariant in the gradient space. In other words, standardization of the smoothed MR image by Euclidean operations, that is rotation and translation, yields the same result as smoothing of the standardized image. This is a common requirement for image processing operations.

The conspicuous efficiency of (ms)POAS results from using the whole information of the measurement space $\mathbb{R}^3 \times \mathbb{S}^2$. This enables a simultaneous smoothing in the voxel and the gradient space. Especially in the first iteration steps, smoothing on the sphere considerably stabilizes the procedure without blurring at structural borders. This makes the method feasible for experimental data with a very low signal-to-noise ratio. In the case of high SNR, (ms)POAS leaves the data almost unchanged as we demonstrated in Figure 6.18 (b). In Section 6.5, a series of examples with simulated and experimental data illustrated the effectiveness of (ms)POAS. As it turned out, the procedure yields similar results as the averaged image of repeated scans. Therefore, the method (ms)POAS allows a reduction of acquisition time, while still providing a similar data quality.

However, smoothing on the sphere possibly introduces a bias since extrema may be weakened. Therefore, (ms)POAS benefits from a high number of applied diffusion-weighting gradient directions, which limits the spherical bias and stabilizes adaptation in the first iteration steps. For a low number of applied gradients, say less than 20, the parameter κ_0 should be chosen such that (ms)POAS restricts smoothing to the voxel space. This requires a sufficient image contrast to compensate for the missing stabilization of adaptation by spherical smoothing. The multi-shell procedure msPOAS additionally benefits from sampling different q-shells. Particularly for large b-values, the additional information from the inner q-shells improves adaptation. On the one hand, application of identical gradient schemes for all b-values renders spherical interpolation unnecessary, which allows to accelerate the method. On the other hand, varying gradient schemes provide a higher spatial resolution, although the interpolation of missing values may slightly bias the statistical penalty.

We should critically discuss the assumption of statistically independent data. Many methods for pre-processing use registration to compensate for artifacts which are caused by motion and eddy currents, for instance. The registration matches via affine transformations and interpolation different slices, scans, or brains in order to provide comparability [Mohammadi et al., 2010]. Generally, this introduces spatial correlation, but it avoids spurious discontinuities, which (ms)POAS could identify in the case of unregistered data. In our experience, (ms)POAS benefits from registered data, while the effects of the spatial correlation seem to be negligible. A combination of registration methods with (ms)POAS could further improve the results since registration and adaptive smoothing may benefit from each other.

Another questionable assumption is the structural assumption of the underlying Propagation-Separation Approach, which supposes a locally constant model. This is certainly not satisfied, but it can be considered as approximation of a piecewise polynomial model, which is more realistic. The univariate example in Figure 5.12 (page 74) of a piecewise polynomial parameter function with non-central chi-distributed observations shows that the Propagation-Separation Approach forces the estimation function into a step function if the maximal location bandwidth $h^{(k^*)}$ is sufficiently large. A similar behavior can be observed for (ms)POAS in Figure 6.19 (page 132), where we consider the behavior of the estimator during iteration in a region of interest of the first experimental data set. The same figures also show that intermediate steps of the iteration yield good results in accordance with the locally polynomial model. As in Section 5.4, we conclude that an appropriate choice of k^* might improve the results. Usually, it is distinguishable by visual inspection of the diffusion-weighted images whether the algorithm reduces noise, or whether it started to create artificial steps in order to force the smoothed estimation function into the locally constant model. The numerical results suggest that, even without stopping, the estimation bias is bounded by the resulting step function for large values of k^*.

As mentioned in § 6.1.3, there are many voxels which contain crossings of several fiber bundles with distinct directions. In this case, the measured signal is determined by all fibers present in the respective voxel. This leads to different signal intensities in voxels with varying numbers of fiber directions. Hence, at best, a homogeneous region as detected by (ms)POAS includes all voxels which contain a certain combination of fiber directions, which is a subset of all voxels which contain one of the involved fiber bundles completely. This reduces the size of the homogeneous regions and consequently the amount of smoothing in comparison with the unrealistic situation where we had complete knowledge of the partial volume effects. Nevertheless, (ms)POAS avoids blurring at these structural borders with the only effect of a non-optimal amount of smoothing.

6.6.2 Impact of the parameters

In § 6.2.2, we specified the choices of the various method parameters. Most of them can be chosen in reasonable ranges without causing trouble if varied slightly. In the **R**-package **dti** by Tabelow and Polzehl [2013], the kernel functions K_{loc} and K_{ad}, the sequence of location bandwidths $\{h^{(k)}\}_{k=1}^{k^*}$, and the balancing parameter κ are fixed in agreement with § 6.2.2. Then, we recommend to choose $h^{(0)} = 1$, $k^* = 12$, and $\kappa_0 \in (0,1)$ such that $N_g(1-\cos(\kappa_0)) \in [5,10]$, where $N_g = \sum_{b \in B} |G_b|/\mathfrak{B}$ denotes the mean number of applied gradients per shell.

In the following, we discuss the impact of the adaptation bandwidth λ and the data dependent input parameters σ and L', denoting the noise standard deviation and the effective number of MR receiver coils, respectively. Recall that σ and L' are used in the adaptive weights of the estimator to determine the statistical penalty, which depends via the Kullback-Leibler divergence on the probability distribution of the standardized observations S/σ. If σ and L' are unknown, they should be estimated separately by an appropriate method.

In this thesis, we assumed σ and L' to be homogeneous over the voxel space. This assumption may be violated, for instance, in the case of parallel imaging, where $L' > 1$, see § 6.1.4. Basically, (ms)POAS could be extended to heteroscedasticity as we will discuss in § 6.6.6. As we observe in Figure 6.20 (page 133), (ms)POAS is relatively robust against misspecification of L'. This is a very helpful property since L' is usually unknown and even more difficult to estimate from the data than the standard deviation.

The adaptation bandwidth λ is the crucial parameter of the procedure. In § 6.2.2, we proposed a choice given by the propagation condition. This choice is determined by the probability distribution of the observations as demonstrated in Chapter 3. In Table 6.1 (page 128), we observed that the propagation condition is (almost) invariant with respect to the fixed parameter θ_0. Therefore, we can choose λ independent of the noise standard deviation σ, which only influences the size of the true parameter function $\theta(.)$. Additionally, we observe in the same table that the propagation condition is almost robust against the realization of the sample, the number of applied diffusion-weighting gradients per q-shell, and the degrees of freedom $2L'$. The latter follows from the small impact of L' in the Gaussian approximation of the Kullback-Leibler divergence, see Equation (6.19) (page 98). Only the number of applied b-values noticeable influenced the propagation level.

In the (ms)POAS algorithm, the adaptation bandwidth λ interacts with the standard deviation σ since both parameters only appear in the adaptive weights, see Equations (6.15) (page 97) and (6.16). More precisely, they influence the procedure in a similar manner. Therefore, to some extent, a misspecification of σ can be balanced by a subsequent adjustment of λ. Similarly, a data-dependent adjustment of λ can help to compensate possibly violated assumptions or other uncertainties which may influence the procedure. Then, the choice of λ given by the propagation condition serves as a starting point or, if the propagation level ϵ of λ is sufficiently small, as an upper bound which covers the worst case.

6.6.3 Comparison of POAS and msPOAS

The method POAS is a special case of msPOAS, where a single b-value $b \geq 0$ is fixed. In contrast, for smoothing of data with various b-values, we would apply POAS to each shell separately, while msPOAS allows simultaneous smoothing of all shells. To our knowledge, msPOAS is the first algorithm which combines information in position, orientation, and from all shells. This includes an explicit coupling of the S_0-image with the other shells. As a consequence, msPOAS on single-shell data does not equal POAS, where the S_0-image is smoothed separately.

The implementations of POAS and msPOAS in the **R**-package **dti** show additional differences due to some changes which we introduced to improve and accelerate msPOAS compared to the previously implemented POAS algorithm. Nevertheless, we emphasize that all resulting differences between POAS and the restriction of msPOAS to a single-shell can be considered to be exchangeable modules or building blocks of the procedure. This is an important feature of the algorithm, which enables to meet specific properties of data. First, in the implementation of msPOAS, we use another distance function, see Notation 6.64. This is a left-invariant pseudometric, which ensures the Euclidean invariance in the voxel space and the rotation invariance in the gradient space. Additionally, it is much simpler than the distance function in the implementation of POAS, see Notation 6.63, which was based on the left-invariant vector fields of the special Euclidean motion group $\mathrm{SE}(3)$. Second, we introduced a simplified approximation of the Kullback-Leibler divergence which appears in the statistical penalty of the adaptive weights. This can be considered as the symmetrized Kullback-Leibler divergence between the approximating Gaussian distributions with coinciding first and second moments. It considerably accelerates the computations, while also yielding better results than the numerical approximation used in POAS, see Becker et al. [2012, App. B]. Additionally, the Gaussian approximation motivated the usage of a weighted arithmetic mean in msPOAS instead of the weighted quadratic mean for the estimator in POAS.

The superiority of the multi-shell approach becomes obvious when comparing the results of POAS and msPOAS on experimental data. Particularly for the multi-shell data set, msPOAS outperforms POAS as shown in Figures 6.20 and 6.21 (page 133). Here, the large b-value in the outer shell yields a considerable signal attenuation. As a consequence, discontinuities are scaled down, while the noise levels are comparable. Obviously, the resulting very low SNR does not suffice for successful adaptation by POAS. In contrast, msPOAS benefits from the high SNR on the S_0-image and the additional information of the inner shell. In Figure 6.18 (page 131), we see, for a single-shell data set, that both procedures behave similar.

6.6.4 Alternative perspectives on diffusion-weighted data

The presented approach for adaptive smoothing of dMRI data is based on a specific perspective on the data. In this chapter, we described the data by a standard orientation score, that is by a real valued function on the measurement space $\mathbb{R}^3 \times \mathbb{S}^2$. To our knowledge, only McGraw et al. [2009] followed the same approach, but they applied spatial and spherical smoothing separately. Hagmann et al. [2006] and later on Duits and Franken [2011], Duits et al. [2011], Franken [2008] and Jonasson et al. [2007] considered real valued functions on $\mathbb{R}^3 \times \mathbb{S}^2$ in order to derive fiber tracks. However, these functions were derived from diffusion models which yield, for each voxel, a real valued function on the sphere, such as the diffusion tensor or QBI, see § 6.1.3. In contrast, we consider the diffusion-weighted images prior to any modeling. Of course, this is not the only possible perspective on dMRI data.

