Weierstraß-Institut
für Angewandte Analysis und Stochastik
Leibniz-Institut im Forschungsverbund Berlin e. V.

The Propagation-Separation Approach:
Consequences of model misspecification

Saskia Becker

submitted: November 27, 2013

Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: saskia.becker@wias-berlin.de

ABSTRACT. The article presents new results on the Propagation-Separation Approach by Polzehl and Spokoiny [2006]. This iterative procedure provides a unified approach for nonparametric estimation, supposing a local parametric model. The adaptivity of the estimator ensures sensitivity to structural changes. Originally, an additional memory step was included into the algorithm, where most of the theoretical properties were based on. However, in practice, a simplified version of the algorithm is used, where the memory step is omitted. Hence, we aim to justify this simplified procedure by means of a theoretical study and numerical simulations. In our previous study [Becker and Mathé, 2013], we analyzed the simplified Propagation-Separation Approach, supposing piecewise constant parameter functions with sharp discontinuities. Here, we consider the case of a misspecified model.

1. INTRODUCTION

In statistics, local modeling is one of the most commonly used approaches for nonparametric estimation, see, for instance, Simonoff [1996] and Fan and Gijbels [1996]. Local models can be described by weights which depend on the explanatory variables (design) only. An alternative approach for local modeling is based on weighting schemes that depend (additionally) on the response variables (observations). This helps to avoid blurring at discontinuities. As it turned out, the comparison of noisy observations in single points suffers from a lack of robustness, see Buades et al. [2005] and the references therein. Therefore, the Propagation-Separation Approach by Polzehl and Spokoiny [2006] uses a multiscale approach with iteratively updated weights that benefit from the previously aggregated information about the underlying structure. This enables the detection of discontinuities. Within homogeneous regions the method yields similar results as non-adaptive smoothing.

The Propagation-Separation Approach relates to Lepski's method [Lepskiĭ, 1990; Mathé and Pereverzev, 2006]. Furthermore, it extends the Adaptive Weights Smoothing (AWS) procedure [Polzehl and Spokoiny, 2000], whose theoretical properties were restricted to additive Gaussian noise. In contrast, the Propagation-Separation Approach supposes a local likelihood model. Hence, it is applicable to a large variety of problems. It has been successfully applied in the context of image denoising [Becker et al., 2012; Li et al., 2011, 2012; Polzehl and Spokoiny, 2008; Tabelow et al., 2008], time series analysis [Divine et al., 2008], density estimation, and classification [Polzehl and Spokoiny, 2006], for example. Despite the practical use of this method, only few properties are known. The aim of this article is to provide a better understanding of the Propagation-Separation Approach, the involved parameters, its theoretical properties, and its behavior in practice.

For the verification of theoretical properties, we suppose a local exponential family model. This provides an explicit expression of the Kullback-Leibler divergence on which the algorithm is based. The model includes, for instance, the Gaussian regression and the inhomogeneous Bernoulli, exponential, and Poisson models [Polzehl and Spokoiny, 2006, §2]. In practice, the procedure only requires a metric on the design space and the existence of an appropriate approximation of the Kullback-Leibler divergence.

For the sake of computational simplicity the algorithm is formulated with respect to a local constant model. However, it can be generalized to local linear and local polynomial models as well. In our previous study [Becker and Mathé, 2013], we concentrated on the case of piecewise constant functions with sharp discontinuities. Here, we will analyze consequences of a misspecified model. As in [Becker and Mathé, 2013], we omit the additional memory step and avoid Assumption (S0) in [Polzehl and Spokoiny, 2006]. Assumption (S0) is problematic since it requires the data-driven weights of the estimator to be statistically independent of the observations. The memory step was included into the algorithm in order to ensure a certain stability of estimates. However, in applications of the Propagation-Separation Approach it has been omitted. As it turned out, its practical use is questionable, while the algorithm provides the desired behavior even without the memory step.

