Electronic Journal of Statistics
Vol. 7 (2013) 2702–2736
ISSN: 1935-7524
DOI: 10.1214/13-EJS860

A different perspective on the Propagation-Separation Approach

Saskia M. A. Becker* **and Peter Mathé**

Weierstrass Institute for Applied Analysis and Stochastics
Mohrenstrasse 39, 10117 Berlin, Germany
e-mail: saskia.becker@wias-berlin.de; peter.mathe@wias-berlin.de

Abstract: The Propagation-Separation Approach is an iterative procedure for pointwise estimation of local constant and local polynomial functions. The estimator is defined as a weighted mean of the observations with data-driven, iteratively updated weights. Within homogeneous regions it ensures a similar behavior as non-adaptive smoothing (propagation), while avoiding smoothing among distinct regions (separation). In order to enable a proof of stability of estimates, the authors of the original study introduced an additional memory step aggregating the estimators of the successive iteration steps. Here, we study theoretical properties of the simplified algorithm, where the memory step is omitted. In particular, we introduce a new strategy for the choice of the adaptation parameter yielding propagation and stability for local constant functions with sharp discontinuities.

AMS 2000 subject classifications: 62G05.
Keywords and phrases: Structural adaptive smoothing, propagation, separation, local likelihood, exponential families.

Received February 2013.

Contents

1. Introduction

The Propagation-Separation Approach [15] is an adaptive method for nonparametric estimation. This iterative procedure relates to Lepski's method [6, 11] and

*S. Becker was partially supported by the Stiftung der Deutschen Wirtschaft (SDW).

extends the Adaptive Weights Smoothing (AWS) procedure from Polzehl and Spokoiny [14]. The Propagation-Separation Approach supposes a local parametric model. It is especially powerful in case of large homogeneous regions and sharp discontinuities. However, it can be extended to local linear or local polynomial parameter functions, as well. Hence, the method 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 the local constant model is often satisfied.

In this study, we aim to provide a better understanding of the procedure and its properties. The crucial point of the algorithm is the choice of the adaptation bandwidth. We present a new formulation of what is known as propagation condition ensuring an appropriate choice. This formulation allows the verification of propagation and stability of estimates for local constant parameter functions with sharp discontinuities.

In comparison to the study of Polzehl and Spokoiny [15], there are two important differences which we want to emphasize. First, we avoid the problematic Assumption S0 on which the theoretical results in [15] were partially based on. This assumption requires the statistical independence of the adaptive weights from the observations. Theoretically, this can be ensured by means of the standard splitting technique. However, in practice, such a split is questionable due to the iterative approach of the algorithm. Second, we omit the memory step which was included into the algorithm to enable a theoretical study. In each iteration step, the new estimate is compared with the estimate from the previous iteration step. In case of a significant difference the new estimate is replaced by a value between the two estimates, providing a smooth transition, that is relaxation. This is related to the work of Belomestny and Spokoiny [4] about spatial aggregation of local likelihood estimates. The theoretical results in [15] are mainly based on the memory step. However, we show for piecewise constant functions that the adaptivity of the method yields similar results even if the memory step is removed from the algorithm. This gains importance as it turned out, that for practical use the memory step is questionable. Therefore, in later application of the algorithm, the memory step has been omitted, see e. g. Becker et al. [2], Li et al. [9, 8], Tabelow et al. [19], Divine et al. [5] still yielding the desired behavior in practice.

We will show that, for a local constant model, the simplified algorithm behaves very similar as before. Here, we aim to justify the omittance of the memory step, but we do not compare the results with other estimation methods or evaluate the estimation error since this has been done in previous works by Polzehl and Spokoiny, see [14, 15, 16]. Instead, we deduce similar properties for the simplified algorithm as they have been shown for the original procedure in [15]. We compare the theoretical results and discuss the impact of the memory step for the case that the unknown parameter function complies with the local constant model. Consequences of model misspecification will be analyzed in a separate study.

The outline is as follows. After a short introduction of the model and the estimation procedure we introduce a new parameter choice strategy for the adaptation bandwidth. Then, we consider some numerical examples that illustrate the general behavior of the algorithm. The main properties, as these are propagation, separation and stability of estimates, will be verified in Section 3 for piecewise constant parameter functions with sharp discontinuities. Here, we take advantage of our new choice of the adaptation bandwidth, which provides the desired properties under homogeneity. For piecewise constant parameter functions, the algorithm separates the homogeneity regions and treats each of them as under homogeneity, provided that the discontinuities are sufficiently large to be detected. In Section 4, we justify our new choice of the adaptation bandwidth by analyzing its invariance w.r.t. the unknown parameter function and by discussing some further questions concerning its application in practice. We finish with a discussion of the question whether the memory step is needed and if, where.

We use two results from Polzehl and Spokoiny [15] which do not base on Assumption S0. These are given in Appendix A. In order to improve readability we give longer proofs in Appendix B.

2. Model and methodology

In this section we briefly introduce the setting of our study and the estimation procedure resulting from the Propagation-Separation Approach. The behavior of the algorithm depends on the adaptation bandwidth, and we introduce a new strategy for its choice.

2.1. Model

We consider a local parametric model.

Notation 2.1 (Setting)**.** Let $Z_1, \dots, Z_n$ be independent random variables with $Z_i = (X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$. Here, the metric space $\mathcal{X}$ denotes the design space and $\mathcal{Y} \subseteq \mathbb{R}$ the observation space. The observations Y_i are assumed to follow the distribution $\mathbb{P}_{\theta(X_i)} \in \mathcal{P}$, where $\mathcal{P}$ denotes some parametric family of probability distributions and $\theta : \mathcal{X} \to \Theta \subseteq \mathbb{R}$ is the parameter function that we aim to estimate. We suppose the deterministic design $\{X_i\}_{i=1}^n$ to be known.

Typical examples of this general setting are Gaussian regression or the inhomogeneous Bernoulli, Exponential, and Poisson models, see [15, Section 2] for a detailed description. In general, the procedure may work for any vector space $\mathcal{Y} \subseteq M$ with $Y_i \sim \mathbb{P}_{\theta(X_i)}$, $\theta : \mathcal{X} \to \Theta \subseteq M$, where M is a metric space. Following Polzehl and Spokoiny [15] we suppose the parametric family to be an exponential family with standard regularity conditions. This allows an explicit expression of the Kullback-Leibler divergence simplifying our following analysis.

Assumption 1 (Local exponential family model)**.** $\mathcal{P} = (\mathbb{P}_\theta, \theta \in \Theta)$ is an exponential family with a convex parameter set $\Theta \subseteq \mathbb{R}$ and non-decreasing functions $C, B \in C^2(\Theta, \mathbb{R})$ 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 $\mathbb{P}$ denotes a dominating measure, $p(y)$ is some non-negative function on $\mathcal{Y}$, $T : \mathcal{Y} \to \mathbb{R}$, and $B'(\theta) = \theta\, C'(\theta)$. For the parameter θ it holds

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

Remark 2.2.

- In [15, Assumption (A1)], the authors assumed $T(y) \equiv y$, i.e. the identity map. Any invertible transformation T leaves the Kullback-Leibler divergence unchanged. Since the results in Equations (A.2) and (A.1), see Appendix A, depend on the Kullback-Leibler divergence only, they remain valid for invertible maps T. In this study, we consider the general case explicitly in order to clarify, where this transformation T comes into play.
- Equation (2.1), i.e. $\mathbb{E}_\theta[T(Y)] = \theta$, can be achieved via reparametrization with $\theta := t(\vartheta)$, where $t(\vartheta) := \mathbb{E}_\vartheta[T(Y)]$. For invertible functions $t(.)$ this allows estimation of ϑ by the adaptive estimator $\tilde{\vartheta} := t^{-1}(\tilde{\theta})$. Additionally, it follows for all $\vartheta_1, \vartheta_2 \in \Theta$ that $\mathcal{KL}(\vartheta_1, \vartheta_2) = \mathcal{KL}(\theta_1, \theta_2)$, where $\theta_i = t(\vartheta_i)$, $i = 1, 2$. If $t(\vartheta)$ is invertible and linear in ϑ, then we get $\mathcal{KL}(\tilde{\vartheta}, \mathbb{E}\tilde{\vartheta}) = \mathcal{KL}(\tilde{\theta}, \mathbb{E}\tilde{\theta})$. Hence the algorithm remains unmodified and the results in Sections 3 and 4 below remain valid if $t(\vartheta)$ is invertible and linear in ϑ.
- A list of parametric families satisfying Assumption (1), probably after reparametrization, is given in Table 1.
- We suppose Assumption (1) throughout this article while all later Assumptions will be required for specific results only.

In our subsequent analysis the notions of the Kullback–Leibler divergence, given here as

$$\mathcal{KL}\left(\mathbb{P}_\theta, \mathbb{P}_{\theta'}\right) := \int \ln\left(\frac{d(\mathbb{P}_\theta)}{d(\mathbb{P}_{\theta'})}\right) \mathbb{P}_\theta(dy), \quad \theta, \theta' \in \Theta,$$

and the Fisher information

$$I(\theta) := -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} \log p(y,\theta)\right], \quad \theta \in \Theta,$$

will be important.

Lemma 2.3 (Fisher information and Kullback-Leibler divergence)**.** *Under Assumption (1) we have that* $I(\theta) = C'(\theta)$, $\theta \in \Theta$. *Moreover, the following holds.*

- *For every compact and convex subset* $\Theta_\varkappa \subseteq \Theta$ *there is a constant* $\varkappa \geq 1$ *such that*

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

TABLE 1
One-parametric exponential families which satisfy Assumption (1), possibly after reparametrization

$\mathcal{P}$, support(f_ϑ)	Θ	$p(y)$	$T(y)$	$C(\vartheta)$	$B(\vartheta)$	$\mathbb{E}_\vartheta[T(Y)]$
$\mathcal{N}(\vartheta,\sigma^2)$ $y\in\mathbb{R}$	$\mathbb{R}$	$\frac{e^{-y^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}}$	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	$-\frac{1}{\vartheta}$	$n\ln\vartheta$	$n\,\vartheta$
Rayleigh(ϑ) $y\in[0,\infty)$	$(0,\infty)$	y	y^2	$-\frac{1}{2\vartheta^2}$	$2\ln\vartheta$	$2\vartheta^2$
Weibull(ϑ,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
Pareto(x_m,ϑ) $y\in[x_m,\infty)$	$(1,\infty)$	$\frac{1}{y}$	$\ln\left(\frac{y}{x_m}\right)$	$-\vartheta$	$-\ln(\vartheta)$	$\frac{1}{\vartheta}$
Poiss(ϑ) $y:=k\in\mathbb{N}$	$(0,\infty)$	$1/k!$	k	$\ln\vartheta$	ϑ	ϑ
Bin(n,ϑ) $y:=k\in 1:n$	$(0,1]$	$\binom{n}{k}$	k	$\ln\left(\frac{\vartheta}{1-\vartheta}\right)$	$-n\ln(1-\vartheta)$	$n\,\vartheta$
NegativeBin(r,ϑ) $y:=k\in\mathbb{N}$	$(0,1]$	$\binom{k+r-1}{k}$	k	$\ln\vartheta$	$-r\ln(1-\vartheta)$	$\frac{r\vartheta}{1-\vartheta}$
Bernoulli(ϑ) $y:=k\in\{0,1\}$	$(0,1]$	1	k	$\ln\left(\frac{\vartheta}{1-\vartheta}\right)$	$-\ln(1-\vartheta)$	ϑ

- *The Kullback-Leibler divergence is convex w.r.t. the first argument. It satisfies the following explicit formula and locally a quadratic approximation given as*

$$\mathcal{KL}(\mathbb{P}_\theta,\mathbb{P}_{\theta'}) = \theta\,[C(\theta)-C(\theta')]-[B(\theta)-B(\theta')] \tag{2.3}$$

$$= I(\theta)\,[\theta-\theta']^2/2 + r(\theta_*)(\theta-\theta')^3/6, \tag{2.4}$$

where $r(\theta_*) := -I'(\theta_*)/[I(\theta_*)]^3$ *with* θ_* *between* θ *and* θ'.