For instance, we could restrict (ms)POAS to spatial smoothing by choosing the balancing parameter κ_0 sufficiently small. This coincides with a description of the data by the function $S_{(b,\vec{g})} : \mathbb{R}^3 \to \mathbb{R}$, where $b \geq 0$ and $\vec{g} \in G_b$ are fixed. It leads to a loss of efficiency due to the missing stabilization of the adaptation by spherical smoothing and a smaller maximal achievable variance reduction, as a consequence, of the reduced sample size. Hence, spatial smoothing with fixed gradients requires a sufficiently high SNR. There again, we could restrict (ms)POAS to spherical smoothing by choosing κ_0 sufficiently large and $h^{(0)}$ sufficiently small. However, spherical smoothing of dMRI data introduces a bias, which is compensated in (ms)POAS by the increasing amount of spatial smoothing.

Alternatively, the diffusion-weighted data can be described as a map $S_b : \mathbb{R}^3 \to (\mathbb{S}^2 \to \mathbb{R})$. This maps each voxel to a real valued function on the sphere, where the b-value $b > 0$ is fixed. Various diffusion models are based on this description of the signal function, which allows voxelwise estimation of quantities such as the diffusion tensor or the ODF in QBI. Tournier et al. [2004] used it for voxelwise estimation of a certain fiber orientation density function by spherical deconvolution via a spherical and rotational harmonic decomposition for regularization. These approaches only use the voxelwise information. Nevertheless, this setting again combines the whole information of the measurement space in position and orientation. Application of the Propagation-Separation Approach would require the replacement of the Kullback-Leibler divergence in the statistical penalty by a distance between functions on the sphere. Here, one could follow the approach by Rozenholc et al. [2010] which is based on adaptive testing [Spokoiny, 1996]. The approach can be extended to dMRI data of the above form, using spherical wavelets as introduced, for instance, by Lounsbery [1995], and Schröder and Sweldens [1995]. In comparison with (ms)POAS, the above description of the data is less efficient due to the missing stabilization by spherical smoothing and probably smaller homogeneous regions as the adaptation criterion is more restrictive. Therefore, we would not expect improved smoothing results by describing the dMRI data in the form $S_b : \mathbb{R}^3 \to (\mathbb{S}^2 \to \mathbb{R})$.

6.6.5 Other approaches for smoothing of diffusion-weighted MRI

The comparison of different smoothing methods for dMRI data is difficult. First, there have been many proposals from different research areas, making it very hard, or even impossible, to come to know of all of them. Second, many proposals are complex, and the associated software is not always available. Third, all procedures should be applied to the same data sets. Here, it is recommendable to consider a certain number of different settings with low and high SNR, small and large sample sizes, varying numbers of applied diffusion-weighting gradients, different numbers of MR receiver coils and so forth. However, experimental data and in particular data with a specific setting are not freely accessible. Additionally, many procedures are sensitive to the choice of parameters. Therefore, a fair comparison can often only be made in cooperation with the originator. Finally, there is no general quality criterion. Several smoothing methods induce artifacts, but their consequences for a subsequent modeling and analysis have not been studied yet. The efficiency of the method, its capability of edge preservation, the computation time, and the robustness of parameter choices are important aspects. Which one is more important and how should they be weighted for a final rating?

In principle, we can state the following. As mentioned above, it is advantageous to smooth the diffusion-weighted data before modeling. Among these smoothing methods, the Gaussian filter is the most widely-used procedure in the neuroimaging community [Westin et al., 1999]. It is easy to handle and has very short computation time. However, isotropic smoothing blurs the fine structure observed in dMRI. This requires very small bandwidths which reduces efficiency.

In order to overcome this problem, many more sophisticated methods have been proposed. There are several approaches using wavelets, where the choice of the applied wavelet basis is crucial for the efficiency of the procedure [Anand and Sahambi, 2010; Delakis et al., 2007; Nowak, 1999; Pizurica et al., 2006]. The corresponding algorithms are fast, but they introduce artifacts which are known as Gibbs phenomenon. Therefore, they are often combined with other methods to compensate for these artifacts [Lohmann et al., 2010; Manjón et al., 2008]. Similarly, the efficiency of Bayesian approaches mainly depends on the a priori assumption [Awate and Whitaker, 2007]. However, they are very time consuming and computationally demanding and therefore less widely-used. Other methods are based on the total variation norm or modifications of it, see, for instance, Blomgren and Chan [1998], Guo and Huang [2009], and Rudin et al. [1992]. McGraw et al. [2009] demonstrated that a total variation based scheme for spatial smoothing of dMRI data benefits from previous spherical smoothing. Probably, a coupling of spatial and spherical smoothing could further improve the results.

Anisotropic diffusion is often and successfully used for noise reduction of dMRI data, see Aja-Fernández et al. [2009], Ding et al. [2005], Lysaker et al. [2003], Parker et al. [2000], and Perona and Malik [1990]. This is also time consuming, although it is usually faster than the Bayesian procedures. The results strongly depend on the stopping parameter since infinite diffusion results in a homogeneous setting. Blurring of structural boarders is reduced, but cannot be fully avoided. Admittedly, the anisotropy requires the calculation of derivative gradients, which may fail in the case of a very low SNR. Heuristically, the Propagation-Separation Approach and, as a consequence, (ms)POAS yield similar results as methods based on anisotropic diffusion without the need for gradient values. Another possibility to overcome this problem was proposed by Lohmann et al. [2010], where wavelet filtering and anisotropic diffusion are combined. Then, the wavelet induced artifacts are compensated by subsequent anisotropic diffusion, which is in turn stabilized by the previous wavelet smoothing. On the single-shell data set, the method by Lohmann et al. [2010] yield similar results as (ms)POAS, see Figure 6.20 (page 133). On an unpublished data set with even lower SNR, the (ms)POAS algorithm was the superior method.

Finally, there are several procedures using non-local means [Aja-Fernandez et al., 2008; Buades et al., 2005a; Coupé et al., 2008, 2013; Wiest-Daesslé et al., 2008] or non-local maximum likelihood estimates [He and Greenshields, 2009; Rajan et al., 2011]. In Section 2.3, we compared the Propagation-Separation Approach with the non-local means method. For smoothing of diffusion-weighted MRI data, the Propagation-Separation Approach seems to be more intuitive. It is constructed to detect discontinuities in a locally polynomial setting, providing propagation within homogeneous regions and separation at structural borders. The adaptation criterion of the non-local means method requires similar blockwise neighborhoods. For a three-dimensional image with fine structure, this is a restrictive criterion, which possibly reduces the maximal achievable variance reduction considerably without obvious benefit. In comparison with weighted means, maximum likelihood estimates have the advantage of providing the non-centrality parameter instead of the expected value. However, through an (estimated) noise standard deviation σ and a known (or approximated) effective number of MR receiver coils L', the non-centrality parameters can be calculated by Equation (6.7) (page 90), using the estimated expected value. Probably, this is more advantageous than direct estimation by the computationally demanding maximum likelihood approach, which as well requires knowledge of σ and L'.

The structural adaptive smoothing algorithm by Tabelow et al. [2008] is also based on the Propagation-Separation Approach. Nevertheless, there are two important differences to the (ms)POAS procedure. First, the method by Tabelow et al. [2008] is restricted to spatial smoothing in the voxel space. Second, it uses the diffusion tensor model for adaptation. This leads to a loss of information with the result that separation fails in certain situations as illustrated in Figure 6.17 (page 130). Additionally, for a very low SNR, the estimation of the diffusion tensor model becomes considerably unstable, which negatively affects smoothing as shown in Figure 6.18 (c) (page 131).

In a nutshell, (ms)POAS can be considered as a very promising algorithm for the denoising of diffusion-weighted MRI data. Its main drawbacks are its computation time, and that the various mutual reactions in the underlying Propagation-Separation Approach can be confusing initially. Nevertheless, its advantages are apparent. It is very efficient due to the simultaneous smoothing in position and orientation and, for msPOAS, the coupling of all measured q-shells for adaptation. The algorithm provides satisfactory results for experimental data with a very low SNR, where other smoothing procedures fail. Its capability of edge preservation was confirmed on simulated and on experimental data, where possible artifacts can be controlled by an appropriate stopping. Finally, its implementation in the **R**-package **dti** by Tabelow and Polzehl [2013] is freely available, and its parameter choices are robust.

6.6.6 Future research

The presented study raises several questions for future research.

As discussed in the last subsection, there is a lack of an appropriate quality criterion which allows a meaningful and detailed comparison of various smoothing methods. Therefore, it would be very interesting to review and analyze existing quality measures and to develop and justify new ones. This would allow a detailed comparison and evaluation of the existing approaches for the denoising of dMRI data. Additionally, one could study the impact of a coupling between spherical and spatial smoothing or between different shells for other denoising methods than the Propagation-Separation Approach.

In order to further improve the smoothing results of (ms)POAS, the method could be extended to a heteroscedastic setting with a locally varying noise standard deviation σ and varying effective numbers of receiver coils L'. Samsonov and Johnson [2004] proposed a denoising method for spatially varying noise levels, using non-linear diffusion. Guo and Huang [2009] and Manjón et al. [2010] generalized total variation based denoising and non-local means methods to the case of a heteroscedastic variance. Of course, this requires local estimates of both values, which is still a challenging problem. Here, the Propagation-Separation Approach could be used for simultaneous variance estimation and position-orientation adaptive smoothing.

Additionally, the impact of the Kullback-Leibler divergence could be studied. In § 6.2.1, we gave several reasons for its application. However, a simpler distance could provide similar or even better results, especially as (ms)POAS only uses a rough approximation of the Kullback-Leibler divergence between two non-central chi-distributions.

Alternatively, the weighted mean in (ms)POAS could be replaced by a locally weighted maximum likelihood estimator. This is computationally demanding, but it provides the non-centrality parameter instead of the expected value. We recall that the measured signals S correspond to the expected values of the data, while the feature of interest is the non-centrality parameter θ of the associated non-central chi-distribution. Basically, this bias can be corrected, using Equation (6.7) (page 90). However, for satisfactory results, this requires local estimates of σ and L'. As a first step, the advantages and drawbacks of both approaches should be evaluated by a detailed comparison of methods which estimate the non-centrality parameter and methods that estimate the expected value from which the non-centrality parameter is calculated.

Another avenue for future research would be the combination of data registration and noise reduction by (ms)POAS. As mentioned above, (ms)POAS benefits from registered data, while registration can be improved, using smoothed data. Hence, registration could be updated in each iteration step of (ms)POAS, using the smoothing results of the last iteration step. Then, registration could benefit from smoothing and vice versa. Possible challenges of this approach might be the introduced statistical dependence of the observations and the computation time as such a combination would result in an elaborate procedure.