The outline is as follows. First, we will recall the local exponential family model and the original algorithm of the Propagation-Separation Approach. Then, we will introduce a parameter choice strategy for the adaptation bandwidth that allows the verification of propagation and a certain stability of estimates for functions with bounded variability within well-separated regions. Moreover, we will define an *associated step function* which approximates the estimation function for sufficiently large location bandwidths. In Section 4, we will provide further details concerning the practical application of the newly introduced inhomogeneous propagation condition. Our subsequent numerical simulations illustrate the formation of the associated step function. All examples suggest the convergence of the Propagation-Separation Approach. However, this property could not be proven theoretically for reasons that we will discuss in Section 6. In Appendix A, we will recall some auxiliary results by Polzehl and Spokoiny [2006] and Becker and Mathé [2013]. Longer proofs will be given in Appendix B.

2. MODEL AND METHODOLOGY

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

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

For the sake of simplicity, we assume the design to be known and the observation space as well as the parameter set to be one-dimensional, that is $\mathcal{Y}, \Theta \subseteq \mathbb{R}$. Basically, the Propagation-Separation Approach can be applied on any measurable vector space $\mathcal{Y} \subseteq M$ with $Y_i \sim \mathbb{P}_{\theta(X_i)}$ for every $i \in \{1, ..., n\}$ and $\theta : \mathcal{X} \to \Theta \subseteq M$, where M is endowed with a possibly asymmetric distance function.

The algorithm is based on the Kullback-Leibler divergence. This has an explicit expression under the following assumption, which was supposed in [Becker and Mathé, 2013; Polzehl and Spokoiny, 2006], as well. A list of parametric families that satisfy this assumption is given in [Becker and Mathé, 2013, Table 1]. We use the common notation

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

Assumption 1 (Local exponential family model). The parametric family $\mathcal{P} = \{\mathbb{P}_\theta\}_{\theta \in \Theta}$ in Notation 2.1 is an exponential family. More precisely, there are two functions $C, B \in C^2(\Theta, \mathbb{R})$, a non-negative function $p : \mathcal{Y} \to [0, \infty)$, and a sufficient statistic $T : \mathcal{Y} \to \mathbb{R}$ such that

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

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

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

We recall the notions of the Fisher information

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

and of the Kullback–Leibler divergence

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

The Propagation-Separation Approach estimates iteratively the unknown parameter function $\theta(.)$. Here, we consider its simplest version, supposing the parameter function $\theta(.)$ to be piecewise constant with sharp discontinuities. The pointwise estimator equals a weighted mean of the observations. In each iteration step k the adaptive weights are readjusted using the previously aggregated information. More precisely, the adaptive weights are defined as a product of two kernels. The *location kernel* describes for every point $X_i \in \mathcal{X}$ the increasing neighborhood $U_i^{(k)} \subseteq \mathcal{X}$ under consideration, leading to an advancing variance reduction. The *adaptation kernel* uses the Kullback-Leibler divergence for a comparison of the pointwise parameter estimates from the previous iteration step. This avoids blurring at structural borders. An additional *memory step* ensures a certain stability of estimates. In each iteration step, the *memory penalty* compares, for every design point, the new estimate with the previous one. In case of a significant difference, the new estimate is relaxed, replacing it by a value between the two estimates. The memory step provides a smooth transition of the pointwise estimates during iteration. We emphasize that the Propagation-Separation Approach does not use adaptive parameters. It is adaptive in the sense that the returned estimator function is based on structure-adaptive weights that describe the homogeneity regions of the unknown parameter function $\theta(.)$. See Algorithm 1, stated below, for a formal description and Polzehl and Spokoiny [2006] and Becker and Mathé [2013] for more details.

Notation 2.2. Suppose Assumption 1. We fix three non-increasing kernel functions

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

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

$$(2.2) \qquad \overline{\theta}_i^{(k)} := \sum_{j=1}^{n} \overline{w}_{ij}^{(k)} T(Y_j) / \overline{N}_i^{(k)},$$

the non-adaptive estimator of θ_i, where $\overline{w}_{ij}^{(k)} := K_{\mathrm{loc}} \left(\delta(X_i, X_j) / h^{(k)} \right)$ and $\overline{N}_i^{(k)} := \sum_j \overline{w}_{ij}^{(k)}$.