Proof sketch. The first assertion follows with $B'(\theta) = \theta C'(\theta)$. Then, Equation (2.2) holds due to the compactness of $\Theta_\varkappa$ and $C \in C^2(\Theta, \mathbb{R})$. The convexity is satisfied since the second derivative of the Kullback-Leibler divergence is non-negative

$$\tfrac{\partial^2}{\partial\theta^2} \mathcal{KL}\left(\mathbb{P}_\theta, \mathbb{P}_{\theta'}\right) = C'(\theta) > 0.$$

The local approximation follows from the reparametrization $v := C(\theta)$ and $D(v) := B(\theta)$ by Taylor's Theorem, where the remainder is given in Lagrange form. □

The set $\Theta_\varkappa$ should be sufficiently large such that $\theta(X_i) \in \Theta_\varkappa$ holds for all $i \in \{1, \dots, n\}$. Later on, we need that even the corresponding estimators are elements of $\Theta_\varkappa$, see Section 3.3. Explicit choices of $\varkappa$ and their consequences are discussed in Example A.2 for several probability distributions.

2.2. Methodology of the Propagation-Separation Approach

The Propagation-Separation Approach provides pointwise estimates of the unknown parameter function $\theta(.)$ introduced in Notation 2.1. In other words, for every design point X_i with $i \in \{1, \dots, n\}$ it yields a local estimator of the unknown parameter $\theta(X_i)$.

The algorithm is iterative, and in each iteration step the pointwise estimator of the parameter function is defined as a weighted mean of the observations. In each design point the weights are chosen adaptively as product of two kernel functions. The *location kernel* acts on the design space $\mathcal{X}$, and the *adaptation kernel* compares the pointwise parameter estimates of the previous iteration step in terms of the Kullback-Leibler divergence. For each of the two kernels, a bandwidth controls how much information is taken into account. The location bandwidth increases along the number of iterations. Starting at a small vicinity, in each iteration step the considered region is extended. The increasing number of included observations enables a monotone variance reduction during iteration, while the adaptation kernel leads to a decreasing or (in case of model misspecification) bounded estimation bias. It will be clear from the subsequent analysis that, by doing so, one obtains similar results as non-adaptive smoothing within homogeneity regions (propagation) and avoids smoothing across structural borders (separation). We turn to a formal description, and we start with introducing some notation.

Notation 2.4.

- $\theta_i := \theta(X_i)$;
- Δ denotes a metric on $\mathcal{X}$;
- $\mathcal{KL}(\theta, \theta') := \mathcal{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'})$ is the Kullback-Leibler divergence of the probability distributions $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$ with parameter $\theta, \theta' \in \Theta$;
- $K_{\text{loc}}, K_{\text{ad}} : [0, \infty) \to [0, 1]$ are non-increasing kernels with compact support $[0, 1]$ and $K_.(0) = 1$, where K_{loc} denotes the location and K_{ad} the adaptation kernel;

- $\{h^{(k)}\}_{k=0}^{k^*}$ is an increasing sequence of bandwidths for the location kernel with $h^{(0)} > 0$;
- $\lambda > 0$ is the bandwidth of the adaptation kernel;
- $U_i^{(k)} := \{X_j \in \mathcal{X} : \Delta(X_i, X_j) \leq h^{(k)}\}$.

For comparison and for the initialization of the algorithm we define the non-adaptive estimator $\overline{\theta}_i^{(k)}$.

Definition 2.5 (Non-adaptive estimator)**.** Let $i \in \{1, \dots, n\}$ and $k \in \{0, \dots, k^*\}$. The non-adaptive estimator $\overline{\theta}_i^{(k)}$ of θ_i is defined by

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

with weights $\overline{w}_{ij}^{(k)} := K_{\mathrm{loc}}(\Delta(X_i, X_j)/h^{(k)})$, and $\overline{N}_i^{(k)} := \sum_j \overline{w}_{ij}^{(k)}$.

Corollary 2.6 (Relation to maximum likelihood estimation)**.** *Assumption (1) implies that the standard local weighted maximum likelihood estimator*

$$\theta_i^{(\mathrm{MLE})} := \mathrm{argsup}_\theta L(\overline{W}_i^{(k)}, \theta) \quad \textit{with} \quad L(\overline{W}_i^{(k)}, \theta) := \sum_j \overline{w}_{ij}^{(k)} \log p(Y_j, \theta),$$

where $\overline{W}_i^{(k)} := \{\overline{w}_{ij}^{(k)}\}_j$*, equals the non-adaptive estimator* $\overline{\theta}_i^{(k)}$ *in Definition 2.5. Moreover, it follows for the "fitted log-likelihood" with* $\theta \in \Theta$ *that*

$$L(\overline{W}_i^{(k)}, \theta_i^{(\mathrm{MLE})}, \theta) := L(\overline{W}_i^{(k)}, \theta_i^{(\mathrm{MLE})}) - L(\overline{W}_i^{(k)}, \theta) = \overline{N}_i^{(k)} \mathcal{KL}\left(\overline{\theta}_i^{(k)}, \theta\right).$$

Now, we present the (slightly modified) algorithm of the Propagation-Separation Approach allowing $T(y) \neq y$ and omitting the memory step [15, Section 3.2] by setting $\eta_i \equiv 1$. Modifications to obtain the original algorithm as in [15, Section 3.3] are discussed in Remark 2.8, below. More details can be found in [15, Section 3].

Algorithm 1 (Propagation-Separation Approach)**.**

1. Input parameters: Sequence of location bandwidths $\{h^{(k)}\}_{k=0}^{k^*}$, adaptation bandwidth λ.
2. Initialization: $\tilde{\theta}_i^{(0)} := \overline{\theta}_i^{(0)}$ and $\tilde{N}_i^{(0)} := \overline{N}_i^{(0)}$ for all $i \in \{1, \dots, n\}$, $k := 1$.
3. Iteration: Do for every $i = 1, \dots, n$

$$\tilde{\theta}_i^{(k)} := \sum_{j=1}^{n} \tilde{w}_{ij}^{(k)} T(Y_j) / \tilde{N}_i^{(k)} \tag{2.5}$$

   with weights $\tilde{w}_{ij}^{(k)} := K_{\mathrm{loc}}(\Delta(X_i, X_j)/h^{(k)}) \cdot K_{\mathrm{ad}}(s_{ij}^{(k)}/\lambda)$, where $s_{ij}^{(k)} := \tilde{N}_i^{(k-1)} \mathcal{KL}(\tilde{\theta}_i^{(k-1)}, \tilde{\theta}_j^{(k-1)})$ and $\tilde{N}_i^{(k)} := \sum_j \tilde{w}_{ij}^{(k)}$.
4. Stopping: Stop if $k = k^*$ and return $\tilde{\theta}_i^{(k^*)}$ for all $i \in \{1, \dots, n\}$, otherwise increase k by 1.

Remark 2.7 (Choice of the input parameters)**.**

- 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 Sections 2.3, 4 and [15, Sections 3.4 and 3.5]. The choice $\lambda = \infty$ (formally) yields non-adaptive smoothing, while a small adaptation bandwidth λ leads to adaptation to noise such that the adaptive estimator equals the observation, i.e. $\tilde{\theta}_i^{(k)} = Y_i$.
- The initial location bandwidth $h^{(0)}$ should be sufficiently small in order to avoid smoothing among distinct homogeneous regions, before adaptation starts. In practice, any choice of $h^{(0)}$ such that $U_i^{(0)} = \{X_i\}$ for every $i \in \{1, \dots, n\}$ seems to be recommendable. Its drawback is discussed in Remark A.3.
- The sequence of bandwidth $\{h^{(k)}\}_{k=0}^{k^*}$ can be chosen such that $h^{(k)} := a^k h^{(0)}$ with $a \approx 1.25^{1/d}$, where d denotes the dimension of the design space $\mathcal{X}$, see [15, Section 3.4]. Alternatively, we could ensure a constant variance reduction of the estimator, see [2].
- The procedure provides an intrinsic stopping criterion yielding a certain stability of estimates, see Section 3 and the simulations in Figures 1 and 2. Hence, the maximal bandwidth $h^{(k^*)}$, specified by the maximal number of iterations k^*, is only bounded by the available computation time.

Note, that the input parameters, the non-adaptive weights $\overline{w}_{ij}^{(k)}$ and their sum $\overline{N}_i^{(k)}$ are deterministic while the adaptive weights $\tilde{w}_{ij}^{(k)}$ and their sum $\tilde{N}_i^{(k)}$ are random due to the data-driven statistical penalty $s_{ij}^{(k)}$. In particular, we emphasize that the Propagation-Separation Approach does not use adaptive parameters. It is adaptive in the sense that the returned estimator function $\tilde{\theta}^{(k^*)}(.)$ is based upon structure-adaptive weights $\tilde{w}_{ij}^{(k)}$, which describe the homogeneity regions of the unknown parameter function θ.

Remark 2.8 (Original procedure)**.** In Algorithm 1, we omitted the memory step. In order to get the original version of the Propagation-Separation Approach as introduced in [15, Section 3.3] the memory step can be included in the following manner. Denoting the aggregated estimator by $\hat{\theta}_i^{(k)}$ the procedure is initialized with $\hat{\theta}_i^{(0)} := \overline{\theta}_i^{(0)}$ and $\hat{N}_i^{(0)} := \overline{N}_i^{(0)}$ for all i, see item (2) in Algorithm 1. Then, we relax the adaptive estimator $\tilde{\theta}_i^{(k)}$ in Equation (2.5) by adding in item (3) of Algorithm 1 the additional step

$$\hat{\theta}_i^{(k)} := \eta_i \tilde{\theta}_i^{(k)} + (1 - \eta_i)\hat{\theta}_i^{(k-1)}, \quad \text{where} \quad \eta_i := (1 - \eta_0) K_{\text{me}}\left(m_i^{(k)}/\tau\right). \tag{2.6}$$

This uses the memory kernel $K_{\text{me}} : [0, \infty) \to [0, 1]$, the memory bandwidth $\tau > 0$, the minimal memory effect $\eta_0 \in [0, 1)$ and the memory penalty

$$m_i^{(k)} := \overline{N}_i^{(k-1)} \mathcal{KL}(\tilde{\theta}_i^{(k)}, \hat{\theta}_i^{(k-1)}) \quad \text{with} \quad \overline{N}_i^{(k-1)} = \sum_j \overline{w}_{ij}^{(k)},$$

which measures the difference between the new adaptive estimator $\tilde{\theta}_i^{(k)}$ and the aggregated estimator $\hat{\theta}_i^{(k-1)}$ of the previous iteration step. Additionally, we replace the statistical penalty

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

and define

$$\hat{N}_i^{(k-1)} := \eta_i \tilde{N}_i^{(k-1)} + (1-\eta_i)\hat{N}_i^{(k-1)}.$$

This leads to the modified output $\hat{\theta}_i^{(k^*)}$ for all $i \in \{1, \dots, n\}$ in item (4). In Section 5, we discuss the impact of the memory step. There, we concentrate on the question whether the memory step is needed to obtain the properties of the algorithm shown in [15, Section 5] and if, where.

Both the original and the simplified Propagation-Separation Approach provide a sequence of estimates with, in general, decreasing variances. Here, the adaptivity of the weights may be interpreted as a stopping criterion that leads to a similar model selection as Lepski's method [6, 11]. More precisely, the Propagation-Separation Approach and Lepski's method yield an estimator which balances the trade-off between the decreasing variance and the increasing bias when different homogeneity regions are included into the estimator.