Finally, we mention some implications of the presented results for the regularization of inverse problems on geometrical structures. Particularly in the context of image processing, there appear several inverse problems such as deconvolution or the estimation of derivatives. These are often ill-posed, and hence require regularization, as many imaging modalities suffer from significant random noise. For instance, in the case of a linear ill-posed problem with additive noise, one aims to reconstruct the component f from the noisy observations $g = Af + \epsilon$, where the invertibility is impeded since $g \notin \mathrm{range}(A)$ is not in the non-closed range of the linear, compact, and bounded operator A. Then, smoothing of the observation g can provide a component $g_{\text{smooth}} \in \mathrm{range}(A)$ and hence a reconstruction of the unknown component f via $f_{\text{smooth}} := A^{-1} g_{\text{smooth}}$. The presented results suggest that smoothing of geometrical structures can benefit from an appropriate handling of the specific geometry of the observed objects. In the case of oriented objects, the applied operations should be left-invariant and at best structural adaptive. We apply our left-invariant smoothing method (ms)POAS as preparation for the solution of the respective inverse problem for the subsequent analysis or the modeling of the data. A different approach modifies the inverse operator A^{-1} instead of the observation g, where regularization methods as Tikhonov regularization or the spectral cut-off may be applied. This raises the question whether the usual regularization methods, see Engl et al. [1996] for an overview, allow an appropriate handling of a specific geometry, and which modifications are possibly required for the respective application.

6.7 Proofs

Proof of Proposition 6.18. The map

$$\mathrm{SO}(3)/\mathrm{stab}(\vec{e}_z) \ni \{\mathbf{R}_{(\alpha',\beta,\gamma)} : \alpha' \in (-\pi,\pi]\} \mapsto \vec{u}(\beta,\gamma) \in \mathbb{S}^2$$

defines an isomorphism since $\mathbf{R}_{(\alpha',\beta,\gamma)}\vec{e}_z = \vec{u}(\beta,\gamma)$ for all $\alpha' \in (-\pi,\pi]$, which ensures that this map and its inverse are bijective and continuously differentiable, see also Bröcker and tom Dieck [1995, (4.7)]. Additionally, for all $\vec{v} \in \mathbb{R}^3$, it holds $[\vec{v}]_{\vec{0}} = \{\vec{v} + \vec{0}\}$ and consequently $\mathbb{R}^3/\{\vec{0}\} \cong \mathbb{R}^3$. Hence, we get the isomorphy

$$\mathbb{R}^3 \times \mathbb{S}^2 \cong (\mathbb{R}^3/\{\vec{0}\}) \rtimes (\mathrm{SO}(3)/\mathrm{stab}(\vec{e}_z)) = \mathrm{SE}(3)/(\{\vec{0}\} \rtimes \mathrm{stab}(\vec{e}_z)).$$

□

Proof of Lemma 6.20. We show the existence of $g_1, g_2 \in \mathrm{SE}(3)$ with

$$[g_1]_{\mathrm{stab}(\vec{e}_z)} \cdot_{\mathrm{SE}(3)} [g_2]_{\mathrm{stab}(\vec{e}_z)} \neq [g_1 \cdot_{\mathrm{SE}(3)} g_2]_{\mathrm{stab}(\vec{e}_z)},$$

where $\cdot_{\mathrm{SE}(3)}$ denotes the group operation of $\mathrm{SE}(3)$. For this purpose, we consider the rotation matrices $\mathbf{R}_z^{(1)}, \mathbf{R}_z^{(2)} \in \mathrm{stab}(\vec{e}_z)$ and the group elements $g_1, g_2 \in \mathrm{SE}(3)$ with $g_i = (\vec{v}_i, \mathbf{R}_i)$, $i = 1, 2$. It holds

$$\begin{aligned}\left[g_1 \cdot_{\mathrm{SE}(3)} \left(\vec{0}, \mathbf{R}_z^{(1)}\right)\right] \cdot_{\mathrm{SE}(3)} \left[g_2 \cdot_{\mathrm{SE}(3)} \left(\vec{0}, \mathbf{R}_z^{(2)}\right)\right] &= \left(\vec{v}_1, \mathbf{R}_1 \cdot \mathbf{R}_z^{(1)}\right) \cdot_{\mathrm{SE}(3)} \left(\vec{v}_2, \mathbf{R}_2 \cdot \mathbf{R}_z^{(2)}\right) \\ &= \left(\vec{v}_1 + \mathbf{R}_1 \cdot \mathbf{R}_z^{(1)} \cdot \vec{v}_2, \mathbf{R}_1 \cdot \mathbf{R}_z^{(1)} \cdot \mathbf{R}_2 \cdot \mathbf{R}_z^{(2)}\right)\end{aligned}$$

and, for all $\mathbf{R}_z \in \mathrm{stab}(\vec{e}_z)$, we have

$$\left[g_1 \cdot_{\mathrm{SE}(3)} g_2\right] \cdot_{\mathrm{SE}(3)} \left(\vec{0}, \mathbf{R}_z\right) = (\vec{v}_1 + \mathbf{R}_1 \cdot \vec{v}_2, \mathbf{R}_1 \cdot \mathbf{R}_2 \cdot \mathbf{R}_z).$$

This yields the assertion with $\vec{v}_2 \neq \vec{e}_z$ and $\mathbf{R}_z^{(1)} \neq \mathbf{1}$. □

Proof of Proposition 6.33. First, we show that $\mathcal{W}^*[\mathbb{L}_2(\mathrm{SE}(d),\mathbb{R})] = \mathbb{L}_2(\mathbb{R}^d,\mathbb{R})$. Jensen's inequality yields, for every $U \in \mathbb{L}_2(\mathrm{SE}(d),\mathbb{R})$, that

$$\int_{\mathbb{R}^d} [\mathcal{W}^*[U](\vec{v})]^2 \, d\vec{v} = \int_{\mathbb{R}^d} \left[\int_{\mathrm{SO}(d)} U(\vec{v},\mathbf{R}) d\mu(\mathbf{R})\right]^2 d\vec{v} \leq \int_{\mathbb{R}^d} \int_{\mathrm{SO}(d)} [U(\vec{v},\mathbf{R})]^2 \, d\mu(\mathbf{R}) d\vec{v} < \infty$$

and consequently $\mathcal{W}^*[U] \in \mathbb{L}_2(\mathbb{R}^d,\mathbb{R})$. This verifies that $\mathcal{W}^*[\mathbb{L}_2(\mathrm{SE}(d),\mathbb{R})] \subseteq \mathbb{L}_2(\mathbb{R}^d,\mathbb{R})$. Vice versa, for every $f \in \mathbb{L}_2(\mathbb{R}^d,\mathbb{R})$, there is an orientation score $U_f \in \mathbb{L}_2(\mathrm{SE}(d),\mathbb{R})$ given by $U_f(\vec{v},\mathbf{R}) := f(\vec{v})$ for all $\mathbf{R} \in \mathrm{SO}(3)$, and we have $\mathcal{W}^*[\mathbb{L}_2(\mathrm{SE}(d),\mathbb{R})] \supseteq \mathbb{L}_2(\mathbb{R}^d,\mathbb{R})$. This leads to the above assertion. Additionally, for all $g \in \mathrm{SE}(d)$ and every $U \in \mathbb{L}_2(\mathrm{SE}(d),\mathbb{R})$, the left-invariance of Φ implies by definition of Γ and Equation (6.28) that

$$\begin{aligned}\mathcal{E}_g \circ \Gamma \circ \mathcal{W}^* \circ U &= \mathcal{E}_g \circ \mathcal{W}^* \circ \Phi \circ U &&= \mathcal{W}^* \circ \mathcal{L}_g \circ \Phi \circ U \\ &= \mathcal{W}^* \circ \Phi \circ \mathcal{L}_g \circ U &&= \Gamma \circ \mathcal{W}^* \circ \mathcal{L}_g \circ U = \Gamma \circ \mathcal{E}_g \circ \mathcal{W}^* \circ U,\end{aligned}$$

which terminates the proof. □

Proof of Proposition 6.34. The first assertion follows in the same manner as the special case $d = 2$ that was proven by Franken [2008, §2.7.3]. Hence, we concentrate on the second assertion. By definition of Φ_K, for every $U \in \mathbb{L}_2(\mathrm{SE}(d), \mathbb{R})$ and all $g, m \in \mathrm{SE}(d)$, it holds

$$\begin{aligned} \mathcal{L}_m \circ [\Phi_K \circ U](g) &= [\Phi_K U](m^{-1}g) \\ = \int_{\mathrm{SE}(3)} K(h^{-1}m^{-1}g)U(h)d\mu(h) &\overset{n:=mh}{=} \int_{\mathrm{SE}(3)} K(n^{-1}g)U(m^{-1}n)d\mu(n) \\ = \int_{\mathrm{SE}(3)} K(n^{-1}g)[\mathcal{L}_m \circ U](n)d\mu(n) &= \Phi_K \circ [\mathcal{L}_m \circ U](g), \end{aligned}$$

and we have $\mathcal{L}_m \circ \Phi_K = \Phi_K \circ \mathcal{L}_m$. □

Proof of Theorem 6.38. First, we show the left-invariance of the operator ϕ_k, $k \in \{0, ..., k^*\}$, in Notation 6.37, which relates to the POAS procedure. Let the applied distance δ_κ be left-invariant. For all $S_b \in \mathbb{L}_2(V \times G_b, \mathbb{R})$ with $b \in B_0$ and every $v \in \mathrm{SE}(3)$ and $m \in \mathbb{R}^3 \times \mathbb{S}^2$ with $L_v^{-1}[m] \in V \times G_b$, we observe that

$$(\mathcal{L}_v \circ \phi_k)[S_b](m) = \phi_k[S_b](L_v^{-1}[m]) = \Big(\sum_{n \in V \times G_b} \tilde{w}^{(k)}_{L_v^{-1}[m],n} S_b^2(n) / \tilde{N}^{(k)}_{L_v^{-1}[m]}\Big)^{1/2},$$

and

$$\begin{aligned} \phi_k[\mathcal{L}_v \circ S_b](m) &= \Big(\sum_{z : L_v^{-1}[z] \in V \times G_b} \tilde{w}^{(k)}_{mz} S_b^2(L_v^{-1}[z]) / \tilde{N}^{(k)}_m\Big)^{1/2} \\ &\overset{n:=L_v^{-1}[z]}{=} \Big(\sum_{n \in V \times G_b} \tilde{w}^{(k)}_{m,L_v[n]} S_b^2(n) / \tilde{N}^{(k)}_m\Big)^{1/2}. \end{aligned}$$