Algorithm 1 (The original Propagation-Separation Algorithm)**.**

  1 Input parameters: Sequence of bandwidths $\{h^{(k)}\}_{k=0}^{k^*}$, adaptation bandwidth λ,
     the memory bandwidth τ, and the minimal memory effect η_0.

  2 Initialization: $\hat{\theta}_i^{(0)} := \overline{\theta}_i^{(0)}$ and $\hat{N}_i^{(0)} := \overline{N}_i^{(0)}$ for all $i \in \{1, ..., n\}$, $k := 1$.

  3 Iteration: Calculate, for every $i, j = 1, ..., n$,
     the non-adaptive weights $\overline{w}_{ij}^{(k)} := K_{\mathrm{loc}} \left(\delta(X_i, X_j) / h^{(k)} \right)$,
     the statistical penalty $s_{ij}^{(k)} := \hat{N}_i^{(k-1)} \mathcal{KL}(\hat{\theta}_i^{(k-1)}, \hat{\theta}_j^{(k-1)})$,
     the adaptive weights $\tilde{w}_{ij}^{(k)} := \overline{w}_{ij}^{(k)} \cdot K_{\mathrm{ad}} \left(s_{ij}^{(k)} / \lambda \right)$,
     the sum of the adaptive weights $\tilde{N}_i^{(k)} := \sum_j \tilde{w}_{ij}^{(k)}$,
     and the adaptive estimator

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

  4 Memory step: Calculate, for every $i, j = 1, ..., n$,
     the sum of the non-adaptive weights $\overline{N}_i^{(k)} := \sum_j \overline{w}_{ij}^{(k)}$,
     the memory penalty $m_i^{(k)} := \overline{N}_i^{(k)} \mathcal{KL}(\tilde{\theta}_i^{(k)}, \hat{\theta}_i^{(k-1)})$,

the relaxation weight $\eta_i^{(k)} := (1 - \eta_0) K_{\mathrm{me}}\left(m_i^{(k)}/\tau\right)$,

the relaxed estimator

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

and the relaxed sum of the adaptive weights $\hat{N}_i^{(k)} := \eta_i^{(k)} \tilde{N}_i^{(k)} + (1 - \eta_i^{(k)})\hat{N}_i^{(k-1)}$.

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

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

Notation 2.3 (Simplified algorithm)**.** In the rest of this article, we refer to Algorithm 1 as the *original algorithm* with aggregated estimates $\{\hat{\theta}_i^{(k)}\}_{i,k}$. The formal choice $\eta_i^{(k)} := 1$ for every $i \in \{1, ..., n\}$ and each $k \in \{1, ..., k^*\}$ omits the memory step, leading to the *simplified algorithm* with adaptive estimates $\{\tilde{\theta}_i^{(k)}\}_{i,k}$.

The choice of the input parameters is crucial for the behavior of the algorithm. Since the initial estimator $\tilde{\theta}_i^{(0)}$ is non-adaptive the corresponding location bandwidth $h^{(0)}$ should be small. A choice of $h^{(0)}$ such that $\overline{w}_{ij}^{(0)} = 0$ for all $X_i \neq X_j$ avoids blurring at the boundaries of the homogeneity regions. The subsequent bandwidths $\{h^{(k)}\}_{k=1}^{k^*}$ should be increasing. For instance, they may ensure a constant variance reduction of the estimator [Becker et al., 2012] or an exponential growth [Polzehl and Spokoiny, 2006] of the mean number of design points $X_j \in \mathcal{X}$ with non-zero weights $\overline{w}_{ij}^{(k)} \neq 0$, $X_i \in \mathcal{X}$. The maximal number of iterations k^* specifies the maximal location bandwidth $h^{(k^*)}$. This is mainly bounded by the available computation time. However, in the case of model misspecification the resulting estimation bias can be reduced by an accurate stopping criterion as we will discuss in § 6.3.