2.3. Propagation condition

As mentioned above, an appropriate choice of the adaptation bandwidth λ is crucial for the behavior of the algorithm. Polzehl and Spokoiny [15, Section 3.5] suggested a choice, 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 free 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 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, in the setting of Algorithm 1, the verification of propagation and stability under (local) homogeneity. Additionally, it provides a better interpretability than earlier formulations, see e.g. [17]. In [18], the authors presented a similar approach in the context of model selection using a *propagation condition* for the choice of the critical values $\mathfrak{z}_k$ determining some confidence intervals. Here, we consider quantiles instead of confidence intervals.

Under homogeneity, i.e. if $\theta(.) \equiv \theta$, Equation (A.1) in Appendix A shows that the non-adaptive estimator satisfies $\mathbb{P}(\overline{N}_i^{(k)} \mathcal{KL}(\overline{\theta}_i^{(k)}, \theta) > z) \leq 2e^{-z}$ for all $i \in \{1, \dots, n\}$ and every $k \in \{0, \dots, k^*\}$. Hence, $\mathcal{KL}(\overline{\theta}_i^{(k)}, \theta)$ decreases at

least with rate $\overline{N}_i^{(k)}$. The following condition in Definition 2.9 ensures a similar behavior for the adaptive estimator. We introduce the function

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

defined as

$$\mathfrak{Z}_\lambda(k, p; \theta, i) := \inf \left\{ z > 0 : \mathbb{P}\left(\overline{N}_i^{(k)} \mathcal{KL}(\tilde{\theta}_i^{(k)}(\lambda), \theta) > z \right) \leq p \right\}, \tag{2.7}$$

where $\tilde{\theta}_i^{(k)}(\lambda)$ denotes the adaptive estimator in position Xi resulting from the Propagation-Separation Approach with adaptation bandwidth $\lambda > 0$ and observations $Y_j \sim \mathbb{P}_\theta$ for all $j \in \{1, \dots, n\}$, i.e. $\theta(.) \equiv \theta$.

Definition 2.9 (Propagation condition)**.** We say that λ is chosen in accordance with the propagation condition at level $\epsilon > 0$ for $\theta \in \Theta$ if the function $\mathfrak{Z}_\lambda(., p; \theta, i)$ in Equation (2.7) is non-increasing for all $p \in (\epsilon, 1)$ and every $i \in \{1, \dots, n\}$.

As before, the propagation condition is formulated w.r.t. 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 all values θ_i with $i \in \{1, \dots, n\}$. At best, the choice of λ by the propagation condition is invariant w.r.t. the underlying parameter θ. The study in Section 4.1 points out that this is the case for Gaussian and exponential distribution and as a consequence for log-normal, Rayleigh, Weibull, and Pareto distribution. Else, we recommend to identify some parameter θ^* yielding a sufficiently large choice of the adaptation bandwidth λ such that the propagation condition remains valid for all values θ_i, $i \in \{1, \dots, n\}$, see Section 4.1 for more details.

Remark 2.10.

- In Section 4.1, we consider some examples of the propagation condition with Gaussian, exponential and Poisson distribution, see Figures 3, 4, and 5.
- In Theorem 1 we need ϵ to be strictly smaller than $1/n$. However, this is based on a quite rough upper bound. In practice, it seems advantageous to choose ϵ appropriately for the respective application.
- The probability $\mathbb{P}(\overline{N}_i^{(k)} \mathcal{KL}(\tilde{\theta}_i^{(k)}(\lambda), \theta) > z)$ cannot be calculated exactly. In Section 4.2, we introduce an appropriate approximation which can be used in practice.

The propagation condition yields a lower bound for the choice of λ. 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 2.11. Let $\lambda_{\min} > 0$ be fixed and consider the set

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

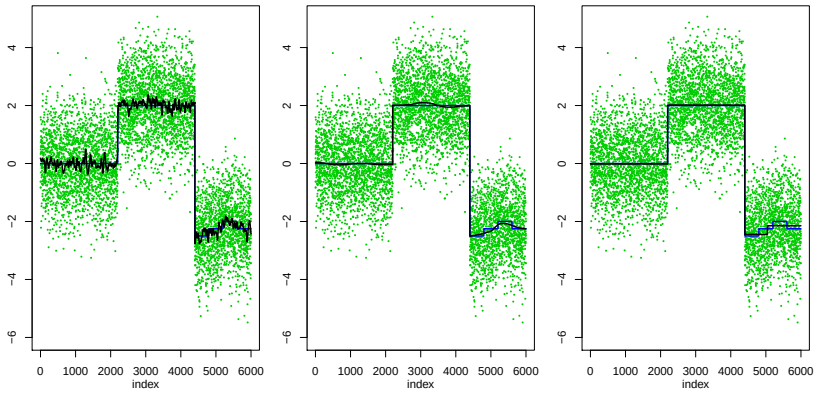

FIG 1. *Results of Algorithm 1 (black line) for the piecewise constant parameter function* $\theta_1(.)$ *(blue line) with adaptation bandwidth* $\lambda_1 = 14.6$ *and location bandwidths (f.l.t.r.)* $h_1 = 26.6, 388, 5640$. *The green circles represent the Gaussian observations.*

Then, we introduce

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

2.4. *Some heuristic observations*

In order to provide some intuition, we illustrate the general behavior of Algorithm 1 on two examples, see Figures 1 and 2. We apply the R-package `aws` [13]. Here, the memory step is omitted by default. It can be included setting `memory = TRUE`.

On $\mathcal{X} := \{1, \ldots, 6000\}$, the first test function is piecewise constant

$$\theta_1(x) := \begin{cases} 0, & \text{if } x \in \{1, \ldots, 2200\} \\ 2, & \text{if } x \in \{2201, \ldots, 4400\} \\ -2.5, & \text{if } x \in \{4401, \ldots, 4800\} \\ -2.25, & \text{if } x \in \{4801, \ldots, 5200\} \\ -2, & \text{if } x \in \{5201, \ldots, 5600\} \\ -2.25, & \text{if } x \in \{5601, \ldots, 6000\} \end{cases}$$

and the second one is piecewise polynomial

$$\theta_2(x) := \begin{cases} x/1500, & \text{if } x \in \{1, \ldots, 1500\} \\ 4 + ((x/100 - 27)/6)^2/2, & \text{if } x \in \{1501, \ldots, 4500\} \\ -1 - (x/300 - 15), & \text{if } x \in \{4501, \ldots, 6000\}. \end{cases}$$

The observations follow a Gaussian distribution, i.e. $Y_i \sim \mathcal{N}(\theta(X_i), 1)$.

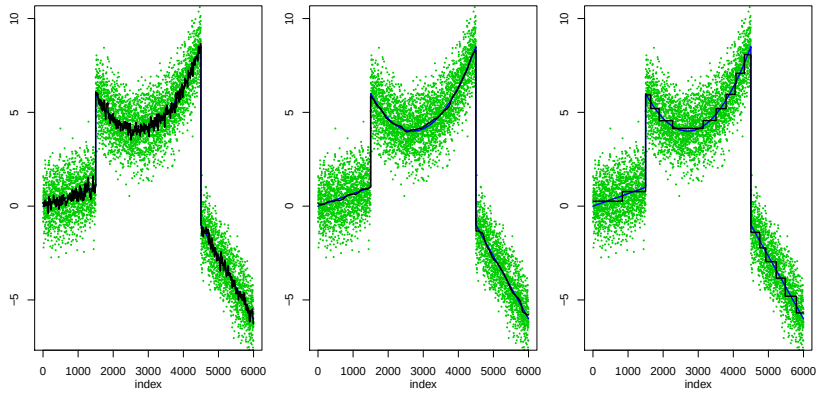

FIG 2. *Results of Algorithm 1 (black line) for the piecewise polynomial parameter function* $\theta_2(.)$ *(blue line) with adaptation bandwidth* $\lambda_2 = 14.6$ *and location bandwidths (f.l.t.r.)* $h_2 = 13.6, 127, 5640$. *The green circles correspond to the Gaussian observations.*

The plots were provided by the function `aws` setting `hmax` $:= h^{(k^*)} := 6000$ and `lkern = "Triangle"`, such that

$$K_{\text{loc}}(x) := (1 - x^2)_+ \quad \text{and} \quad K_{\text{ad}}(x) := \min\{1, 2 - 2x\}_+. \tag{2.8}$$

The adaptation bandwidth $\lambda = 14.6$ was chosen in accordance with the propagation condition of level $\epsilon = 5 \cdot 10^{-4}$, see Section 2.3 for the definition and Section 4.1 for a study of its invariance w.r.t. the parameter θ in case of Gaussian observations.

In Figure 1, we show the results for the piecewise constant function $\theta_1(.)$ with increasing location bandwidths $h_1 = 26.6, 388, 5640$ corresponding to the iteration steps $k_1 = 14, 24, 41$. Figure 2 is based on the piecewise smooth function $\theta_2(.)$ setting $h_2 = 13.6, 127, 5640$, that is $k_2 = 14, 24, 41$. For both examples, it holds $k^* = 41$ representing the final iteration step. In the steps $k_1 = k_2 = 24$ the MSE is minimal.

We summarize the following heuristic observations.

- Homogeneous regions with sufficiently large discontinuities are separated by the algorithm leading to a consistent estimator, see $x \in \{1, \ldots, 4400\}$ in Figure 1.
- If the discontinuities are too small, separation fails. Then, different homogeneous regions are treated as one yielding a bounded estimation bias. This is illustrated in the right part of Figure 1, where $x \in \{4401, \ldots, 6000\}$.
- In Figure 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 depends mainly on the smoothness of the parameter function $\theta(.)$ and the adaptation bandwidth λ. However, the estimation bias can be reduced by an accurate

stopping criterion. The maximal location bandwidth $h^{(k^*)}$ should be chosen such that the non-adaptive estimator in Definition 2.5 behaves good within regions without discontinuities. Then, supposing an appropriate choice of the adaptation bandwidth λ, within these regions, Algorithm 1 would yield similar results as non-adaptive smoothing while smoothing among distinct regions would be avoided as sharp discontinuities could be detected by the adaptive weights. Such a choice of k^* can be advantageous under model misspecification, but this is beyond the scope of this article. In case that a local constant model with sharp discontinuities is valid we will deduce stability results in Proposition 3.1 and Section 3.3 showing that no stopping criterion is needed.

Thus, the heuristic properties are quite clear. However, the iterative approach complicates a theoretical verification considerably. Therefore, in Section 3 we concentrate on piecewise constant functions with sharp discontinuities. Here, our new propagation condition, see Section 2.3, ensures propagation within homogeneous regions and stability of estimates due to separation of distinct regions. The case of model misspecification will be analyzed in an upcoming study.

3. Theoretical properties

Now, we analyze the behavior of the algorithm in more detail. First, we consider a homogeneous setting, where propagation and stability of estimates follow as direct consequence of the propagation condition. Then, we show the separation property. For locally constant parameter functions with sufficiently sharp discontinuities this restricts smoothing to the respective homogeneous regions yielding again propagation and a certain stability of estimates. Throughout this section, we assume that we have identified λ and ϵ such that the propagation condition holds.