We show that the adaptive weights of POAS satisfy

$$\tilde{w}^{(k)}_{L_v^{-1}[m],n}[S_b] = \tilde{w}^{(k)}_{m,L_v[n]}[\mathcal{L}_v \circ S_b] \quad \text{for all } m, n \in \mathbb{R}^3 \times \mathbb{S}^2 \text{ and every } v \in \mathrm{SE}(3), \tag{6.33}$$

where we explicitly distinguish whether the weights are calculated with respect to the estimator of S_b or with respect to $\mathcal{L}_v \circ S_b$. Then, it follows $\tilde{N}^{(k)}_{L_v^{-1}[m]}[S_b] = \tilde{N}^{(k)}_m[\mathcal{L}_v \circ S_b]$ as

$$\sum_{n \in V \times G_b} \tilde{w}^{(k)}_{L_v^{-1}[m],n}[S_b] = \sum_{n \in V \times G_b} \tilde{w}^{(k)}_{m,L_v[n]}[\mathcal{L}_v \circ S_b] \overset{z:=L_v[n]}{=} \sum_{z : L_v^{-1}[z] \in V \times G_b} \tilde{w}^{(k)}_{mz}[\mathcal{L}_v \circ S_b],$$

which leads to the left-invariance of the operator ϕ_k via

$$\begin{aligned} (\mathcal{L}_v \circ \phi_k)[S_b](m) &= \Big(\sum_{n \in V \times G_b} \{\tilde{w}^{(k)}_{L_v^{-1}[m],n}[S_b]\} S_b^2(n) / \{\tilde{N}^{(k)}_{L_v^{-1}[m]}[S_b]\}\Big)^{1/2} \\ &\overset{\text{Eq. (6.33)}}{=} \Big(\sum_{n \in V \times G_b} \{\tilde{w}^{(k)}_{m,L_v[n]}[\mathcal{L}_v \circ S_b]\} S_b^2(n) / \{\tilde{N}^{(k)}_m[\mathcal{L}_v \circ S_b]\}\Big)^{1/2} \\ &= \phi_k[\mathcal{L}_v \circ S_b](m). \end{aligned}$$

Hence, it remains to show Equation (6.33), where we proceed by induction, starting at $k = 0$. It holds, for all $m, n \in \mathbb{R}^3 \times \mathbb{S}^2$, that $\tilde{w}^{(0)}_{mn} = \overline{w}^{(0)}_{mn}$. Additionally, the choices of $h^{(0)}$ and $\kappa^{(0)}$ satisfy, for every $m \in \mathbb{R}^3 \times \mathbb{S}^2$, that $h^{(0)}(b, \vec{g}_m) \equiv 1$ and $\kappa^{(0)}(b, \vec{g}_m) \equiv \kappa_0$ for some $\kappa_0 > 0$. Thus, it follows from the left-invariance of δ_κ that

$$\tilde{w}^{(0)}_{L_v^{-1}[m],n} = K_{\mathrm{loc}}\left(\delta_{\kappa_0}(L_v^{-1}[m], n)\right) = K_{\mathrm{loc}}\left(\delta_{\kappa_0}(m, L_v[n])\right) = \tilde{w}^{(0)}_{m,L_v[n]}.$$

Supposing that the assertion is valid at iteration step $k-1$, we may deduce the following. It holds $h^{(k)}(b,\vec{g}_{L_v^{-1}[m]}) = h^{(k)}(b,\vec{g}_m)$ since $h^{(k)}$ is determined by solving the equation

$$\frac{\sum\limits_{n\in V\times G_b}[K_{\mathrm{loc}}(\delta_{\kappa_0/h^{(k)}}(m,n)/h^{(k)})]^2}{[\sum\limits_{n\in V\times G_b}K_{\mathrm{loc}}(\delta_{\kappa_0/h^{(k)}}(m,n)/h^{(k)})]^2} \overset{!}{=} 1.25^{-k}\,\frac{\sum\limits_{n\in V\times G_b}[K_{\mathrm{loc}}(\delta_{\kappa_0}(m,n))]^2}{[\sum\limits_{n\in V\times G_b}K_{\mathrm{loc}}(\delta_{\kappa_0}(m,n))]^2},$$

which only depends on m via the left-invariant distance δ_κ. Then, for the non-adaptive weights, we get with $\kappa^{(k)}(b,\vec{g}_m) := \kappa_0/h^{(k)}(b,\vec{g}_m)$, that

$$\begin{aligned}
\overline{w}^{(k)}_{L_v^{-1}[m],n} &= K_{\mathrm{loc}}\left(\delta_{\kappa^{(k)}(b,\vec{g}_{L_v^{-1}[m]})}(L_v^{-1}[m],n)/h^{(k)}(b,\vec{g}_{L_v^{-1}[m]})\right)\\
&= K_{\mathrm{loc}}\left(\delta_{\kappa^{(k)}(b,\vec{g}_m)}(m,L_v[n])/h^{(k)}(b,\vec{g}_m)\right)\\
&= \overline{w}^{(k)}_{m,L_v[n]}.
\end{aligned}$$

The induction hypothesis yields

$$\begin{aligned}
&\tilde{w}^{(k)}_{L_v^{-1}[m],n}[S_b]\\
=\ &\overline{w}^{(k)}_{L_v^{-1}[m],n}\cdot K_{\mathrm{ad}}\left(\tilde{N}^{(k-1)}_{L_v^{-1}[m]}[S_b]\cdot\mathcal{KL}\left(\phi_{k-1}[S_b](L_v^{-1}[m]),\phi_{k-1}[S_b](n)\right)/\lambda\right)\\
=\ &\overline{w}^{(k)}_{m,L_v[n]}\cdot K_{\mathrm{ad}}\left(\tilde{N}^{(k-1)}_{m}[\mathcal{L}_v\circ S_b]\cdot\mathcal{KL}\left(\mathcal{L}_v\circ\phi_{k-1}[S_b](m),\mathcal{L}_v^{-1}\circ\phi_{k-1}[\mathcal{L}_v\circ S_b](n)\right)/\lambda\right)\\
=\ &\overline{w}^{(k)}_{m,L_v[n]}\cdot K_{\mathrm{ad}}\left(\tilde{N}^{(k-1)}_{m}[\mathcal{L}_v\circ S_b]\cdot\mathcal{KL}\left(\phi_{k-1}[\mathcal{L}_v\circ S_b](m),\phi_{k-1}[\mathcal{L}_v\circ S_b](L_v[n])\right)/\lambda\right)\\
=\ &\tilde{w}^{(k)}_{m,L_v[n]}[\mathcal{L}_v\circ S_b].
\end{aligned}$$

The left-invariance of the operator φ_k, which relates to msPOAS, follows in an analogous manner. In msPOAS, we use a weighted arithmetic mean instead of the weighted quadratic mean in POAS. Additionally, we use a simplified approximation of the Kullback-Leibler divergence and a modified statistical penalty, which sums the Kullback-Leibler divergences of all b-values. These modifications do not effect the proof of left-invariance. Recall that we assumed to have identical gradient schemes. Hence, the interpolation does not apply.

Conversely, the left-invariance of the operator ϕ_k implies Equation (6.33) and furthermore

$$\delta_\kappa(L_v^{-1}[m],n) = \delta_\kappa(m,L_v[n])$$

for every $\kappa>0$, all $m,n\in\mathbb{R}^3\times\mathbb{S}^2$, and each $v\in\mathrm{SE}(3)$. This is equivalent to the left-invariance of the applied distance. The same holds with φ_k instead of ϕ_k.

Finally, we emphasize that the bandwidths $\{h^{(k)}(b,\vec{g}_m)\}_{k=0}^{k^*}$ do not use the embedding of $\mathbb{R}^3\times\mathbb{S}^2$ into $\mathrm{SE}(3)$. Additionally, we recall that the balancing parameter κ only depends on the gradient $\vec{g}_m$ via the location bandwidths, and it follows that POAS and msPOAS are well-defined with respect to the embedding if and only if the distance δ_κ is well-defined. □

Proof of Lemma 6.57. The first assertion was proven by Varopoulos et al. [1992, §III.4]. The notation of Proposition 6.55 yields with $\{p_{ij}\}_{ij} := \mathrm{diag}\{1,1,1,1,1,0\}$ that

$$\sum_{i=1}^{5}|\phi_i(s)|^2 = \mathbf{T}_{\varphi(s)}(\dot{\phi}(s),\dot{\phi}(s)).$$

Then, it holds

$$\begin{aligned}\Delta(g_1,g_2) &= \inf\{\int_0^1 \left[\mathbf{T}_{\varphi(s)}(\dot\phi(s),\dot\phi(s))\right]^{1/2} ds : \phi:[0,1]\to \mathrm{SE}(3) \text{ absolutely continuous,}\\ &\qquad \phi(0)=g_1, \phi(1)=g_2, \text{ and } \dot\phi(s)=\sum_{i=1}^{5}\phi_i(s)A_i|_{\phi(s)} \text{ almost everywhere}\},\end{aligned}$$

and the assertion follows from Proposition 6.55, where the left-invariance yields

$$\Delta(g_1,g_2)=\Delta(g_2^{-1}\cdot_{\mathrm{SE}(3)} g_1, e)=|g_2^{-1}\cdot_{\mathrm{SE}(3)} g_1|$$

for all $g_1,g_2\in\mathrm{SE}(3)$. □

Proof of Proposition 6.65. The metric properties of $\|\vec v_1-\vec v_2\|+\kappa^{-1}\arccos\langle\vec u_1,\vec u_2\rangle$ were justified by Hagmann et al. [2006]. The distance δ_κ depends on the absolute value of $\langle\vec u_1,\vec u_2\rangle$, which reduces δ_κ to a pseudometric. Hence, it remains to justify the left-invariance of δ_κ. Let $m:=(\vec w,\mathbf{R})\in\mathrm{SE}(3)$ and $g_i:=(\vec v_i,\vec u_i)\in\mathbb{R}^3\times\mathbb{S}^2$, $i=1,2$. Then, we get by the rotation invariance of the Euclidean metric and of the scalar product that

$$\begin{aligned}\delta_\kappa\left(L_m^{-1}[g_1],L_m^{-1}[g_2]\right) &= \delta_\kappa\left[(\mathbf{R}^{-1}(\vec v_1-\vec w),\mathbf{R}^{-1}\vec u_1),(\mathbf{R}^{-1}(\vec v_2-\vec w),\mathbf{R}^{-1}\vec u_2)\right]\\ &= \|\mathbf{R}^{-1}(\vec v_1-\vec w-\vec v_2+\vec w)\|+\kappa^{-1}\arccos|\langle\mathbf{R}^{-1}\vec u_1,\mathbf{R}^{-1}\vec u_2\rangle|\\ &= \|\vec v_1-\vec v_2\|+\kappa^{-1}\arccos|\langle\vec u_1,\vec u_2\rangle|\\ &= \delta_\kappa(g_1,g_2),\end{aligned}$$

leading to the assertion. □

Bibliography

I. Aganj, C. Lenglet, G. Sapiro, E. Yacoub, K. Ugurbil, and N. Harel. Reconstruction of the orientation distribution function in single- and multiple-shell q-ball imaging within constant solid angle. *Magnetic Resonance in Medicine*, 64(2):554–566, 2010. ISSN 1522-2594. URL http://dx.doi.org/10.1002/mrm.22365.