The adaptation bandwidth λ specifies the amount of adaptation. For $\lambda \to \infty$ the algorithm results in non-adaptive estimates as defined in Equation (2.2) (over-smoothing), while small values lead to adaptation to noise (under-smoothing). Polzehl and Spokoiny [2006, §3.5] introduced a choice of λ by a strategy, called the *propagation condition*. This recommends to use the smallest value for λ that provides under homogeneity a similar behavior as non-adaptive smoothing. We use a revised formulation that was introduced in [Becker and Mathé, 2013, §2.3]. It is based on the function $\mathfrak{Z}_\lambda : \{0, ..., k^*\} \times (0, 1) \times \Theta \times \{1, ..., n\} \to [0, \infty)$ given by

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

where $\lambda > 0$ is fixed. Here, $\tilde{\theta}_i^{(k)}(\lambda)$ denotes the adaptive estimator in the position $X_i \in \mathcal{X}$, resulting from the simplified algorithm in Notation 2.3 with the adaptation bandwidth λ and observations $Y_j \overset{\mathrm{iid}}{\sim} \mathbb{P}_\theta$ for all $j \in \{1, ..., n\}$ with $\theta(.) \equiv \theta$.

Definition 2.4 (Homogeneous propagation condition)**.** We say that the adaptation bandwidth $\lambda > 0$ is chosen in accordance with the homogeneous propagation condition at level $\epsilon > 0$ for $\theta \in \Theta$ if the function $\mathfrak{Z}_\lambda(., p; \theta, i)$ is non-increasing for all $p \in (\epsilon, 1)$ and every $i \in \{1, ..., n\}$.

The study in [Becker and Mathé, 2013, §4.1] points out that the choice of λ by the homogeneous propagation condition is invariant with respect to the underlying parameter θ for the Gaussian and the exponential distribution and, as a consequence, for the log-normal, Rayleigh, Weibull, and Pareto distributions. Else,

some parameter θ^* which yields a sufficiently large choice should be identified, such that the homogeneous propagation condition holds for all (unknown) parameters θ_i with $i \in \{1, ..., n\}$ as well. Hence, the homogeneous propagation condition allows a choice of λ by simulations and hence independent of the data at hand. We refer the reader to Polzehl and Spokoiny [2006, §3.4 & 3.5] and Becker and Mathé [2013, §2.3 & 4.1] for more details.

3. THEORETICAL PROPERTIES UNDER MODEL MISSPECIFICATION

We consider the local exponential family model in Assumption 1 (page 2) and the simplified procedure in Notation 2.3. In [Becker and Mathé, 2013, §2.4], its general behavior was illustrated on two examples. There, we observed the following. For piecewise constant parameter functions $\theta(.)$, the algorithm detects sufficiently sharp discontinuities providing a consistent estimation function. For small discontinuities this separation property fails. This leads to a bounded estimation bias since the algorithm treats non-separated homogeneity regions as one yielding similar results as non-adaptive smoothing. For piecewise smooth parameter functions $\theta(.)$ the algorithm results in a step function which is mainly determined by the local smoothness of the parameter function $\theta(.)$ and the adaptation bandwidth λ. An appropriate stopping criterion may reduce the corresponding estimation bias.

Our theoretical study in [Becker and Mathé, 2013] focused on piecewise constant functions with sharp discontinuities. In this article, we aim to verify the mentioned heuristic observations for the case of model misspecification. First, we will generalize the propagation condition for the choice of the adaptation bandwidth to inhomogeneous settings with bounded variability. Then, the propagation and the stability property will follow for parameter functions with (piecewise) bounded variability in a similar manner as under (piecewise) homogeneity, see Becker and Mathé [2013, §3.1 & 3.3]. Furthermore, we will introduce a specific step function which approximates the adaptive estimates, resulting from the simplified Propagation-Separation Approach.