3.1. Propagation and stability under homogeneity

We show for a homogeneous setting that the propagation condition yields with Equation (A.1) in Appendix A 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 3.1. *Suppose $\theta(.) \equiv \theta$, Assumption (1), and let the adaptation bandwidth λ be chosen in accordance with the propagation condition at level ϵ for $\theta \in \Theta$. Then, for each $i \in \{1,\dots,n\}$, $k \in \{0,\dots,k^*\}$, and all $z > 0$, it holds*

$$\mathbb{P}\left(\overline{N}_i^{(k)}\mathcal{KL}\left(\tilde{\theta}_i^{(k)},\theta\right) > z\right) \le \max\left\{2e^{-z},\epsilon\right\}. \tag{3.1}$$

In particular, we get for all $k' \ge k$ that

$$\mathbb{P}\left(\overline{N}_i^{(k')}\mathcal{KL}\left(\tilde{\theta}_i^{(k')},\theta\right) > z\right) \le \max\left\{\mathbb{P}\left(\overline{N}_i^{(k)}\mathcal{KL}\left(\tilde{\theta}_i^{(k)},\theta\right) > z\right),\epsilon\right\}. \tag{3.2}$$

Proof. Equation (3.2) follows from the propagation condition, which ensures that the function $\mathfrak{Z}_\lambda(.,p;\theta,i)$ in Equation (2.7) is non-increasing for all $p \in (\epsilon,1)$ and every $i \in \{1,\dots,n\}$. Since, see item (2) in Algorithm 1, we have $\tilde\theta_i^{(0)} = \overline\theta_i^{(0)}$ this yields

$$\begin{aligned}\mathbb{P}\left(\overline{N}_i^{(k)}\mathcal{KL}\left(\tilde\theta_i^{(k)},\theta\right) > z\right) &\overset{\text{Eq.(3.2)}}{\leq} \max\left\{\mathbb{P}\left(\overline{N}_i^{(0)}\mathcal{KL}\left(\overline\theta_i^{(0)},\theta\right) > z\right),\epsilon\right\}\\ &\overset{\text{Eq. (A.1)}}{\leq} \max\left\{2e^{-z},\epsilon\right\},\end{aligned}$$

leading to the assertion. □

Proposition 3.1 yields with $z := \mu\log(n)$ and $\epsilon := c_\epsilon n^{-\mu}$, where $c_\epsilon > 0$ and $\mu > 2$ that

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

If $h^{(k^*)}$ is sufficiently large such that $\overline{N}_i^{(k^*)}$ is of order n this leads with Equation (2.4) to the root-n consistency of $\tilde\theta_i^{(k)}$ up to a log-factor. This additional log-factor results from the adaptivity as discussed in [7]. However, asymptotic results are problematic in this context as we discuss in Section 5. Therefore, we prefer to consider Proposition 3.1 as error bound for a fixed iteration step k, where the results for local homogeneity in Section 3.3 are based on.

3.2. Separation property

For considerably different parameter values the corresponding adaptive weights become zero, see Proposition 3.2 below. The result is similar to the first part of [15, Theorem 5.9]. It implies that different homogeneous regions with sufficiently large discontinuities will be separated by the algorithm. In particular, we will see, that the lower bound for the discontinuities allowing exact separation of the distinct regions depends mainly on the adaptation bandwidth λ and the achieved quality of estimation in the previous iteration step. Remember that the adaptive weights $\tilde w_{ij}^{(k)}$ and their sum $\tilde N_i^{(k)}$ are random. In the proofs, we apply Equation (A.2) in Appendix A, which requires that $\tilde\theta_i^{(k)} \in \Theta_\varkappa$.

Proposition 3.2 (Separation property)**.** *Suppose Assumptions (1). We consider two points X_{i_1} and X_{i_2} providing in iteration step k the estimation accuracy $\mathcal{KL}(\tilde\theta_{i_m}^{(k)},\theta_{i_m}) \leq z_m^{(k)} := z/\overline{N}_{i_m}^{(k)}$ with some constant $z > 0$ and $\theta_{i_m},\tilde\theta_{i_m}^{(k)} \in \Theta_\varkappa$ with $\Theta_\varkappa$ as in Lemma 2.3, $m = 1,2$. If*

$$\mathcal{KL}^{1/2}(\theta_{i_1},\theta_{i_2}) > \varkappa\left(\sqrt{\lambda/\tilde N_{i_1}^{(k)}} + \sqrt{z_1^{(k)}} + \sqrt{z_2^{(k)}}\right) \tag{3.3}$$

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

Proof. Due to the compact support of the adaptation kernel K_{ad}, it suffices to show that the statistical penalty introduced in item (3) of Algorithm 1 satisfies $s_{i_1 i_2}^{(k+1)} > \lambda$. Equation (A.2) in Appendix A yields for $\mathcal{KL}(\tilde{\theta}_{i_m}^{(k)}, \theta_{i_m}) \leq z_m^{(k)}$ with $m = 1, 2$ that

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

such that

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

by Equation (3.3). □

Remark 3.3. The lower bound (3.3) holds if

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

This emphasizes the impact of the involved sample sizes.

3.3. Propagation and stability under local homogeneity

Next, we consider a locally homogeneous setting with sharp discontinuities, formally described in Assumption (2). In this case, smoothing is restricted to the homogeneous regions leading to similar results as under homogeneity, that is, to propagation and to stability of estimates. We introduce some auxiliary notions.

Notation 3.4.

- $\mathfrak{C}(M)$ is the smallest connected set that includes the respective set M, i.e.

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

- We call the discrete set $M := \{X_{l_j}\}_{j=1}^m \subseteq \mathcal{X}$ connected if

$$X_j \in M \Leftrightarrow X_j \in \mathfrak{C}(M) \qquad \text{for all} X_j \in \mathcal{X}.$$

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

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

Assumption 2 (Structural assumption)**.** There is a non-trivial partition $\mathcal{V} := \{\mathcal{V}_i\}_i$ of $\mathcal{X}$ into maximal homogeneity regions, i.e. for each $X_i \in \mathcal{X}$ there are a convex neighborhood $\mathcal{V}_i \subseteq \mathcal{X}$ and a constant $\varphi_i > 0$ such that

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

The convexity of the neighborhoods $\{\mathcal{V}_i\}_{i=1}^n$ ensures the comparability of the homogeneous setting in Proposition 3.1 and the setting within each of these neighborhoods. A violation of this condition may lead to another 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 Equation (A.1) and hence the probability condition do not depend thereon.

We deduce the propagation property for the present case of local homogeneity. Here, we should take into account that the considered neighborhood $U_i^{(k)}$ 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 3.5. We define for each $i \in \{1, \dots, n\}$ and $k \in \{0, \dots, k^*\}$ the effective sample size and its local minimum

$$\overline{n}_i^{(k)} := \sum_{X_j \in \mathcal{V}_i \cap U_i^{(k)}} \overline{w}_{ij}^{(k)} \qquad \text{and} \qquad n_i^{(k)} := \min_{X_j \in U_i^{(k)}} \overline{n}_j^{(k)}. \tag{3.4}$$

As it turns out, the quantities $n_i^{(k)}$ determine the minimal stepsizes φ_i such that a discontinuity will be detected. 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 the events

$$\mathcal{B}^{(k)}(z) := \left\{ \overline{n}_i^{(k)} \mathcal{KL}(\tilde{\theta}_i^{(k)}, \theta_i) \leq z \quad \text{for all } i \right\}, \qquad z > 0.$$

and

$$M^{(k')}(z) := \bigcap_{k=0}^{k'} \bigcap_{i=1}^{n} \left\{ \varphi_i > \varkappa \left[\sqrt{\lambda / \tilde{N}_i^{(k)}} + 2\sqrt{z / n_i^{(k)}} \right] \right\}, \tag{3.5}$$

where $\varphi_i > 0$ is as in Assumption (2). In $\mathcal{B}^{(k)}(z)$ the estimation error is bounded from above and in $M^{(k')}(z)$ the discontinuities are sufficiently large for separation.

We confine our subsequent analysis to the favorable realizations $\{T(Y_i) \in \Theta_\varkappa \text{ for all } i\}$ and quantify the probability $p_\varkappa$ of its complementary set, see Appendix A for further details. We also restrict the range of $\theta(.)$ by the subset $\Theta^* \subseteq \Theta$ which may influence the value of $p_\varkappa$.

Theorem 1 (Propagation property under local homogeneity)**.** *Suppose Assumptions (1) and (2) and let the bandwidth λ be chosen in accordance with the propagation condition at level ϵ for all θ_i, $i \in \{1, \dots, n\}$. If $\mathbb{P}\left(M^{(k')}(z)\right) > 0$ then it holds*

$$\begin{aligned} & \mathbb{P}\left(\mathcal{B}^{(k')}(z) | M^{(k')}(z)\right) \\ \geq \; & 1 - \left[p_\varkappa + (k'+1) \max\left\{2ne^{-z}, n\epsilon\right\}\right] / \mathbb{P}\left(M^{(k')}(z)\right), \end{aligned} \tag{3.6}$$

where $p_\varkappa$ is as in Notation A.1.

Remark 3.6.

- In Equation (3.6), we observe an additional factor $(k+1)$, which appeared in the propagation property of Polzehl and Spokoiny [15] as well, see Equation (5.1) in Section 5, below. This factor results from the proof only and might be avoidable. In particular, we notice that the given bound is not sharp as we did not take advantage of the intersections of the sets $(\mathcal{B}^{(k)}(z))^c$ in Equation (B.2) of the proof. The above theorem provides a meaningful result for $z \geq q\log(n)$ and $\epsilon := c_\epsilon n^{-q}$ with $c_\epsilon > 0$ and $q > 1$. It is the better the smaller $p_\varkappa$ and the larger the discontinuities φ_i which implies that $\mathbb{P}(M^{(k')}(z))$ is close to one. Remind that $p_\varkappa = 0$ for Gaussian and log-normal distributed observations.
- Separation 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, denoted by $n_i^{(k)}$, determines the lower bound (3.5) for the discontinuities that provide an exact separation of the distinct homogeneous regions. This bound is closely related to Equation (3.3) that involves only two points such that the term $2/\sqrt{n_i^{(k)}}$ from Equation (3.5) can be replaced by

$$\left(1/\sqrt{\overline{N}_{i_1}^{(k)}} + 1/\sqrt{\overline{N}_{i_2}^{(k)}}\right)$$

  having the same effect.

Finally, we deduce a similar result as in Equation (3.2) under local homogeneity. Thus, we infer from the estimation quality in iteration step k_1 on the estimation quality in step $k_2 > k_1$. To this end, we apply again the separation property, see Proposition 3.2. This requires sure knowledge on the previously achieved estimation quality. Therefore, we consider the conditional probability and verify an exponential bound.

Proposition 3.7 (Stability of estimates under local homogeneity)**.** *In the situation of Theorem 1, it holds for all $k_1, k_2 \in \{0, \dots, k^*\}$ with $k_1 < k_2 \leq k'$ such that $2p_\varkappa + (k_2+1)\max\{2ne^{-z}, n\epsilon\} < \mathbb{P}(M^{(k_2)}(z))$ that*

$$\begin{aligned}&\mathbb{P}\left(\mathcal{B}^{(k_2)}(z) | \mathcal{B}^{(k_1)}(z) \cap M^{(k_2)}(z)\right)\\ \geq\ &\frac{\mathbb{P}\left(M^{(k_2)}(z)\right) - 2p_\varkappa - (k_2+1)\max\left\{2ne^{-z}, n\epsilon\right\}}{\mathbb{P}\left(M^{(k_2)}(z)\right) - p_\varkappa - (k_1+1)\max\left\{2ne^{-z}, n\epsilon\right\}}.\end{aligned}$$

Remark 3.8. The assumptions on the choices of k_1 and k_2 ensure that the lower bound in Proposition 3.7 is larger than zero and smaller than one. This lower bound for the conditional probability $\mathbb{P}(\mathcal{B}^{(k_2)}(z)|\mathcal{B}^{(k_1)}(z) \cap M^{(k_2)}(z))$ improves the lower bound of $\mathbb{P}(\mathcal{B}^{(k_2)}(z)|M^{(k_2)}(z))$ in Theorem 1 if $p_\varkappa$ is small and $\mathbb{P}(M^{(k_2)}(z))$ is large, that is, if $\Theta_\varkappa$ and the discontinuities are sufficiently large. However, the result allows a comparison of the established lower bounds only, but not of the exact probabilities.

4. Justification of the propagation condition

In this section, we dwell into the propagation condition and discuss its application in practice.