S. Aja-Fernandez, C. Alberola-Lopez, and C. F Westin. Noise and signal estimation in magnitude MRI and Rician distributed images: A LMMSE approach. *IEEE Transactions on Image Processing*, 17(8): 1383–1398, 2008. ISSN 1057-7149. doi: 10.1109/TIP.2008.925382.

S. Aja-Fernández, A. Tristán-Vega, and C. Alberola-López. Noise estimation in single- and multiple-coil magnetic resonance data based on statistical models. *Magnetic Resonance Imaging*, 27(10):1397–1409, 2009. URL http://dx.doi.org/10.1016/j.mri.2009.05.025.

M. Akkouchi. On the convolution of exponential distributions. *Journal of the Chungcheong Mathematical Society*, 21(4):501–510, 2008.

A.L. Alexander, K.M. Hasan, M. Lazar, J.S. Tsuruda, and D.L. Parker. Analysis of partial volume effects in diffusion-tensor MRI. *Magnetic Resonance in Medicine*, 45(5):770–780, 2001. ISSN 1522-2594. URL http://dx.doi.org/10.1002/mrm.1105.

C.S. Anand and J.S. Sahambi. Wavelet domain non-linear filtering for MRI denoising. *Magnetic Resonance Imaging*, 28(6):842–861, 2010. doi: 10.1016/j.mri.2010.03.013.

Y. Assaf and P.J. Basser. Composite hindered and restricted model of diffusion (CHARMED) MR imaging of the human brain. *NeuroImage*, 27(1):48–58, 2005. pii: S1053-8119(05)00225-9.

H.-E. Assemlal, D. Tschumperlé, L. Brun, and K. Siddiqi. Recent advances in diffusion MRI modeling: Angular and radial reconstruction. *Medical Image Analysis*, 15(4):369–396, 2011. pii: S1361-8415(11)00024-7.

S.P. Awate and R.T. Whitaker. Feature-preserving MRI denoising: a nonparametric empirical Bayes approach. *IEEE Transactions on Medical Imaging*, 26(9):1242–1255, 2007. ISSN 0278-0062. doi: 10.1109/TMI.2007.900319.

P.J. Basser and S. Pajevic. Statistical artifacts in diffusion tensor MRI (DT-MRI) caused by background noise. *Magnetic Resonance in Medicine*, 44(1):41–50, 2000. ISSN 1522-2594. URL http://dx.doi.org/10.1002/1522-2594(200007)44:1<41::AID-MRM8>3.0.CO;2-O.

P.J. Basser, J. Mattiello, and D. Le Bihan. Estimation of the effective self-diffusion tensor from the NMR spin echo. *Journal of Magnetic Resonance, Series B*, 103(3):247–254, 1994a. ISSN 1064-1866. doi: http://dx.doi.org/10.1006/jmrb.1994.1037. URL http://www.sciencedirect.com/science/article/pii/S1064186684710375.

P.J. Basser, J. Mattiello, and D. LeBihan. MR diffusion tensor spectroscopy and imaging. *Biophysical Journal*, 66(1):259–267, 1994b. ISSN 0006-3495. doi: http://dx.doi.org/10.1016/S0006-3495(94)80775-1. URL http://www.sciencedirect.com/science/article/pii/S0006349594807751.

S.M.A. Becker. The Propagation-Separation approach: Consequences of model misspecification. *WIAS Preprint*, 1877, 2013. URL http://www.wias-berlin.de/publications/wias-publ/.

S.M.A. Becker and P. Mathé. A different perspective on the Propagation-Separation Approach. *Electronic Journal of Statistics*, 7:2702–2736, 2013. ISSN 1935-7524. doi: 10.1214/13-EJS860.

S.M.A. Becker, K. Tabelow, H.U. Voss, A. Anwander, R.M. Heidemann, and J. Polzehl. Position-orientation adaptive smoothing of diffusion weighted magnetic resonance data (POAS). *Medical Image Analysis*, 16(6):1142–1155, 2012. ISSN 1361-8415. doi: http://dx.doi.org/10.1016/j.media.2012.05.007. URL http://www.sciencedirect.com/science/article/pii/S1361841512000631.

S.M.A. Becker, K. Tabelow, S. Mohammadi, N. Weiskopf, and J. Polzehl. Adaptive smoothing of multi-shell diffusion-weighted magnetic resonance data by msPOAS. *WIAS Preprint*, 1809, 2013. URL http://www.wias-berlin.de/publications/wias-publ/.

T.E.J. Behrens, M.W. Woolrich, M. Jenkinson, H. Johansen-Berg, R.G. Nunes, S. Clare, P.M. Matthews, J.M. Brady, and S.M. Smith. Characterization and propagation of uncertainty in diffusion-weighted MR imaging. *Magnetic Resonance in Medicine*, 50(5):1077–1088, 2003. ISSN 1522-2594. URL http://dx.doi.org/10.1002/mrm.10609.

D. Belomestny and V. Spokoiny. Spatial aggregation of local likelihood estimates with applications to classification. *The Annals of Statistics*, 35(5):2287–2311, 2007. URL http://dx.doi.org/10.1214/009053607000000271.

P. Blomgren and T.F. Chan. Color TV: total variation methods for restoration of vector-valued images. *IEEE Transactions on Image Processing*, 7(3):304–309, 1998. ISSN 1057-7149. doi: 10.1109/83.661180.

P.A. Bottomley, T.H. Foster, R.E. Argersinger, and L.M. Pfeifer. A review of normal tissue hydrogen NMR relaxation times and relaxation mechanisms from 1–100 MHz: dependence on tissue type, NMR frequency, temperature, species, excision, and age. *Medical physics*, 11:425, 1984. URL http://dx.doi.org/10.1118/1.595535.

R.N. Bracewell. *The Fourier transform and its applications*. McGraw-Hill New York, second edition, 1978. ISBN 9780070070134.

T. Bröcker and T. tom Dieck. *Representations of compact Lie groups*, volume 98 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1995. ISBN 0-387-13678-9. Translated from the German manuscript, Corrected reprint of the 1985 translation.

A. Buades, B. Coll, and J.-M. Morel. A review of image denoising algorithms, with a new one. *Multiscale Modeling & Simulation*, 4(2):490–530, 2005a. URL http://epubs.siam.org/doi/abs/10.1137/040616024.

A. Buades, B. Coll, and J.-M. Morel. A non-local algorithm for image denoising. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 60–65, 2005b. ISBN 0-7695-2372-2. doi: 10.1109/CVPR.2005.38.

P.T. Callaghan. *Principles of Nuclear Magnetic Resonance Microscopy*. Oxford University Press, 2007. ISBN 978 0 19 853997.

M.F. Carfora. Interpolation on spherical geodesic grids: a comparative study. *Journal of Computational and Applied Mathematics*, 210:99–105, 2007. ISSN 0377-0427. doi: http://dx.doi.org/10.1016/j.cam.2006.10.068. URL http://www.sciencedirect.com/science/article/pii/S0377042706006522.

E. Caruyer, J. Cheng, C. Lenglet, G. Sapiro, T. Jiang, and R. Deriche. Optimal design of multiple q-shells experiments for diffusion MRI. In *MICCAI Workshop on Computational Diffusion MRI- CDMRI 2011*, Toronto, Canada, 2011. URL http://hal.inria.fr/inria-00617663.

J. Cheng, A. Ghosh, T. Jiang, and R. Deriche. Model-free and analytical EAP reconstruction via spherical polar fourier diffusion MRI. In T. Jiang, N. Navab, J.P.W. Pluim, and M.A. Viergever, editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2010*, volume 6361 of *Lecture Notes in Computer Science*, pages 590–597. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-15704-2. URL http://dx.doi.org/10.1007/978-3-642-15705-9_72.

C.A. Clark, G.J. Barker, and P.S. Tofts. Magnetic resonance diffusion imaging of the human cervical spinal cord in vivo. *Magnetic Resonance in Medicine*, 41(6):1269–1273, 1999. ISSN 1522-2594. URL http://dx.doi.org/10.1002/(SICI)1522-2594(199906)41:6<1269::AID-MRM26>3.0.CO;2-2.

P. Coupé, P. Yger, S. Prima, P. Hellier, C. Kervrann, and C. Barillot. An optimized blockwise nonlocal means denoising filter for 3-D magnetic resonance images. *IEEE Transactions on Medical Imaging*, 27(4):425–441, 2008. ISSN 0278-0062. doi: 10.1109/TMI.2007.906087.

P. Coupé, J.V. Manjón, M. Chamberland, M. Descoteaux, and B. Hiba. Collaborative patch-based super-resolution for diffusion-weighted images. *NeuroImage*, 83(0):245 – 261, 2013. ISSN 1053-8119. doi: http://dx.doi.org/10.1016/j.neuroimage.2013.06.030. URL http://www.sciencedirect.com/science/article/pii/S1053811913006678.

I. Delakis, O. Hammad, and R.I. Kitney. Wavelet-based de-noising algorithm for images acquired with parallel magnetic resonance imaging (MRI). *Physics in Medicine and Biology*, 52(13):3741, 2007. URL http://dx.doi.org/10.1088/0031-9155/52/13/006.

C.-A. Deledalle, L. Denis, and F. Tupin. Iterative weighted maximum likelihood denoising with probabilistic patch-based weights. *IEEE Transactions on Image Processing*, 18(12):2661–2672, 2009. ISSN 1057-7149. doi: 10.1109/TIP.2009.2029593.