3.1. **Inhomogeneous propagation condition.** The homogeneous propagation condition in Definition 2.4 bounds the probability of adaptation to noise, supposing a constant parameter function. In [Polzehl and Spokoiny, 2006, §3], this was used to verify propagation and a certain stability of estimates for (piecewise) constant parameter functions. In order to extend these properties to (piecewise) bounded parameter functions, we will formulate an inhomogeneous propagation condition. Like before under homogeneity, we will consider an artificial data set. Then, we aim to ensure a similar behavior of the algorithm as for non-adaptive estimation for every locally varying function which satisfies a pre-specified variability bound.

Our inhomogeneous propagation condition is motivated by Theorem A.12 in Appendix A. This can be considered as the inhomogeneous analog of Theorem A.8, where the homogeneous propagation condition was based on. For the non-adaptive estimator, Theorem A.12 establishes the exponential bound

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

for all $z > 0$, $i \in \{1, ..., n\}$, and $k \in \{0, ..., k^*\}$, where we refer the reader to the Notations A.2 and A.9 for the definitions of $\varkappa \geq 1$ and $\breve{p}_\varkappa \in [0, 1]$. This result implies that the Kullback-Leibler divergence $\mathcal{KL}(\overline{\theta}_i^{(k)}, \mathbb{E}\overline{\theta}_i^{(k)})$ decreases, in probability, at least with rate $\overline{N}_i^{(k)}$. We observe that

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

whereas

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

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

Notation 3.1. We set

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

Next we specify the considered inhomogeneous setting. Following Polzehl and Spokoiny [2006, §5.2], we presume that the variability of the parameter function $\theta(.)$ is smaller in order than the rate of convergence $\overline{N}_i^{(k)}$ in Theorem A.12. Here, we even require the rate $\max_{j'}\overline{N}_{j'}^{(k)}$ in order to ensure that $\overline{N}_i^{(k)}/\max_{j'}\overline{N}_{j'}^{(k)} \leq 1$ for every $i \in \{1, ..., n\}$. More precisely, we require the existence of a constant $\varphi \geq 0$ such that

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

for every $i \in \{1, ..., n\}$ and each $k \in \{0, ..., k^*\}$. In this subsection, we require Equation (3.1) with $k := k^*$ for all $i, j \in \{1, ..., n\}$, but, in § 3.2, we will only consider the points in a certain neighborhood, for instance all $X_j \in U_i^{(k)}$ with $k \in \{0, ..., k^*\}$. For brevity, we denote $\varphi_0 := \varphi/\max_i(\overline{N}_i^{(k^*)})^{1/2}$.

We proceed as under homogeneity, see Definition 2.4 for comparison.

Notation 3.2. For every $\lambda > 0$, we consider the function

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

defined by

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

where $\mathcal{E}\tilde{\theta}_i^{(k)}$ is as in Notation 3.1, and $\tilde{\theta}_i^{(k)}(\lambda)$ denotes the adaptive estimator in the position $X_i \in \mathcal{X}$, resulting from the simplified algorithm in Notation 2.3 with the adaptation bandwidth λ and the statistically independent observations $Y_j \sim \mathbb{P}_{\theta_j} \in \mathcal{P}, j \in \{1, ..., n\}$.

In order to enable the application of Equation (A.1) and Lemma A.3, we restrict the range of the parameter function $\theta(.)$. Thus, we introduce a subset $\Theta^* \subseteq \Theta$ with $\{\theta_j\}_{j=1}^n \in (\Theta^*)^n$.

Definition 3.3 (Inhomogeneous propagation condition). Let $\epsilon > 0$ and $\varphi_0 \geq 0$ be constants. The adaptation bandwidth $\lambda > 0$ satisfies the inhomogeneous propagation condition at probability level ϵ and variability level φ_0 for the parameter set $\Theta^* \subseteq \Theta$ if the function $\hat{\mathfrak{z}}_\lambda(., p; \{\theta_j\}_{j=1}^n, i)$ is non-increasing for all $p \in (\epsilon, 1)$, every $i \in \{1, ..., n\}$, and each parameter function $\theta(.)$ with $\{\theta_j\}_{j=1}^n \in (\Theta^*)^n$ and $\mathcal{KL}(\theta_j, \theta_{j'}) \leq \varphi_0^2$ for all $j, j' \in \{1, ..., n\}$.