4.1. Invariance of the propagation condition w.r.t. the parameter

The propagation condition in Definition 2.9 is formulated w.r.t. the unknown parameter $\theta \in \Theta$. In this section, we evaluate its variability w.r.t. this parameter. To this end, we start with a more general problem yielding a sufficient criterion. This criterion suggests the invariance of the propagation condition w.r.t. the parameter θ in case of Gaussian and exponential distribution and as a consequence of log-normal, Rayleigh, Weibull, and Pareto distribution. Additionally, we discuss the choice of λ if the associated function $\mathfrak{Z}_\lambda$ in Equation (2.7) is not invariant w.r.t. the parameter θ, where we concentrate on the Poisson distribution.

We introduce a general criterion for the invariance of the composition of two functions w.r.t. some parameter θ.

Proposition 4.1. *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 $\Omega^f_\theta := \{y : (y, \theta) \in \Omega^f\}$, $f_\theta : \Omega^f_\theta \to \mathbb{R}$ with $f_\theta(y) := f(y, \theta)$, and analogous Ω^g and g_θ. Then, we suppose $g_\theta(\Omega^g_\theta) \subseteq \Omega^f_\theta$ and $|\frac{\partial g_\theta}{\partial y}| > 0$, such that the composition $f_\theta \circ g_\theta^{-1} : g_\theta(\Omega^g_\theta) \to \mathbb{R}$ is well-defined. The function*

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

is invariant w.r.t. θ if a variable $\zeta(y, \theta)$ and functions $\tilde{f}$ and $\tilde{g}$ exist such that

$$\tilde{f}(\zeta) = f_\theta(y) \quad \text{and} \quad \tilde{g}(\zeta) = g_\theta(y). \tag{4.1}$$

Now, we are well prepared to evaluate the invariance of the propagation condition in Definition 2.9, and hence of the choice of λ, w.r.t. the parameter θ. The estimator is defined as linear combination of the terms $T(Y_j)$, where the adaptive and the non-adaptive estimator differ only 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 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 case of the *adaptive* estimator we cannot follow the same approach. This would require knowledge about the probability distribution of the random variables $\tilde{w}_{ij}^{(k)} T(Y_j)/\tilde{N}_i^{(k)}$, where the adaptive weights follow an unknown distribution. Furthermore, these variables are not statistically independent. 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 w.r.t. θ if the same holds true w.r.t. the non-adaptive estimator. The single observation case is treated first.

Lemma 4.2. *Let $\mathcal{P} = \{\mathbb{P}_\theta\}_{\theta\in\Theta}$ with $\Theta \subseteq \mathbb{R}$ be a parametric family of continuous probability distributions. Suppose that $Y \sim \mathbb{P}_\theta$ and $T(Y) \in \Theta$ almost surely, and that the density f_θ^Y of Y is continuously differentiable. Consider the random variable $Z := g_\theta(Y) := \mathcal{KL}(\mathbb{P}_{T(Y(\omega))}, \mathbb{P}_\theta)$, and assume that $\frac{\partial g_\theta}{\partial y} \neq 0$. The density f_θ^Z of Z is invariant w.r.t. the parameter θ if a variable $\zeta(y,\theta)$ and functions $\tilde{f}$ and $\tilde{g}$ exist such that*

$$\tilde{f}(\zeta) = f_\theta^Y(y) \cdot \left|\frac{\partial g_\theta}{\partial y}(y)\right|^{-1} \quad \text{and} \quad \tilde{g}(\zeta) = g_\theta(y). \tag{4.2}$$

Proof. The assertion follows as special case of Proposition 4.1 with

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

since $\mathbb{P}_\theta(|\frac{\partial g_\theta}{\partial y}(y)| > 0) = \mathbb{P}_\theta(T(Y) \neq \theta) = 1$. □

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

Example 4.3. We consider the same setting as in Lemma 4.2. In the following cases, the density of Z is invariant w.r.t. the parameter θ.

- $\mathcal{P} = \{\mathcal{N}(\theta, \sigma^2)\}_{\theta\in\Theta}$ with $\sigma > 0$ fixed: Equation (2.3) and Table 1 yield for the Kullback-Leibler divergence of $\mathbb{P}_\theta, \mathbb{P}_{\theta'} \in \mathcal{P}$ the explicit formula

$$\mathcal{KL}(\theta, \theta') = \frac{(\theta-\theta')^2}{2\sigma^2} \quad \text{such that} \quad \frac{\partial g_\theta}{\partial y}(y) = \frac{y-\theta}{\sigma^2}.$$

  Since $f_\theta^Y(y) = \exp(-\frac{(y-\theta)^2}{2\sigma^2})/\sqrt{2\pi\sigma^2}$ we get the invariance w.r.t. θ from Lemma 4.2 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: It holds $f_\theta^Y(y) = \frac{y^{p-1}\, e^{-y/\theta}}{\theta^p \Gamma(p)}$, such that

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

  Thus, Lemma 4.2 can be applied with

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

This extends to non-adaptive linear combinations as follows. Lemma 4.2 can be applied w.r.t. the non-adaptive estimator with $Y := \overline{\theta}_i^{(k)}$ 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 $Y_j \sim \mathbb{P}_\theta \in \mathcal{P}$. Hence, it depends directly on the function $T(.)$ introduced in Assumption (1).

Theorem 2. *Let $\mathcal{P} = \{\mathbb{P}_\theta\}_{\theta\in\Theta}$ with $\Theta \subseteq \mathbb{R}$ be a parametric family of probability distributions. We consider the random variable*

$$Z := g_\theta(\overline{\theta}_i^{(k)}) := \left[\omega \mapsto \mathcal{KL}\left(\mathbb{P}_{\overline{\theta}_i^{(k)}(\omega)}, \mathbb{P}_\theta\right)\right],$$

where $\overline{\theta}_i^{(k)}$ denotes the non-adaptive estimator depending on the observations $Y_j \overset{\text{iid}}{\sim} \mathbb{P}_\theta$ with $j \in \{1,\dots,n\}$ and some $\theta \in \Theta$. The density of Z is invariant w.r.t. the parameter θ in the following cases.

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

The density of the convolution of exponential distributions has been studied for instance in [1].

Remark 4.4. The following is known from Example 4.3: The random variable $[\omega \mapsto \mathcal{KL}(\mathbb{P}_{T(Y(\omega))}, \mathbb{P}_\theta)]$ is invariant w.r.t. the parameter θ if the observations follow a Gamma distribution. However, the probability distribution of the corresponding non-adaptive estimator has a quite sophisticated form [10, 12], where the corresponding summands could not been proven to be invariant w.r.t. θ. Though, in case of a location kernel that attains only values in $\{0, 1\}$ we get

$$Y_j \overset{\text{iid}}{\sim} \Gamma(p, \theta) \implies \overline{\theta}_i^{(k)} \sim \Gamma(\overline{N}_i^{(k)} p, \theta/\overline{N}_i^{(k)}) \qquad \text{if } \overline{w}_{ij}^{(k)} \in \{0, 1\} \text{ for all } j.$$

This yields via Example 4.3 the invariance w.r.t. θ. The same holds true for the Erlang and scaled chi-squared distribution since

$$\text{Erlang}(n, 1/\theta) = \Gamma(n, \theta)$$

and

$$Y \sim \Gamma(k/2, 2\theta/k) \text{ if } kY/\theta \sim \chi^2(k) = \Gamma(k/2, 2).$$

The new propagation condition is included into the R-package `aws` [13]. Simulation tests yield smaller values of the adaptation bandwidth λ than the previous version of the propagation condition, hence allowing for better smoothing results with a smaller estimation bias.

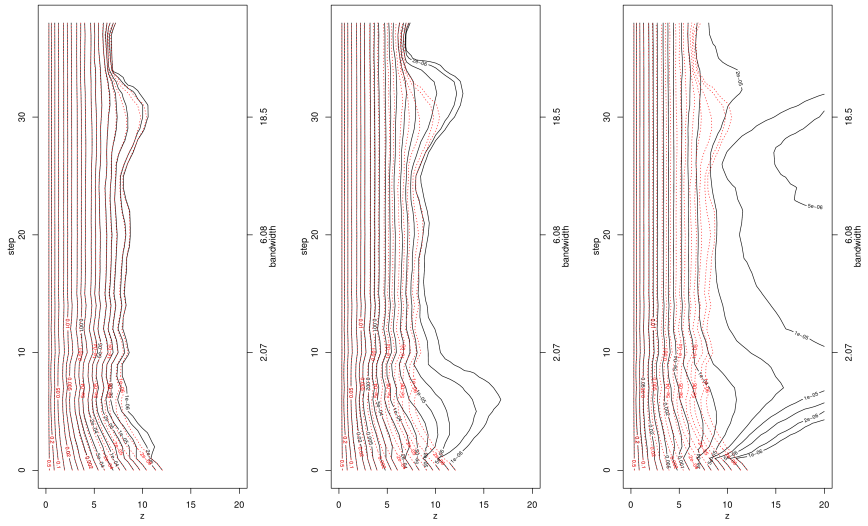

FIG 3. *Plots of the propagation condition for the Gaussian distribution with (f.l.t.r.)* $\lambda = 22.4, 13.6, 9.72$. *The isolines of the probability* p *for values between* 10^{-6} *and* 0.5 *are plotted w.r.t. the location bandwidth* $h^{(k)}$ *described by the iteration step* k *and the corresponding value* $z = \mathfrak{Z}_\lambda(k, p; \theta = 1, i)$ *for some* $i \in \{1, \dots, n\}$. *The black solid lines represent the isolines of the adaptive estimator, the red dotted lines correspond to the non-adaptive estimator.*

In Figures 3 and 4, we show some examples to illustrate the close relation of the adaptive and the non-adaptive estimator under a satisfied propagation condition. The plots have been realized using the function `awstestprop` on a two-dimensional design with 5000×5000 points and the same kernels as in Equation (2.8). 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, we show the results for the Gaussian distribution with three different values of λ. In Figure 4, we consider the same setting w.r.t. the exponential distribution. Both Theorem 2 and the numerical simulations suggest the invariance of the propagation condition w.r.t. the parameter θ.

Finally, we discuss how to proceed if the function $\mathfrak{Z}_\lambda$ in Equation (2.7) varies with the parameter θ. We want to ensure that our choice of the adaptation bandwidth λ is in accordance with the propagation condition for all θ_i, $i \in \{1, \dots, n\}$. Certainly, we do not know the exact parameters $\{\theta_i\}_i$. Instead, we could analyze the monotonicity of the optimal choice $\lambda_{\mathrm{opt}}(\epsilon, \theta, \lambda_{\min})$, see Remark 2.10, 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, we get for every θ 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 finite number of

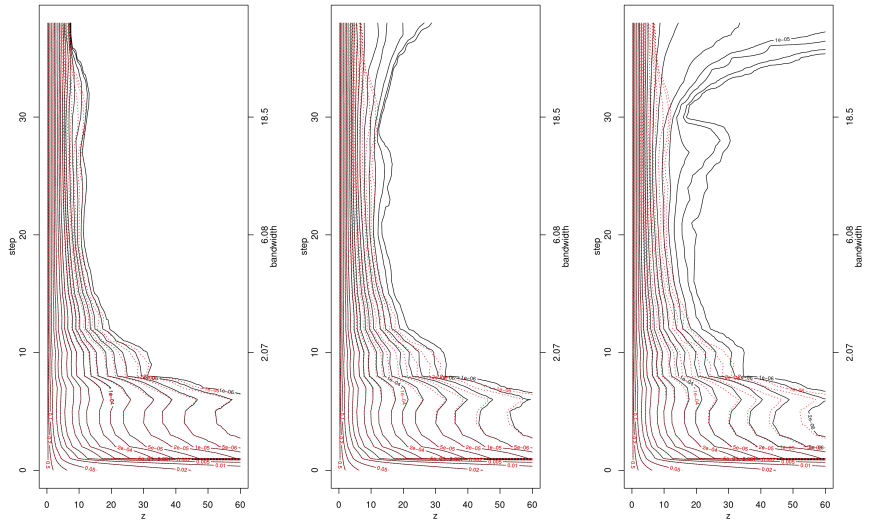

FIG 4. *Plots of the propagation condition for the exponential distribution with (f.l.t.r.)* $\lambda = 13.2, 10.2, 8.78$.