M. Descoteaux, R. Deriche, D. Le Bihan, J.-F. Mangin, and C. Poupon. Multiple q-shell diffusion propagator imaging. *Medical Image Analysis*, 15(4):603–621, 2011. ISSN 1361-8415. doi: http://dx.doi.org/10.1016/j.media.2010.07.001. URL http://www.sciencedirect.com/science/article/pii/S1361841510000939.

O. Dietrich, J.G. Raya, S.B. Reeder, M. Ingrisch, M.F. Reiser, and S.O. Schoenberg. Influence of multichannel combination, parallel imaging and other reconstruction techniques on MRI noise characteristics. *Magnetic Resonance Imaging*, 26(6):754–762, 2008. ISSN 0730-725X. doi: http://dx.doi.org/10.1016/j.mri.2008.02.001. URL http://www.sciencedirect.com/science/article/pii/S0730725X08000337.

Z. Ding, J.C. Gore, and A.W. Anderson. Reduction of noise in diffusion tensor images using anisotropic smoothing. *Magnetic Resonance in Medicine*, 53(2):485–490, 2005. ISSN 1522-2594. URL http://dx.doi.org/10.1002/mrm.20339.

D.V. Divine, J. Polzehl, and F. Godtliebsen. A propagation-separation approach to estimate the autocorrelation in a time-series. *Nonlinear processes in geophysics*, 15(4):591–599, 2008. doi: 10.5194/npg-15-591-2008. URL http://www.nonlin-processes-geophys.net/15/591/2008/.

R. Duits. *Perceptual Organization in Image Analysis*. PhD thesis, Eindhoven University of Technology, Eindhoven, The Netherlands, 2005.

R. Duits and E.M. Franken. Left-invariant diffusions on the space of positions and orientations and their application to crossing-preserving smoothing of HARDI images. *International Journal of Computer Vision*, 92(3):231–264, 2011. ISSN 0920-5691. URL http://dx.doi.org/10.1007/s11263-010-0332-z.

R. Duits and M. van Almsick. The explicit solutions of linear left-invariant second order stochastic evolution equations on the 2D Euclidean motion group. *Quarterly of Applied Mathematics*, 66(1):27–67, 2008. ISSN 0033-569X. doi: http://dx.doi.org/10.1090/S0033-569X-07-01066-0.

R. Duits, E. Creusen, A. Ghosh, and T. Dela Haije. Diffusion, convection and erosion on $\mathrm{SE}(3)/(0 \times \mathrm{SO}(2))$ and their applications to the enhancement of crossing fibers. *ArXiv, http://arxiv.org/pdf/1103.0656v5*, 2011.

N. Dungey, A.F.M. ter Elst, and D.W. Robinson. *Analysis on Lie groups with polynomial growth*, volume 214 of *Progress in Mathematics*. Birkhäuser Boston Inc., Boston, MA, 2003. ISBN 0-8176-3225-5.

R. Durrett. *Probability: theory and examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, fourth edition, 2010. ISBN 978-0-521-76539-8.

A.M.A. El-Sayed. Laguerre polynomials of arbitrary (fractional) orders. *Applied Mathematics and Computation*, 109(1):1–9, 2000. ISSN 0096-3003. doi: http://dx.doi.org/10.1016/S0096-3003(98)10112-1. URL http://www.sciencedirect.com/science/article/pii/S0096300398101121.

H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*. Mathematics and Its Applications. Kluwers Academic Publishers, Dordrecht, 1996. ISBN 9780792341574.

J. Fan and I. Gijbels. *Local polynomial modelling and its applications*, volume 66 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1996. ISBN 0-412-98321-4.

J. Fan, M. Farmen, and I. Gijbels. Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 60(3):591–608, 1998. ISSN 1369-7412. URL http://dx.doi.org/10.1111/1467-9868.00142.

E.M. Franken. *Enhancement of crossing elongated structures in images*. PhD thesis, Eindhoven University of Technology, Eindhoven, The Netherlands, 2008.

K.J. Friston, J.T. Ashburner, S.J. Kiebel, T.E. Nichols, and W.D. Penny. *Statistical parametric mapping: the analysis of functional brain images*. Academic Press, London, 2006. ISBN 9780123725608.

M.A. Griswold, P.M. Jakob, R.M. Heidemann, M. Nittka, V. Jellus, J. Wang, B. Kiefer, and A. Haase. Generalized autocalibrating partially parallel acquisitions (GRAPPA). *Magnetic Resonance in Medicine*, 47(6):1202–1210, 2002. ISSN 1522-2594. URL http://dx.doi.org/10.1002/mrm.10171.

W. Guo and F. Huang. Adaptive total variation based filtering for MRI images with spatially inhomogeneous noise and artifacts. In *IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2009*, pages 101–104, 2009. doi: 10.1109/ISBI.2009.5192993.

P. Hagmann. *From diffusion MRI to brain connectomics*. PhD thesis, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, France, 2005.

P. Hagmann, L. Jonasson, T. Deffieux, R. Meuli, J.-P. Thiran, and V.J. Wedeen. Fibertract segmentation in position orientation space from high angular resolution diffusion MRI. *NeuroImage*, 32(2):665–675, 2006. ISSN 1053-8119. doi: http://dx.doi.org/10.1016/j.neuroimage.2006.02.043. URL http://www.sciencedirect.com/science/article/pii/S1053811906001613.

L. He and I.R. Greenshields. A nonlocal maximum likelihood estimation method for Rician noise reduction in MR images. *IEEE Transactions on Medical Imaging*, 28(2):165–172, 2009. ISSN 0278-0062. doi: 10.1109/TMI.2008.927338.

R.M. Heidemann, D.A. Porter, A. Anwander, T. Feiweier, K. Heberlein, T.R. Knösche, and R. Turner. Diffusion imaging in humans at 7T using readout-segmented EPI and GRAPPA. *Magnetic Resonance in Medicine*, 64(1):9–14, 2010. ISSN 1522-2594. URL http://dx.doi.org/10.1002/mrm.22480.

R.M. Heidemann, A. Anwander, T. Feiweier, T.R. Knösche, and R. Turner. k-space and q-space: Combining ultra-high spatial and angular resolution in diffusion imaging using ZOOPPA at 7T. *NeuroImage*, 60(2):967–978, 2012. ISSN 1053-8119. doi: http://dx.doi.org/10.1016/j.neuroimage.2011.12.081. URL http://www.sciencedirect.com/science/article/pii/S1053811912000225.

E. Hewitt and K.A. Ross. *Abstract harmonic analysis. Vol. I*, volume 115 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, second edition, 1979. ISBN 3-540-09434-2.

K.M. Jansons and D.C. Alexander. Persistent angular structure: new insights from diffusion magnetic resonance imaging data. *Inverse Problems*, 19(5):1031, 2003. URL http://dx.doi.org/10.1088/0266-5611/19/5/303.

J.H. Jensen, J.A. Helpern, A. Ramani, H. Lu, and K. Kaczynski. Diffusional kurtosis imaging: the quantification of non-gaussian water diffusion by means of magnetic resonance imaging. *Magnetic Resonance in Medicine*, 53(6):1432–1440, 2005. ISSN 1522-2594. URL http://dx.doi.org/10.1002/mrm.20508.

J.X. Ji, J.B. Son, and S.D. Rane. PULSAR: A Matlab toolbox for parallel magnetic resonance imaging using array coils and multiple channel receivers. *Concepts in Magnetic Resonance Part B: Magnetic Resonance Engineering*, 31B(1):24–36, 2007. ISSN 1552-504X. URL http://dx.doi.org/10.1002/cmr.b.20081.

H. Johansen-Berg and T.E.J. Behrens, editors. *Diffusion MRI: From Quantitative Measurement to In-Vivo Neuroanatomy*. Academic Press, 2009. ISBN 978-0123747099.

L. Jonasson, X. Bresson, J.-P. Thiran, V.J. Wedeen, and P. Hagmann. Representing diffusion MRI in 5-D simplifies regularization and segmentation of white matter tracts. *IEEE Transactions on Medical Imaging*, 26(11):1547–1554, 2007. ISSN 0278-0062. doi: 10.1109/TMI.2007.899168.

D.K. Jones, editor. *Diffusion MRI: Theory, Methods, and Applications*. Oxford University Press, 2010. ISBN 978-0195369779.

D.K. Jones and P.J. Basser. "Squashing peanuts and smashing pumpkins": How noise distorts diffusion-weighted MR data. *Magnetic Resonance in Medicine*, 52(5):979–993, 2004. ISSN 1522-2594. URL http://dx.doi.org/10.1002/mrm.20283.

D.K. Jones, M.A. Horsfield, and A. Simmons. Optimal strategies for measuring diffusion in anisotropic systems by magnetic resonance imaging. *Magnetic Resonance in Medicine*, 42(3):515–525, 1999. doi:

10.1002/(SICI)1522-2594(199909)42:3<515::AID-MRM14>3.0.CO;2-Q.

D.K. Jones, T.R. Knösche, and R. Turner. White matter integrity, fiber count, and other fallacies: The do's and don'ts of diffusion {MRI}. *NeuroImage*, 73(0):239 – 254, 2013. ISSN 1053-8119. doi: http://dx.doi.org/10.1016/j.neuroimage.2012.06.081. URL http://www.sciencedirect.com/science/article/pii/S1053811912007306.

J. Jost. *Riemannian geometry and geometric analysis*. Universitext. Springer-Verlag, Heidelberg, sixth edition, 2011. ISBN 978-3-642-21297-0. URL http://dx.doi.org/10.1007/978-3-642-21298-7.

A. Kamali, A.E. Flanders, J. Brody, J.V. Hunter, and K.M. Hasan. Tracing superior longitudinal fasciculus connectivity in the human brain using high resolution diffusion tensor tractography. *Brain Structure and Function*, pages 1–13, 2013. ISSN 1863-2653. URL http://dx.doi.org/10.1007/s00429-012-0498-y.

V. Katkovnik, A. Foi, K. Egiazarian, and J. Astola. From local kernel to nonlocal multiple-model image denoising. *International Journal of Computer Vision*, 86(1):1–32, 2010. ISSN 0920-5691. URL http://dx.doi.org/10.1007/s11263-009-0272-7.

M. Kleinnijenhuis, M. Barth, D.C. Alexander, A.-M. van Cappellen van Walsum, and D.G. Norris. Structure tensor informed fiber tractography (STIFT) by combining gradient echo MRI and diffusion weighted imaging. *NeuroImage*, 59(4):3941–3954, 2012. ISSN 1053-8119. doi: http://dx.doi.org/10.1016/j.neuroimage.2011.10.078. URL http://www.sciencedirect.com/science/article/pii/S1053811911012420.