Remark 3.4. For $\varphi_0 := 0$ and $\Theta^* := \{\theta\}$, the inhomogeneous propagation condition equals the homogeneous propagation condition in Definition 2.4.

3.2. **Locally varying parameter functions with sharp discontinuities.** We deduce from the inhomogeneous propagation condition and Theorem A.12 in Appendix A an exponential bound for the probability $\mathbb{P}(\overline{N}_i^{(k)}\mathcal{KL}(\tilde{\theta}_i^{(k)}, \theta_i) > z)$ of the Kullback-Leibler divergence between the adaptive estimator $\tilde{\theta}_i^{(k)}$ and its true parameter θ_i to exceed the error bound $z/\overline{N}_i^{(k)}$. The following proposition provides the inhomogeneous analog of [Becker and Mathé, 2013, Proposition 3.1]. It requires slightly different assumptions and yields a different exponent in the exponential bound of the excess probability.

Proposition 3.5 (Propagation and stability under bounded variability)**.** *Let Assumption 1 (page 2) be fulfilled, and let the adaptation bandwidth λ be chosen in accordance with the inhomogeneous propagation condition at probability level $\epsilon > 0$ and variability level $\varphi_0 > 0$ for some set $\Theta^* \subseteq \Theta$ satisfying $\{\theta_i\}_{i=1}^n \in (\Theta^*)^n$. Additionally, we recall Notation A.2 in Appendix A, and we choose $\varkappa \geq 1$ sufficiently large such that $\Theta^* \subseteq \Theta_\varkappa$. If $\mathcal{KL}(\theta_i, \theta_j) \leq \varphi^2 / \max_{j'} \overline{N}_{j'}^{(k_0)} = \varphi_0^2$ holds for all $i, j \in \{1, ..., n\}$ and some fixed $k_0 \in \{1, ..., k^*\}$, then we get*

$$(3.2) \qquad \mathbb{P}\left(\overline{N}_i^{(k)} \mathcal{KL}\left(\tilde{\theta}_i^{(k)}, \theta_i\right) > z\right) \leq \max\left\{2e^{-\left[\sqrt{z}/\varkappa - \varphi\right]^2/\varkappa^2}, \epsilon\right\} + \breve{p}_{\varkappa,0}$$

for each $i \in \{1, ..., n\}$, $k \in \{0, ..., k_0\}$, and all $z > \varkappa^2 \varphi^2$, where $\breve{p}_{\varkappa,0}$ is as in Notation A.9. In particular, for all $k_1 \leq k_2 \leq k_0$, it holds

$$(3.3) \qquad \mathbb{P}\left(\overline{N}_i^{(k_2)} \mathcal{KL}\left(\tilde{\theta}_i^{(k_2)}, \mathcal{E}\tilde{\theta}_i^{(k_2)}\right) > z\right) \leq \max\left\{\mathbb{P}\left(\overline{N}_i^{(k_1)} \mathcal{KL}\left(\tilde{\theta}_i^{(k_1)}, \mathcal{E}\tilde{\theta}_i^{(k_1)}\right) > z\right), \epsilon\right\},$$

where $\mathcal{E}\tilde{\theta}_i^{(k)} = \sum_j \tilde{w}_{ij}^{(k)} \theta_j / \tilde{N}_i^{(k)}$ as in Notation 3.1.

Next we consider piecewise bounded functions with sharp discontinuities. We recall some auxiliary notations.