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, $i \in \{1, \dots, n\}$.

For observations following 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 5, where we consider the same kernels as in Equation (2.8), a regular design with 5000×5000 points, and $h^{(k^*)} = 50$, i.e. 38 iteration steps. In case of Bernoulli distributed observations it seems to be recommendable to ensure the propagation condition for $\theta^* := 0.5$. In both cases the implemented algorithm avoids that the Kullback-Leibler divergence becomes infinity by slightly shifting the estimator.

4.2. The propagation condition in practice

The propagation condition is based on the function $\mathcal{Z}_\lambda$. This depends on the exceedence 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 Equation (A.1) 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 noise is 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 may estimate the above probability by the relative frequency of design points $X_i \in \mathcal{X}$ with

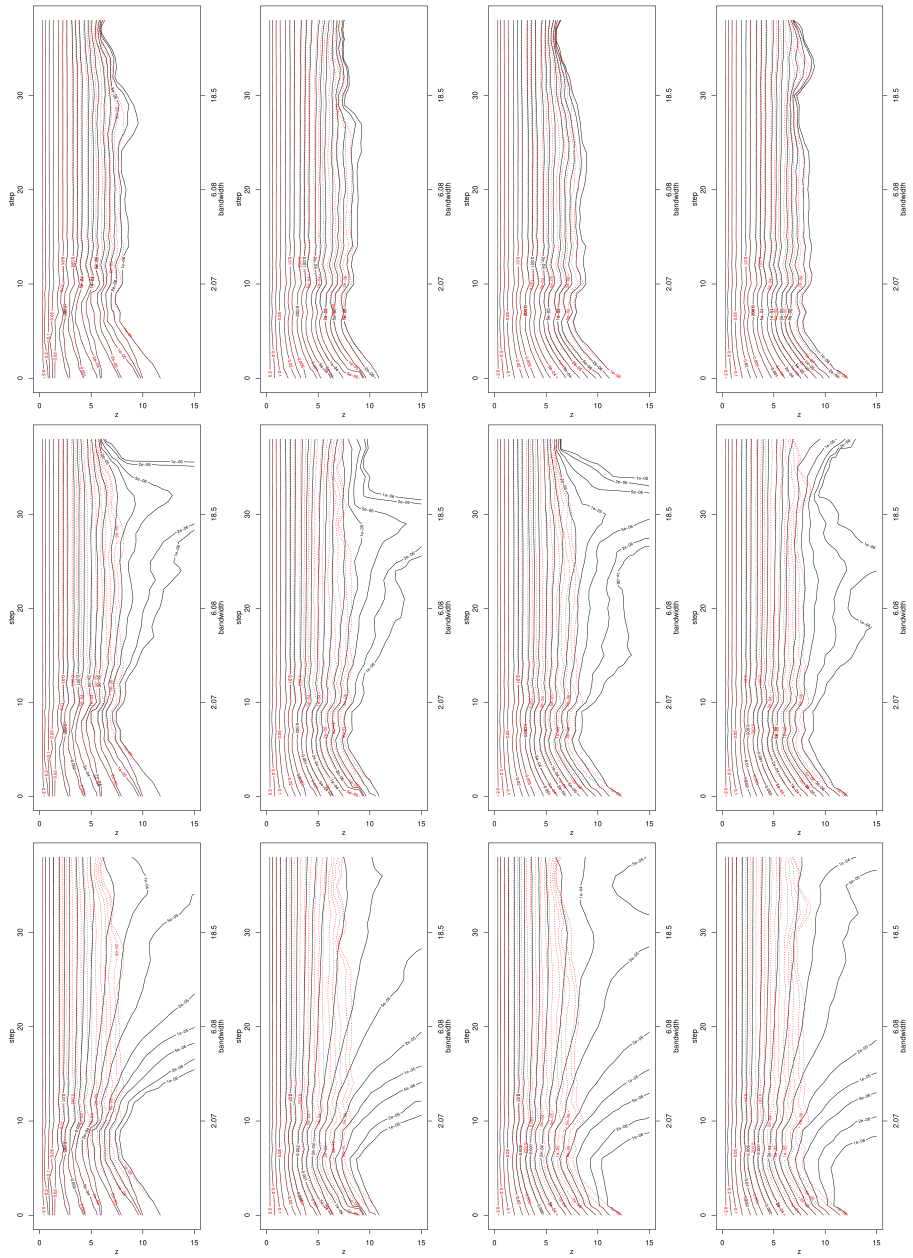

FIG 5. *Plots of the propagation condition for the Poisson distribution with (f.l.t.r.)* $\theta = 1, 10, 100, 1000$ *and (from top to bottom)* $\lambda = 13.2, 9.88, 7.69$ *yielding* $\epsilon_{13.2}(\theta) \leq 10^{-6}$, $\epsilon_{9.88}(\theta) \approx 5 \cdot 10^{-5}$, *and* $\epsilon_{7.69}(\theta) \approx 5 \cdot 10^{-4}$.

$\overline{N}_i^{(k)} \mathcal{KL}(\tilde{\theta}_i^{(k)}(\lambda), \theta) > z$ as we discuss in Definition 4.5 and Lemma 4.6. 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 points $X_i \in \mathcal{X}$ where the final neighborhood $U_i^{(k^*)}$ is not restricted by the boundaries of the considered region $\{X_i\}_{i=1}^n$. 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 4.5 (Approximation)**.** We consider the same setting as in Definition 2.9 and 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

$$\hat{p}_\lambda^{(k)}(z) := n_0^{-1} \sum_{i=1}^{n_0} \mathbf{1}_{M_\lambda^{(k)}(z)}(X_i) \tag{4.3}$$

where $\mathbf{1}$ denotes the indicator function with $\mathbf{1}_M(x) = 1$ if $x \in M$ and $\mathbf{1}_M(x) = 0$, else.

Lemma 4.6. *We consider the same setting as in Definition 2.9 and suppose the conditions of Proposition 3.1 to be satisfied. Then, it holds for each* $j \in \{1, \dots, n_0\}$ *that*

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

and

$$\mathrm{Var}\left[\hat{p}_\lambda^{(k)}(z)\right] \leq \max\{2e^{-z}, \epsilon\}. \tag{4.4}$$

Remark 4.7. Simulations are carried out using an artificial data set that ensures a sufficiently large number of effectively independent regions for estimating the propagation level on the basis of a single realization. Theorem 1 provides a meaningful result only if $\epsilon := c_\epsilon n^{-q}$ with $c_\epsilon > 0$ and $q > 1$. We approximate the probability $\mathbb{P}(\overline{N}_i^{(k)} \mathcal{KL}(\tilde{\theta}_i^{(k)}(\lambda), \theta) > z)$ by the corresponding relative frequency (4.3). This estimate can be calculated for $\epsilon \geq 1/n$ only. Additionally, it becomes unstable if ϵ is close to $1/n$. In 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, we can achieve $\epsilon := c_\epsilon n^{-q}$ with $c_\epsilon > 0$ and $q > 1$ solely by application of the propagation condition on an artificial data set with m design points, where $m \gg n$. In this case, one should evaluate carefully under which conditions the propagation condition generalizes from the artificial data set to the data set at hand.

5. Discussion

Finally, we compare our theoretical results with the study in [15] in order to clarify the impact of the memory step. In the original study by Polzehl and

Spokoiny [15], the authors demonstrated propagation, separation and stability of estimates up to some constant. We will summarize these results briefly. Apart from the separation property, all associated proofs were based on the memory step. Here, we have shown similar properties for the simplified algorithm, where the memory step is omitted. However, our results are restricted to locally constant parameter functions with sharp discontinuities. Theoretical properties of the algorithm in case of model misspecification will be analyzed in an upcoming study.

Both studies include a certain separation property, see [15, Section 5.5] and Proposition 3.2. 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 things, the statistical independence of the adaptive weights from the observations. They then showed for $\theta(.) \equiv \theta$ that

$$\mathbb{P}\left(\overline{N}_i^{(k)} \mathcal{KL}\left(\hat{\theta}_i^{(k)}, \theta\right) \leq \mu \log(n) \quad \forall i\right) > 1 - 2k/n, \qquad \mu \geq 2, \tag{5.1}$$

where $\hat{\theta}_i^{(k)}$ denotes the adaptive estimator after modification by the memory step, see Remark 2.8 and [15, Section 3.2 and 3.3]. For locally almost constant parameters the authors established a similar result. Equation (5.1) could be improved by Proposition 3.1 taking advantage of the new propagation condition introduced in Section 2.3. Setting $z := \mu \log(n)$ and $\epsilon := c_\epsilon n^{-q}$ Proposition 3.1 implies

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

where the additional factor k is avoided. Theorem 1 sheds light on the interplay of propagation and separation during iteration. Here, we do not restrict the analysis to the respective homogeneous region as in Proposition 3.1 and [15]. Instead, we use the separation property to verify the propagation property for piecewise constant 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 in Equation (3.6) differs from Equation (5.1) by the terms $p_\varkappa$ and $\mathbb{P}(M^{(k)}(z))$, only. These are required for the separation of distinct homogeneity regions.

The results on stability of estimates are difficult to compare. Our corresponding results are stated in Propositions 3.1 and 3.7. Polzehl and Spokoiny proved under weak assumptions 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

$$\overline{N}_i^{(k)} \mathcal{KL}\left(\hat{\theta}_i^{(k^*)}, \theta_i\right) \leq c \log(n), \qquad c := \varkappa^2 \left(\sqrt{c_1 C_\tau} + \sqrt{\mu}\right)^2, \tag{5.2}$$

where $\varkappa$ is as in Lemma 2.3, $\tau := C_\tau \log(n)$ denotes the bandwidth of the memory kernel and $c_1 := \varkappa^2 \nu (1 - \sqrt{\nu})^{-2}$ depends on the constant ν satisfying

$\nu_1 \leq \overline{N}_i^{(k-1)}/\overline{N}_i^{(k)} \leq \nu$ with $\nu_1, \nu \in (2/3, 1)$. Hence, the constant c might be quite large. This result allowed to verify, under smoothness conditions on the parameter function $\theta(.)$, the optimal rate of convergence. Equation (5.2) is based upon Lemma 2.3 and consequently requires that $\theta_i, \hat{\theta}_i^{(k)}, \hat{\theta}_i^{(k^*)} \in \Theta_{\varkappa}$. This leads again to the discussion in Section 3.3, not mentioned in [15]. In particular, we refer to Notation A.1, Example A.2 and Remark A.3.

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, such that $\lim_{n\to\infty} \epsilon(n) = 0$. However, the adaptation bandwidth λ depends on the propagation level ϵ. For a fixed sample size, the simulations in Section 4.1 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. Unfortunetely, a detailed analysis with varying sample sizes is hampered by the complexity of the adaptive estimator, which depends on the whole sample via the statistical penalty. The adaptation bandwidth is the crucial parameter which distinguishes the Propagation-Separation Approach from non-adaptive smoothing. Hence, an asymptotic study is useless if we do not know how the increasing sample size affects the adaptation bandwidth or if $\lim_{n\to\infty} \lambda(n) = \infty$.

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 study is restricted to piecewise constant functions. This restriction prohibits a proof of the optimal rate of convergence under smoothness conditions. Additionally, our approach is not constructed to provide asymptotic results, see above. Second, our stability results hold for piecewise constant parameter functions with sharp discontinuities, only. In other words, we lose the general stability of estimates in Equation (5.2). To ensure stability of estimates under model misspecification for the simplified algorithm will be an interesting subject for future research. The examples in Section 2.4 suggest that the estimator results in any case in a step function. However, it remains to show its immutability for large iteration steps, either by a theoretical proof or by introducing an appropriate stopping criterion.