D. Le Bihan. Looking into the functional architecture of the brain with diffusion MRI. *Nature Reviews Neuroscience*, 4(6):469–480, 2003. doi: http://dx.doi.org/10.1038/nrn1119.

O. V. Lepski and V. G. Spokoiny. Optimal pointwise adaptive methods in nonparametric estimation. *The Annals of Statistics*, 25(6):2512–2546, 1997. ISSN 0090-5364. URL http://dx.doi.org/10.1214/aos/1030741083.

O. V. Lepski, E. Mammen, and V. G. Spokoiny. Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *The Annals of Statistics*, 25 (3):929–947, 1997. ISSN 0090-5364. URL http://dx.doi.org/10.1214/aos/1069362731.

O.V. Lepskiĭ. A problem of adaptive estimation in Gaussian white noise. *Akademiya Nauk SSSR. Teoriya Veroyatnosteĭ i ee Primeneniya*, 35(3):459–470, 1990. URL http://dx.doi.org/10.1137/1135065.

Y. Li, H. Zhu, D. Shen, W. Lin, J.H. Gilmore, and J.G. Ibrahim. Multiscale adaptive regression models for neuroimaging data. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 73(4):559–578, 2011. ISSN 1369-7412. URL http://dx.doi.org/10.1111/j.1467-9868.2010.00767.x.

Y. Li, J.H. Gilmore, J. Wang, M. Styner, W. Lin, and H. Zhu. TwinMARM: two-stage multiscale adaptive regression methods for twin neuroimaging data. *IEEE Transaction on Medical Imaging*, 31 (5):1100–1112, 2012. doi: 10.1109/TMI.2012.2185830. URL http://www.ncbi.nlm.nih.gov/pubmed/22287236.

C. Liu, R. Bammer, and M.E. Moseley. Generalized diffusion tensor imaging (GDTI): a method for characterizing and imaging diffusion anisotropy caused by non-Gaussian diffusion. *Israel Journal of Chemistry*, 43(1-2):145–154, 2003. ISSN 1869-5868. URL http://dx.doi.org/10.1560/HB5H-6XBR-1AW1-LNX9.

C. Loader. *Local regression and likelihood*. Statistics and Computing. Springer-Verlag, New York, 1999. ISBN 0-387-98775-4.

G. Lohmann, S. Bohn, K. Müller, R. Trampel, and R. Turner. Image restoration and spatial resolution in 7-tesla magnetic resonance imaging. *Magnetic Resonance in Medicine*, 64(1):15–22, 2010. ISSN 1522-2594. URL http://dx.doi.org/10.1002/mrm.22488.

J.M. Lounsbery. *Multiresolution analysis for surfaces of arbitrary topological type*. PhD thesis, University of Washington, Seattle, WA, USA, 1995. UMI Order No. GAX95-23718.

M. Lysaker, A. Lundervold, and Xue-Cheng Tai. Noise removal using fourth-order partial differential equation with applications to medical magnetic resonance images in space and time. *IEEE Transactions on Image Processing*, 12(12):1579–1590, 2003. ISSN 1057-7149. doi: 10.1109/TIP.2003.819229.

J.V. Manjón, J. Carbonell-Caballero, J.J. Lull, G. García-Martí, L. Martí-Bonmatí, and M. Robles. MRI denoising using non-local means. *Medical Image Analysis*, 12(4):514–523, 2008. ISSN 1361-8415. doi: http://dx.doi.org/10.1016/j.media.2008.02.004. URL http://www.sciencedirect.com/science/article/pii/S1361841508000248.

J.V. Manjón, P. Coupé, L. Martí-Bonmatí, D.L. Collins, and M. Robles. Adaptive non-local means denoising of MR images with spatially varying noise levels. *Journal of Magnetic Resonance Imaging*, 31(1):192–203, 2010. ISSN 1522-2586. URL http://dx.doi.org/10.1002/jmri.22003.

A.M. Mathai. Storage capacity of a dam with gamma type inputs. *Annals of the Institute of Statistical Mathematics*, 34(3):591–597, 1982. ISSN 0020-3157. URL http://dx.doi.org/10.1007/BF02481056.

P. Mathé and S.V. Pereverzev. Regularization of some linear ill-posed problems with discretized random noisy data. *Mathematics of Computation*, 75(256):1913–1929 (electronic), 2006. URL http://dx.doi.org/10.1090/S0025-5718-06-01873-4.

T. McGraw, B. Vemuri, E. Özarslan, Y. Chen, and T. Mareci. Variational denoising of diffusion weighted MRI. *Inverse Problems and Imaging*, 3(4):625–648, 2009. doi: 10.3934/ipi.2009.3.625.

L. Minati and W.P. Weglarz. Physical foundations, models, and methods of diffusion magnetic resonance imaging of the brain: A review. *Concepts in Magnetic Resonance Part A*, 30A(5):278–307, 2007. ISSN 1552-5023. URL http://dx.doi.org/10.1002/cmr.a.20094.

P.P. Mitra and P.N. Sen. Effects of microgeometry and surface relaxation on NMR pulsed-field-gradient experiments: Simple pore geometries. *Physical Review B*, 45:143–156, 1992. URL http://link.aps.org/doi/10.1103/PhysRevB.45.143.

S. Mohammadi, H.E. Möller, H. Kugel, D.K. Müller, and M. Deppe. Correcting eddy current and motion effects by affine whole-brain registrations: evaluation of three-dimensional distortions and comparison with slicewise correction. *Magnetic Resonance in Medicine*, 64(4):1047–1056, 2010. ISSN 1522-2594. URL http://dx.doi.org/10.1002/mrm.22501.

J.N. Morelli, V.M. Runge, T. Feiweier, J.E. Kirsch, K.W. Williams, and U.I. Attenberger. Evaluation of a modified Stejskal-Tanner diffusion encoding scheme, permitting a marked reduction in TE, in diffusion-weighted imaging of stroke patients at 3T. *Investigative Radiology*, 45(1):29–35, 2010. doi: 10.1097/RLI.0b013e3181c65c11.

S. Mori. *Introduction to Diffusion Tensor Imaging*. Elsevier Science, 2007. ISBN 978-0444528285.

S. Mori, B.J. Crain, V.P. Chacko, and P.C.M. van Zijl. Three dimensional tracking of axonal projections in the brain by magnetic resonance imaging. *Annals of Neurology*, 45(2):265–269, 1999. ISSN 1531-8249. URL http://dx.doi.org/10.1002/1531-8249(199902)45:2<265::AID-ANA21>3.0.CO;2-3.

P.G. Moschopoulos. The distribution of the sum of independent gamma random variables. *Annals of the Institute of Statistical Mathematics*, 37(3):541–544, 1985. ISSN 0020-3157. URL http://dx.doi.org/10.1007/BF02481123.

M.E. Moseley, Y. Cohen, J. Kucharczyk, J. Mintorovitch, H.S. Asgari, M.F. Wendland, J. Tsuruda, and D. Norman. Diffusion-weighted MR imaging of anisotropic water diffusion in cat central nervous system. *Radiology*, 176(2):439–445, 1990. doi: 10.1148/radiology.176.2.2367658. URL http://radiology.rsna.org/content/176/2/439.abstract.

H.-G. Müller. Change-points in nonparametric regression analysis. *The Annals of Statistics*, 20(2):737–761, 1992. ISSN 0090-5364. URL http://dx.doi.org/10.1214/aos/1176348654.

R.D. Nowak. Wavelet-based Rician noise removal for magnetic resonance imaging. *IEEE Transactions on Image Processing*, 8(10):1408–1419, 1999. ISSN 1057-7149. doi: 10.1109/83.791966.

F.W.J. Olver, D.W. Lozier, R.F. Boisvert, and C.W. Clark, editors. *NIST handbook of mathematical functions*. Cambridge University Press, Cambridge, 2010. ISBN 9780521192255.

A. Oppelt. *Imaging Systems for Medical Diagnostics: Fundamentals, Technical Solutions and Applications for Systems Applying Ionizing Radiation, Nuclear Magnetic Resonance and Ultrasound*. Publicis Publishing, second edition, 2005. ISBN 978-3895782268.

E. Özarslan and T.H. Mareci. Generalized diffusion tensor imaging and analytical relationships between diffusion tensor imaging and high angular resolution diffusion imaging. *Magnetic Resonance in Medicine*, 50(5):955–965, 2003. ISSN 1522-2594. URL http://dx.doi.org/10.1002/mrm.10596.

G.J. Parker, J.A. Schnabel, M.R. Symms, D.J. Werring, and G.J. Barker. Nonlinear smoothing for reduction of systematic and random errors in diffusion tensor imaging. *Journal of Magnetic Resonance Imaging*, 11(6):702–710, 2000. ISSN 1522-2586. URL http://dx.doi.org/10.1002/1522-2586(200006)11:6<702::AID-JMRI18>3.0.CO;2-A.

P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):629–639, 1990. ISSN 0162-8828. doi: 10.1109/34.56205.

A. Pizurica, A.M. Wink, E. Vansteenkiste, W. Philips, and B.J. Roerdink. A review of wavelet denoising in MRI and ultrasound brain imaging. *Current Medical Imaging Reviews*, 2(2):247–260, 2006. doi: 10.2174/157340506776930665.

J. Polzehl. *aws: Adaptive Weights Smoothing*, 2012. URL http://cran.r-project.org/package=aws. **R**-package version 1.9-3.

J. Polzehl and V. Spokoiny. Spatially adaptive regression estimation: Propagation-separation approach. *WIAS Preprint*, 998, 2004. URL http://www.wias-berlin.de/publications/wias-publ/.

J. Polzehl and V. Spokoiny. Propagation-separation approach for local likelihood estimation. *Probability Theory and Related Fields*, 135(3):335–362, 2006. ISSN 0178-8051. URL http://dx.doi.org/10.1007/s00440-005-0464-1.

J. Polzehl and V. Spokoiny. Structural adaptive smoothing by propagation-separation methods. In *Handbook of Data Visualization*, Springer Handbooks of Computational Statistics, pages 471–492. Springer-Verlag, Berlin Heidelberg, 2008. ISBN 978-3-540-33036-3. URL http://dx.doi.org/10.1007/978-3-540-33037-0_19.

J. Polzehl and V.G. Spokoiny. Adaptive weights smoothing with applications to image restoration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(2):335–354, 2000. ISSN 1467-9868. URL http://dx.doi.org/10.1111/1467-9868.00235.

J. Polzehl and K. Tabelow. Structural adaptive smoothing in diffusion tensor imaging: The **R** package *dti*. *Journal of Statistical Software*, 31(9):1–23, 2009. ISSN 1548-7660. URL http://www.jstatsoft.org/v31/i09.