Notation 3.6. For any set M, we define

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

Then, we call the discrete set $M := \{X_j\}_{j=1}^m \subseteq \mathcal{X}$ convex if $\mathfrak{C}(M)$ is convex and

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

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

Assumption 2. Suppose the existence of a non-trivial partition $\mathcal{V} := \{\mathcal{V}_i\}_i$ of $\mathcal{X}$ such that, for every $X_i \in \mathcal{X}$, there are constants $\phi_i > \varphi_0 \geq 0$ and a convex neighborhood $\mathcal{V}_i \subseteq \mathcal{X}$ which satisfy

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

We recall some notations from [Becker and Mathé, 2013]. The effective sample size concentrates on the case where the considered neighborhood $U_i^{(k)} = \{X_j \in \mathcal{X} : \overline{w}_{ij}^{(k)} > 0\}$ is larger than the corresponding region region $\mathcal{V}_i$.

Notation 3.7. We define, for each $i \in \{1, ..., n\}$ and every $k \in \{0, ..., k^*\}$, the effective sample size and its local minimum

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

As it turns out, the quantities $n_i^{(k)}$ determine a lower bound for the stepsizes ϕ_i which allows the detection of the associated discontinuity by the algorithm. In the following theorem, we consider two events. On the first one, $\mathcal{B}^{(k)}(z)$, the estimation error is bounded from above, and on the second one, $M^{(k)}(z)$, the discontinuities are sufficiently sharp for separation, see Proposition A.13 in Appendix A.

Notation 3.8. Let the constants $\phi_i > 0$, $i \in \{1, ..., n\}$, be as in Assumption 2, and fix $\lambda > 0$ and $z > 0$. Additionally, we recall Notation A.2 and choose $\varkappa \geq 1$ such that $\{\theta_i\}_{i=1}^n \in (\Theta_\varkappa)^n$. Then, we set

$$(3.5) \qquad \mathcal{B}^{(k)}(z) := \bigcap_{i=1}^n \left\{\overline{n}_i^{(k)} \mathcal{KL}(\tilde{\theta}_i^{(k)}, \theta_i) \leq z\right\}, \qquad k \in \{0, ..., k^*\},$$

$M^{(0)}(z) := \Omega$, and

$$(3.6) \qquad M^{(k)}(z) := \bigcap_{k'=0}^{k-1} \bigcap_{i=1}^{n} \left\{ \phi_i > \varkappa \left[\sqrt{\lambda/\tilde{N}_i^{(k')}} + 2\sqrt{z/n_i^{(k')}} \right] \right\}, \quad k \in \{1, ..., k^*\}.$$

Theorem 3.9 (Propagation property under piecewise boundedness)**.** *Suppose Assumptions 1 (page 2) and 2 to be satisfied. Additionally, let the adaptation bandwidth λ, the constant $\varkappa \geq 1$, and the corresponding set $\Theta_\varkappa \subseteq \Theta$ be as in Proposition 3.5 and $h^{(0)} > 0$ sufficiently small such that $\overline{w}_{ij}^{(k)} = 0$ for all $X_i \neq X_j$. Finally, we fix some iteration step $k_0 \in \{0, ..., k^*\}$ and some constant $\varphi \geq 0$ such that $\varphi^2/\max_i \overline{n}_i^{(k_0)} = \varphi_0^2$. If $z > \varkappa^2 \varphi^2$ satisfies $\mathbb{P}\left(M^{(k_0)}(z)\right) > 0$, then it holds*

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

where $\breve{p}_{\varkappa,0}$ is as in Notation A.9.

Remark 3.10. Theorem 3.9 yields a meaningful result for $z \geq \varkappa^2[\varkappa\sqrt{q\log(n)} + \varphi]^2$ and small values of ϵ or, at best, $\epsilon := c_\epsilon n^{-q}$ with $q > 1$ and $c_\epsilon > 0$.

3.3. **Consequences of a violated structural assumption.** The previous results only hold for parameter functions with sharp discontinuities. What happens in the case of a violated structural assumption? In [Becker and Mathé, 2013, §2.4], we observed for simulated examples with Gaussian distributed observations that the estimation function resulted in a step function in the case of a piecewise constant parameter function and as well for a piecewise smooth function. Therefore, we will introduce a specific step function, that we will call the *associated step function* of the Propagation-Separation Approach. Then, we will establish an upper bound for the pointwise Kullback-Leibler divergence between the adaptive estimator of the simplified procedure in Notation 2.3 and the corresponding value of the associated step function.