Without memory step we lose the optimal rate and the stability of estimates under model misspecification. Nevertheless, the essential properties of the algorithm remain valid, that is, propagation and separation. Both properties follow from the adaptivity of the estimator and not from the memory step. Hence, for a local constant model with sufficiently sharp discontinuities the memory step is not needed.

6. Conclusion

This study provides theoretical properties for a simplified version of the Propagation-Separation Approach, where the memory step is omitted from the

algorithm. In particular, we have verified the following results, which may help for a better understanding of the procedure.

In Section 2.3, we introduced an advanced parameter choice strategy for the adaptation bandwidth λ. Its invariance w.r.t. the unknown parameter function is analyzed in Section 4.1, showing for the first time theoretical and numerical results that justify the propagation condition. In practice, this yields a better interpretability of the adaptation bandwidth λ due to the precise information of the propagation level. For instance, in a recent work on structural adaptive smoothing of diffusion-weighted magnetic resonance data the new propagation condition established heuristically a certain stability of the choice of λ w.r.t. the number of measured q-shells, the number of diffusion-weighted gradients and the unknown effective number of MR-receiver coils, see [3, Section 2.5] for more details. In theory, the propagation condition yields strong results on propagation and stability of estimates for piecewise constant functions with sharp discontinuities, see Section 3. This ensures a similar behavior as for the original procedure and consequently substantiates the omittance of the memory step.

Additionally, we studied the interplay of propagation and separation during iteration. Previous work considered these properties only on their own, but not their interaction. This demonstrated that the behavior of the algorithm, and hence the achievable quality of estimation, depend mainly on the extension of the homogeneous regions, on the size of the discontinuities of the parameter function $\theta(.)$, and via the adaptation bandwidth λ on the parametric family $\mathcal{P} = \{\mathbb{P}_\theta\}_{\theta\in\Theta}$ of probability distributions. Future research may concentrate on the case of model misspecification in order to justify the heuristic observations in Section 2.4, mathematically.

Appendix A: Exponential bound and technical lemma

We remind of two results which have been proven in [15, Lemma 5.2, Theorem 2.1].

PS 1 (Exponential bound)**.** *If $\theta(.) \equiv \theta$ and Assumption (1) is satisfied then it holds*

$$\mathbb{P}\left(N\,\mathcal{KL}(\overline{\theta},\theta) > z\right) \le 2e^{-z}, \qquad \forall\, z > 0, \tag{A.1}$$

where $N := \sum_{j=1}^n w_j$ and $\overline{\theta} := \sum_{j=1}^n w_j T(Y_j)/N$ with given weights $w_j \in [0,1]$.

PS 2 (Technical Lemma)**.** *Under Assumption (1) it holds*

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

for any sequence $\theta_0,\theta_1,\ldots,\theta_m \in \Theta_\varkappa$, where $\varkappa > 0$ is as in Lemma 2.3.

The proofs of the results in Section 3.3 rely on Equation (A.2). This requires that $\theta_i, T(Y_i) \in \Theta_\varkappa$ for all $i \in \{1,\ldots,n\}$. However, if $\mathbb{P}_\theta$ has unbounded

support, this cannot be satisfied with probability one. We introduce the probability $p_\varkappa$ that quantifies the probability of the event $\{T(Y_i) \notin \Theta_\varkappa \text{ for some } i\}$. Here, we use for every $\varkappa$ the most convenient choice of the set $\Theta_\varkappa$. Additionally, we restrict the range of $\theta(.)$ by the subset $\Theta^* \subseteq \Theta$. This may influence the respective choice of $\Theta_\varkappa$ and as a consequence the corresponding value $p_\varkappa$, which we introduce, now.

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

$$\mathfrak{p}_\varkappa\left(\{\theta_i\}_{i=1}^n\right) := \inf\{\mathbb{P}\left(\exists\, i \in \{1,\dots,n\} : T(Y_i) \notin \Theta_\varkappa\right) : \{\theta_i\}_{i=1}^n \subseteq \Theta_\varkappa \subseteq \Theta\},$$

with $Y_i \sim \mathbb{P}_{\theta_i}$ for all $i \in \{1,\dots,n\}$ and $\varkappa$ sufficiently large such that $\{\theta_i\}_{i=1}^n \subseteq \Theta_\varkappa$ for some set $\Theta_\varkappa \subseteq \Theta$. Furthermore, we denote the worst choice of $\{\theta_i\}_{i=1}^n \subseteq \Theta^*$ by

$$p_\varkappa \quad := \quad \sup\{\, \mathfrak{p}_\varkappa\left(\{\theta_i\}_{i=1}^n\right) :\, \{\theta_i\}_{i=1}^n \subseteq \Theta^*\}.$$

This leads to a trade-off between $\varkappa$ and $p_\varkappa$ allowing the application of the following results to every exponential family in accordance with Assumption (1). The probability $p_\varkappa$ is the smaller the larger we choose $\varkappa > 0$. 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 A.2.

- For Gaussian and log-normal distributed observations, more precisely for $\mathcal{P} = \{\mathcal{N}(\theta,\sigma^2)\}_{\theta\in\Theta}$ and $\mathcal{P} = \{\log\mathcal{N}(\theta,\sigma^2)\}_{\theta\in\Theta}$, it holds $I(\theta) = 1/\sigma^2$ such that $\varkappa = 1$ and $p_\varkappa = 0$. This is the optimal scenario.
- For Gamma, Erlang, scaled chi-squared, Exponential, Rayleigh, Weibull and Pareto distribution, it holds $I(\theta) = 1/\theta^2$. This leads to quite large values of $\varkappa$ and $p_\varkappa$. More precisely, every $\varkappa > 0$ implies that $\Theta_\varkappa = [a, \varkappa a]$ with $a > 0$. For $\mathcal{P} = \{\mathrm{Exp}(1/\theta)\}_{\theta\in\Theta}$ and $Y \sim \mathbb{P}_\theta$ it follows

  $$\mathbb{P}\left(Y \in \Theta_\varkappa\right) = e^{-a/\theta} - e^{-\varkappa a/\theta}$$

  This depends for $\varphi_0 = 0$ and $\Theta^* = \{\theta\}$ on the explicit choices of θ and $\Theta_\varkappa$ via the quotient a/θ, only. Hence, we get by maximization of $\mathbb{P}(Y \in \Theta_\varkappa)$ w.r.t. a/θ for each value of $\varkappa$ the associated probability $p_\varkappa = 1 - \mathbb{P}(Y \in \Theta_\varkappa)^n$, where

$\varkappa$	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

Remark A.3. Alternatively, we could modify slightly the algorithm replacing Equation (2.5) in item (3) of Algorithm 1 by

$$\tilde\theta_i^{(k)} := \operatorname*{argmin}_{\theta'\in\Theta_\varkappa} \left| \theta' - \sum_{j=1}^n \tilde w_{ij}^{(k)} T(Y_j)/\tilde N_i^{(k)} \right|.$$

This projects the adaptive estimator into the set $\Theta_\varkappa$. Analogous, the initial estimates in item (2) can be defined as the projection of the non-adaptive estimator into $\Theta_\varkappa$ such that

$$\tilde{\theta}_i^{(0)} := \operatorname*{argmin}_{\theta' \in \Theta_\varkappa} \left| \theta' - \overline{\theta}_i^{(0)} \right| .$$

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 $U_i^{(0)}$ contains more design points than X_i for each $i \in \{1, \dots, n\}$. Else, the projection may change the adaptive weights in later iteration steps leading to slightly shifted estimators. On the other hand, initialization with $U_i^{(0)} = \{X_i\}$ avoids smoothing among distinct homogeneous regions before adaptation starts.

Appendix B: Proofs

Proof of Theorem 1. Let M^c denote the complement of the set M and consider the event

$$\Omega_\varkappa := \{T(Y_i) \in \Theta_\varkappa \text{ for all } i \in \{1, \dots, n\}\} .$$

The adaptive estimator is defined as weighted mean of the observations. Therefore, we get for all $k \in \{0, \dots, k^*\}$ that

$$\Omega_\varkappa \subseteq \left\{ \tilde{\theta}_i^{(k)} \in \Theta_\varkappa \text{ for all } i \in \{1, \dots, n\} \right\} . \tag{B.1}$$

We observe that

$$\begin{aligned}
\mathbb{P}\left(\mathcal{B}^{(k)}(z)\right) &= 1 - \mathbb{P}\left((\mathcal{B}^{(k)}(z))^c \cap (\mathcal{B}^{(k-1)}(z) \cap \Omega_\varkappa)^c \right) \\
&\quad - \mathbb{P}\left((\mathcal{B}^{(k)}(z))^c \cap (\mathcal{B}^{(k-1)}(z) \cap \Omega_\varkappa) \right) \\
&\geq 1 - \mathbb{P}\left(\Omega_\varkappa^c\right) - \mathbb{P}\left((\mathcal{B}^{(k-1)}(z))^c \cap \Omega_\varkappa \right) \\
&\quad - n \cdot \mathbb{P}\left(\{ \overline{n}_i^{(k)} \mathcal{KL}\left(\tilde{\theta}_i^{(k)}, \theta_i \right) > z \} \cap \mathcal{B}^{(k-1)}(z) \cap \Omega_\varkappa \right)
\end{aligned}$$

and analogous for the conditional probability

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{B}^{(k)}(z) | M^{(k')}(z)\right) \\
\geq\ & 1 - \left[\mathbb{P}\left(M^{(k')}(z) \right) \right]^{-1} \cdot \Big[p_\varkappa + \mathbb{P}\left((\mathcal{B}^{(k-1)}(z))^c \cap \Omega_\varkappa \cap M^{(k')}(z) \right) \\
&\quad + n \cdot \mathbb{P}\left(\{ \overline{n}_i^{(k)} \mathcal{KL}\left(\tilde{\theta}_i^{(k)}, \theta_i \right) > z \} \cap \mathcal{B}^{(k-1)}(z) \cap \Omega_\varkappa \cap M^{(k')}(z) \right) \Big] .
\end{aligned} \tag{B.2}$$

By definition of the events $M^{(k')}(z)$ in Equation (3.5) and $\Omega_\varkappa$ in (B.1) the conditions of Proposition 3.2 are satisfied on $\mathcal{B}^{(k-1)}(z)$. Therefore, it follows on $\mathcal{B}^{(k-1)}(z) \cap \Omega_\varkappa \cap M^{(k')}(z)$ 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 compartment $\mathcal{V}_i$ and $\mathbb{E}\tilde{\theta}_i^{(k)} = \theta_i$. We get with Proposition 3.1

$$\begin{aligned}&\mathbb{P}\left(\{\overline{n}_i^{(k)}\mathcal{KL}\left(\tilde{\theta}_i^{(k)}, \theta_i\right) > z\} \cap \mathcal{B}^{(k-1)}(z) \cap \Omega_{\varkappa} \cap M^{(k')}(z)\right)\\ &\quad \leq \max\left\{2e^{-z}, \epsilon\right\} \end{aligned} \tag{B.3}$$

for all $k \in \{1, \ldots, k'\}$. Now, we proceed by induction. Since $\tilde{\theta}_i^{(0)} = \overline{\theta}_i^{(0)}$ by item (2) of Algorithm 1 it follows from Equation (A.1) in Appendix A that

$$\mathbb{P}\left(\mathcal{B}^{(0)}(z)\right) \overset{\overline{n}_i^{(0)} \leq \overline{N}_i^{(0)}}{\geq} 1 - n \cdot \mathbb{P}\left(\left\{\overline{N}_i^{(0)}\mathcal{KL}(\overline{\theta}_i^{(0)}, \theta_i) > z\right\}\right) \overset{\text{Eq. (A.1)}}{\geq} 1 - 2\,ne^{-z}.$$