J. Polzehl and K. Tabelow. Beyond the Gaussian model of diffusion-weighted imaging: The package dti. *Journal of Statistical Software*, 44(12):1–26, 2011. ISSN 1548-7660. URL http://www.jstatsoft.org/v44/i12.

J. Polzehl, H.U. Voss, and K. Tabelow. Structural adaptive segmentation for statistical parametric mapping. *NeuroImage*, 52(2):515–523, 2010. ISSN 1053-8119. doi: http://dx.doi.org/10.1016/j.neuroimage.2010.04.241. URL http://www.sciencedirect.com/science/article/pii/S1053811910006531.

K.P. Pruessmann, M. Weiger, M.B. Scheidegger, and P. Boesiger. SENSE: sensitivity encoding for fast MRI. *Magnetic Resonance in Medicine*, 42(5):952–962, 1999. ISSN 0740-3194.

P. Qiu. Discontinuous regression surfaces fitting. *The Annals of Statistics*, 26(6):2218–2245, 1998. ISSN 0090-5364. URL http://dx.doi.org/10.1214/aos/1024691468.

J. Rajan, B. Jeurissen, M. Verhoye, J. Van Audekerke, and J. Sijbers. Maximum likelihood estimation-based denoising of magnetic resonance images using restricted local neighborhoods. *Physics in Medicine and Biology*, 56(16):5221, 2011. doi: 10.1088/0031-9155/56/16/009. URL http://stacks.iop.org/0031-9155/56/i=16/a=009.

P.B. Roemer, W.A. Edelstein, C.E. Hayes, S.P. Souza, and O.M. Mueller. The NMR phased array. *Magnetic Resonance in Medicine*, 16(2):192–225, 1990. ISSN 1522-2594. URL http://dx.doi.org/10.1002/mrm.1910160203.

W. Rossmann. *Lie groups, An introduction through linear groups*, volume 5 of *Oxford Graduate Texts in Mathematics*. Oxford University Press, Oxford, 2002. ISBN 0-19-859683-9.

Y. Rozenholc, M. Reiss, D. Balvay, and C.-A. Cuénod. Growing time-homogeneous neighborhoods for denoising and clustering Dynamic Contrast Enhanced-CT sequences. *HAL-Preprint*, 00712348, 2010. URL http://hal.archives-ouvertes.fr/hal-00712348/.

L.I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992. ISSN 0167-2789. doi: http://dx.doi.org/10.

1016/0167-2789(92)90242-F. URL http://www.sciencedirect.com/science/article/pii/016727899290242F.

A.A. Sagle and R.E. Walde. *Introduction to Lie groups and Lie algebras*. Academic Press, New York, 1973. ISBN 978-0126145502.

A.A. Samsonov and C.R. Johnson. Noise-adaptive nonlinear diffusion filtering of MR images with spatially varying noise levels. *Magnetic Resonance in Medicine*, 52(4):798–806, 2004. ISSN 1522-2594. URL http://dx.doi.org/10.1002/mrm.20207.

P. Schröder and W. Sweldens. Spherical wavelets: Efficiently representing functions on the sphere. In *Proceedings of the 22Nd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '95, pages 161–172, New York, NY, USA, 1995. ACM. ISBN 0-89791-701-4. URL http://doi.acm.org/10.1145/218380.218439.

D.W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, 1992. ISBN 978-0471547709.

J. S. Simonoff. *Smoothing methods in statistics*. Springer Series in Statistics. Springer-Verlag, New York, 1996. ISBN 0-387-94716-7. URL http://dx.doi.org/10.1007/978-1-4612-4026-6.

S. Sinha, U. Sinha, and V.R. Edgerton. In vivo diffusion tensor imaging of the human calf muscle. *Journal of Magnetic Resonance Imaging*, 24(1):182–190, 2006. ISSN 1522-2586. URL http://dx.doi.org/10.1002/jmri.20593.

S.M. Smith and J.M. Brady. SUSAN–A new approach to low level image processing. *International Journal of Computer Vision*, 23(1):45–78, 1997. ISSN 0920-5691. URL http://dx.doi.org/10.1023/A%3A1007963824710.

V. Spokoiny and C. Vial. Parameter tuning in pointwise adaptation using a propagation approach. *The Annals of Statistics*, 37(5B):2783–2807, 2009. ISSN 0090-5364. URL http://dx.doi.org/10.1214/08-AOS607.

V. G. Spokoiny. Estimation of a function with discontinuities via local polynomial fit with an adaptive window choice. *The Annals of Statistics*, 26(4):1356–1378, 1998. ISSN 0090-5364. URL http://dx.doi.org/10.1214/aos/1024691246.

V.G. Spokoiny. Adaptive hypothesis testing using wavelets. *The Annals of Statistics*, 24(6):2477–2498, 1996. ISSN 0090-5364. URL http://dx.doi.org/10.1214/aos/1032181163.

E.O. Stejskal and J.E. Tanner. Spin diffusion measurements: spin echoes in the presence of a time-dependent field gradient. *The Journal of Chemical Physics*, 42(1):288–292, 1965. doi: http://dx.doi.org/10.1063/1.1695690.

K. Tabelow and J. Polzehl. *dti: DTI/DWI Analysis*, 2013. URL http://CRAN.R-project.org/package=dti. R package version 1.1-5.

K. Tabelow, J. Polzehl, V. Spokoiny, and H.U. Voss. Diffusion tensor imaging: Structural adaptive smoothing. *NeuroImage*, 39(4):1763 – 1773, 2008. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2007.10.024. URL http://www.sciencedirect.com/science/article/pii/S1053811907009470.

K. Tabelow, H.U. Voss, and J. Polzehl. Modeling the orientation distribution function by mixtures of angular central Gaussian distributions. *Journal of Neuroscience Methods*, 203(1):200–211, 2012. ISSN 0165-0270. doi: http://dx.doi.org/10.1016/j.jneumeth.2011.09.001. URL http://www.sciencedirect.com/science/article/pii/S0165027011005231.

A. Tabesh, J.H. Jensen, B.A. Ardekani, and J.A. Helpern. Estimation of tensors and tensor-derived measures in diffusional kurtosis imaging. *Magnetic Resonance in Medicine*, 65(3):823–836, 2011. URL http://dx.doi.org/10.1002/mrm.22655.

A.F.M. ter Elst and D.W. Robinson. Weighted subcoercive operators on Lie groups. *Journal of Functional Analysis*, 157(1):88 – 163, 1998. ISSN 0022-1236. doi: DOI:10.1006/jfan.1998.3259. URL http://www.sciencedirect.com/science/article/B6WJJ-45JCC1F-10/2/6db85c11ad35db31c013bd6f676129ec.

P. Thunberg and P. Zetterberg. Noise distribution in SENSE- and GRAPPA-reconstructed images: a computer simulation study. *Magnetic Resonance Imaging*, 25(7):1089–1094, 2007. ISSN 0730-725X. doi: http://dx.doi.org/10.1016/j.mri.2006.11.003. URL http://www.sciencedirect.com/science/article/pii/S0730725X06003973.

R. Tibshirani and T. Hastie. Local likelihood estimation. *Journal of the American Statistical Association*, 82(398):559–567, 1987. ISSN 0162-1459. doi: 10.1080/01621459.1987.10478466.

C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Sixth International Conference on Computer Vision*, pages 839–846. IEEE, 1998. ISBN 81-7319-221-9. doi: 10.1109/ICCV.1998.710815.

J.-D. Tournier, F. Calamante, D.G. Gadian, and A. Connelly. Direct estimation of the fiber orientation density function from diffusion-weighted MRI data using spherical deconvolution. *NeuroImage*, 23(3):1176 – 1185, 2004. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2004.07.037. URL http://www.sciencedirect.com/science/article/pii/S1053811904004100.

D.S. Tuch. *Diffusion MRI of Complex Tissue Structure*. PhD thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, 2002.

D.S. Tuch. Q-ball imaging. *Magnetic Resonance in Medicine*, 52(6):1358–1372, 2004. ISSN 1522-2594. URL http://dx.doi.org/10.1002/mrm.20279.

D.S. Tuch, T.G. Reese, M.R. Wiegell, N. Makris, J.W. Belliveau, and V.J. Wedeen. High angular resolution diffusion imaging reveals intravoxel white matter fiber heterogeneity. *Magnetic Resonance in Medicine*, 48(4):577–582, 2002. ISSN 1522-2594. URL http://dx.doi.org/10.1002/mrm.10268.

N.T. Varopoulos, L. Saloff-Coste, and T. Coulhon. *Analysis and geometry on groups*, volume 100 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1992. ISBN 0-521-35382-3.

M. P. Wand and M. C. Jones. *Kernel smoothing*, volume 60 of *Monographs on Statistics and Applied Probability*. Chapman and Hall Ltd., London, 1995. ISBN 0-412-55270-1.

C.-F. Westin, S.E. Maier, B. Khidhir, P. Everett, F.A. Jolesz, and R. Kikinis. Image processing for diffusion tensor magnetic resonance imaging. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 1999*, volume 1679 of *Lecture Notes in Computer Science*, pages 441–452. Springer Berlin Heidelberg, September 19–22 1999. ISBN 978-3-540-66503-8. URL http://dx.doi.org/10.1007/10704282_48.

N. Wiest-Daesslé, S. Prima, P. Coupé, S.P. Morrissey, and C. Barillot. Rician noise removal by non-local means filtering for low signal-to-noise ratio MRI: Applications to DT-MRI. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2008*, volume 5242 of *Lecture Notes in Computer Science*, pages 171–179. Springer Berlin Heidelberg, 2008. ISBN 978-3-540-85989-5. URL http://dx.doi.org/10.1007/978-3-540-85990-1_21.

L.P. Yaroslavsky. *Digital picture processing. An introduction*, volume 9 of *Springer Series in Information Sciences*. Springer-Verlag, Berlin, 1985. ISBN 978-3642819315.

L. Zhan, D. Franc, V. Patel, N. Jahanshad, Y. Jin, B.A. Mueller, M.A. Bernstein, B.J. Borowski, C.R. Jack, A.W. Toga, and Thompson P.M. Lim, K.O. How do spatial and angular resolution affect brain connectivity maps from diffusion MRI? In *9th IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–4, 2012. ISBN 978-1-4577-1857-1. doi: 10.1109/ISBI.2012.6235469.

List of Figures

Erklärung

Ich erkläre, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Berlin, den 05. Dezember 2013 Saskia Becker