Finally, Equations (B.2) and (B.3) lead for all $k \leq k'$ to

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

This terminates the proof. □

Proof of Proposition 3.7. The lower bound holds since

$$\mathbb{P}\left(\mathcal{B}^{(k_2)}(z) | \mathcal{B}^{(k_1)}(z) \cap M^{(k_2)}(z)\right) = 1 - \frac{\mathbb{P}\left((\mathcal{B}^{(k_2)}(z))^c \cap \mathcal{B}^{(k_1)}(z) \cap M^{(k_2)}(z)\right)}{\mathbb{P}\left(\mathcal{B}^{(k_1)}(z) \cap M^{(k_2)}(z)\right)}$$

and furthermore

$$\begin{aligned}&\mathbb{P}\left((\mathcal{B}^{(k_2)}(z))^c \cap \mathcal{B}^{(k_1)}(z) \cap M^{(k_2)}(z)\right)\\ \leq\ & \mathbb{P}\left((\mathcal{B}^{(k_2)}(z))^c \cap \mathcal{B}^{(k_1)}(z) \cap M^{(k_2)}(z) \cap \Omega_{\varkappa}\right) + p_{\varkappa}\\ =\ & \mathbb{P}\left((\mathcal{B}^{(k_2)}(z))^c \cap \mathcal{B}^{(k_2-1)}(z) \cap \mathcal{B}^{(k_1)}(z) \cap M^{(k_2)}(z) \cap \Omega_{\varkappa}\right)\\ &+ \mathbb{P}\left((\mathcal{B}^{(k_2)}(z))^c \cap (\mathcal{B}^{(k_2-1)}(z))^c \cap \mathcal{B}^{(k_1)}(z) \cap M^{(k_2)}(z) \cap \Omega_{\varkappa}\right) + p_{\varkappa}\\ \leq\ & \mathbb{P}\left((\mathcal{B}^{(k_2)}(z))^c \cap \mathcal{B}^{(k_2-1)}(z) \cap M^{(k_2)}(z) \cap \Omega_{\varkappa}\right)\\ &+ \mathbb{P}\left((\mathcal{B}^{(k_2-1)}(z))^c \cap \mathcal{B}^{(k_1)}(z) \cap M^{(k_2)}(z) \cap \Omega_{\varkappa}\right) + p_{\varkappa}\\ \leq\ & \sum_{k=k_1+1}^{k_2} \mathbb{P}\left((\mathcal{B}^{(k)}(z))^c \cap \mathcal{B}^{(k-1)}(z) \cap M^{(k_2)}(z) \cap \Omega_{\varkappa}\right) + p_{\varkappa}.\end{aligned}$$

Additionally, we know from Equation (B.3) that

$$\mathbb{P}\left((\mathcal{B}^{(k)}(z))^c \cap \mathcal{B}^{(k-1)}(z) \cap \Omega_{\varkappa} \cap M^{(k)}(z)\right) \leq \max\left\{2ne^{-z}, n\epsilon\right\}$$

for every $k \leq k'$. Hence, we get from Equation (3.6) that

$$\begin{aligned} & \mathbb{P}\left(\mathcal{B}^{(k_2)}(z) | \mathcal{B}^{(k_1)}(z) \cap M^{(k_2)}(z)\right) \\ \geq \; & 1 - \frac{p_{\varkappa} + (k_2 - k_1)\max\{2ne^{-z}, n\epsilon\}}{\mathbb{P}\left(M^{(k_2)}(z)\right) - p_{\varkappa} - (k_1+1)\max\{2ne^{-z}, n\epsilon\}} \\ = \; & \frac{\mathbb{P}\left(M^{(k_2)}(z)\right) - 2p_{\varkappa} - (k_2+1)\max\{2ne^{-z}, n\epsilon\}}{\mathbb{P}\left(M^{(k_2)}(z)\right) - p_{\varkappa} - (k_1+1)\max\{2ne^{-z}, n\epsilon\}} \end{aligned}$$

leading to the assertion. □

Proof of Proposition 4.1. Substitution with $y := g_\theta^{-1}(z)$ yields $h(g_\theta(y), \theta) = f(y, \theta)$ for $(y, \theta) \in \Omega^f$ and hence 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 $|\frac{\partial g_\theta}{\partial y}| > 0$ to

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

such that

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

The chain rule implies with Equation (4.1) 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 w.r.t. θ. □

Proof of Theorem 2. The non-adaptive estimator is defined as weighted mean of $T(Y_j)$ with $j = 1, \ldots, n$. We get from Table 1 that

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

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 Gaussian or exponential distributions. Therefore, it suffices to consider Gaussian and exponential distribution.

In the Gaussian case, it follows from the statistical independence of the observations $Y_j \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \sigma^2)$, 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 again Gaussian, and the invariance w.r.t. θ follows analogous to Example 4.3, 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 $Y_j \overset{\text{iid}}{\sim} \text{Exp}(1/\theta)$. 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, i.e.

$$\overline{\theta}_i^{(k)} \sim \Gamma\left(\overline{N}_i^{(k)}, \theta / \overline{N}_i^{(k)}\right).$$

This yields the desired invariance w.r.t. θ via Example 4.3 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 $j, j' \in \{1, \dots, n\}$. If $Y_j \sim \text{Exp}(1/\theta)$ then it holds $a_j Y_j \sim \text{Exp}(1/(\theta a_j))$ for all $a_j > 0$, where we denote $a_j := \overline{w}_{ij}^{(k)} / \overline{N}_i^{(k)}$ for the sake of simplicity. 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 $\overline{Y} := a_1 Y_1 + \dots + a_m Y_m$ with $a_j \neq a_{j'}$ for all $j \neq j'$. Including subsequently observations with equal weights $a_j = a_{j'}$ for some $j, j' \in \{1, \dots, n\}$ we conclude by commutativity, associativity and distributivity of the convolution that

$$f_\theta^{\overline{\theta}_i^{(k)}} = \sum_{j=1}^{m} c_j f_j,$$

where the constants $c_j \in \mathbb{R}$ depend again on $a_1, \dots, a_m$ only. The densities f_j follow the distribution $\Gamma(m, \theta a_j)$, where $m \leq m_j$ and m_j denotes the number of observations $Y_{j'}$ with weights $a_{j'} = a_j$. Thus, we get from Example 4.3 the invariance w.r.t. θ for each summand $c_j f_j$ yielding the assertion for weighted sums of exponentials. □

Proof of Lemma 4.6. It holds by Proposition 3.1 that

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

Furthermore, we get

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

Obviously, it holds for any random variable X with values in $[0,1]$ that $\mathrm{Var}[X] \leq \mathbb{E}[X]$. By definition of $M_\lambda^{(k)}(z)$ this yields

$$\begin{aligned}
\max_{i \in \{1,\dots,n_0\}} \mathbb{E}\left[\mathbf{1}_{M_\lambda^{(k)}(z)}(X_i)\right] &= \max_{i \in \{1,\dots,n_0\}} \mathbb{P}\left(\overline{N}_i^{(k)} \mathcal{KL}(\tilde{\theta}_i^{(k)}(\lambda), \theta) > z\right) \\
&\overset{\text{Prop. 3.1}}{\leq} \max\{2e^{-z}, \epsilon\}
\end{aligned}$$

leading to Equation (4.4). □

Acknowledgments

The authors would like to thank Jörg Polzehl, Vladimir Spokoiny and Karsten Tabelow (WIAS Berlin) for helpful discussions.

References

[1] Akkouchi, M. On the convolution of exponential distributions. *J. Chungcheong Math. Soc.*, 21(4):501–510, 2008.

[2] Becker, S. M. A., Tabelow, K., Voss, H. U., Anwander, A., Heidemann, R. M., and Polzehl, J. Position-orientation adaptive smoothing of diffusion weighted magnetic resonance data (POAS). *Med. Image Anal.*, 16(6):1142–1155, 2012. URL http://dx.doi.org/10.1016/j.media.2012.05.007.

[3] Becker, S. M. A., Tabelow, K., Mohammadi, S., Weiskopf, N., and Polzehl, J. Adaptive smoothing of multi-shell diffusion-weighted magnetic resonance data by msPOAS. *Preprint* no. 1809, WIAS, Berlin, 2013.

[4] Belomestny, D. and Spokoiny, V. Spatial aggregation of local likelihood estimates with applications to classification. *Ann. Statist.*, 35(5):2287–2311, 2007. URL http://dx.doi.org/10.1214/009053607000000271. MR2363972

[5] Divine, D. V., Polzehl, J., and Godtliebsen, F. A Propagation-Separation Approach to estimate the autocorrelation in a time-series. *Nonlinear Processes in Geophysics*, 15(4):591–599, 2008.

[6] Lepskiĭ, O. V. A problem of adaptive estimation in Gaussian white noise. *Teor. Veroyatnost. i Primenen.*, 35(3):459–470, 1990. URL http://dx.doi.org/10.1137/1135065. MR1091202

[7] Lepskiĭ, O. V., Mammen, E., and Spokoiny, V. 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. MR1447734

[8] Li, Y., Zhu, H., Shen, D., Lin, W., Gilmore, J. H., and Ibrahim, J. G. Multiscale adaptive regression models for neuroimaging data. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 73(4):559–578, 2011. URL http://dx.doi.org/10.1111/j.1467-9868.2010.00767.x. MR2853730

[9] Li, Y., Gilmor, J.H., Wang, J., Styner, M., Lin, W., and Zhu, H. Twinmarm: two-stage multiscale adaptive regression methods for twin neuroimaging data. *IEEE Trans. Med. Imaging*, 31(5):1100–1112, 2012. URL http://www.ncbi.nlm.nih.gov/pubmed/22287236.

[10] Mathai, A. M. Storage capacity of a dam with gamma type inputs. *Ann. Inst. Statist. Math.*, 34(3):591–597, 1982. URL http://dx.doi.org/10.1007/BF02481056. MR0695077

[11] Mathé, P. and Pereverzev, S. V. Regularization of some linear ill-posed problems with discretized random noisy data. *Math. Comp.*, 75(256):1913–1929 (electronic), 2006. URL http://dx.doi.org/10.1090/S0025-5718-06-01873-4. MR2240642

[12] Moschopoulos, P. G. The distribution of the sum of independent gamma random variables. *Ann. Inst. Statist. Math.*, 37(3):541–544, 1985. URL http://dx.doi.org/10.1007/BF02481123. MR0818052

[13] Polzehl, J. *aws: Adaptive Weights Smoothing*, 2012. URL http://cran.r-project.org/package=aws. R-package version 1.9-1.

[14] Polzehl, J. and Spokoiny, V. Adaptive weights smoothing with applications to image restoration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62:335–354, 2000. MR1749543

[15] Polzehl, J. and Spokoiny, V. Propagation-Separation Approach for local likelihood estimation. *Probability Theory and Related Fields*, 135:335–362, 2006. MR2240690

[16] Polzehl, J. and Spokoiny, V. Structural Adaptive Smoothing by Propagation-Separation Methods. In *Handbook of Data Visualization*, Springer Handbooks Comp. Statistics, pages 471–492, Springer Berlin Heidelberg, 2008.

[17] Polzehl, J., Voss, H.U., and Tabelow, K. Structural adaptive segmentation for statistical parametric mapping. *NeuroImage*, 52:515–523, 2010.

[18] Spokoiny, V. and Vial, C. Parameter tuning in pointwise adaptation using a propagation approach. *Ann. Statist.*, 37(5B):2783–2807, 2009. MR2541447

[19] Tabelow, K., Polzehl, J., Spokoiny, V., and Voss, H. U. Diffusion tensor imaging: structural adaptive smoothing. *Neuroimage*, 39:1763–1773, 2008.