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

$$(3.7) \qquad \mathcal{H}_i^{(k)} := \left\{ \mathcal{X}_j \in \mathcal{X} : \tilde{w}_{il}^{(k+1)} > 0 \text{ if and only if } \tilde{w}_{jl}^{(k+1)} > 0 \text{ for all } X_l \in \mathcal{X} \right\},$$

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

Definition 3.11. Let $\mathbf{1}$ denote the indicator function, and let $\theta_l^{(k)}$ be the mean value of the n_l estimates $\tilde{\theta}_{l_j}^{(k)}$ corresponding to the design points $\{X_{l_j}\}_{j=1}^{n_l}$ which form the region $H_l^{(k)}$. Then, we call the piecewise constant function

$$(3.8) \qquad \breve{\theta}^{(k)}(X_i) := \sum_{l=1}^{m} \theta_l^{(k)} \mathbf{1}_{H_l^{(k)}}(X_i) \quad \text{with} \quad \theta_l^{(k)} := \frac{1}{n_l} \sum_{j=1}^{n_l} \tilde{\theta}_{l_j}^{(k)}$$

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

The associated step function satisfies the following property.

Lemma 3.12. *For all $i \in \{1, ..., n\}$ and $k \in \{0, ..., k^*\}$, it holds*

$$(3.9) \qquad \mathcal{KL}\left(\breve{\theta}_i^{(k)}, \tilde{\theta}_i^{(k)}\right) \leq \max\{\lambda/\tilde{N}_j^{(k)} : X_j \in \mathcal{H}_i^{(k)}\}.$$

Proof. We know from Lemma A.1 that the Kullback-Leibler divergence is convex with respect to the first argument. Therefore, it holds

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

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

In § 5.2, we will illustrate the formation of the associated step function during iteration. The corresponding simulations suggest its immutability for sufficiently large bandwidths. Additionally, we will see that, in the presented examples, the sets $\{\mathcal{X}_j \in \mathcal{X} : \tilde{w}_{ij}^{(k)} > 0\}$ with $i \in \{1, ..., n\}$ form a well-defined partition of the design space $\mathcal{X}$ if k is sufficiently large. However, both heuristic observations could not be theoretically justified for reasons that we will discuss in § 6.1.

4. The inhomogeneous propagation condition in practice

In this section, we will present further details concerning the practical application of the inhomogeneous propagation condition. In [Becker and Mathé, 2013, §4.2], we explained how the homogeneous propagation condition can be applied in practice. In contrast, the inhomogeneous propagation condition cannot be applied directly if $\varphi_0 > 0$. Here, we need to ensure that the criterion is fulfilled for *every* parameter function satisfying $\{\theta_i\}_{i=1}^n \in (\Theta^*)^n$ and $\mathcal{KL}(\theta_i, \theta_j) \leq \varphi_0^2$ for all $i, j \in \{1, ..., n\}$. Therefore, we recommend to choose some $\lambda > 0$ in accordance with the homogeneous propagation condition and to increase it such that the inhomogeneous propagation condition holds as well. Apart from the Gaussian and log-normal distribution, the practical use of our precise choice is questionable due to the size of the involved constants. Nevertheless, it suggests the *existence* of an appropriate value. Hence, the inhomogeneous propagation condition is in the first instance of theoretical interest. It allows the desired extension of the propagation and the stability property to (piecewise) bounded functions. The justification of our choice will be based on a comparison of the homogeneous and the inhomogeneous propagation condition. In order to avoid confusion, we introduce the following notation.

Notation 4.1. Let the parametric family $\mathcal{P}$ satisfy Assumption 1 (page 2) with a strictly monotonic sufficient statistic T. We fix some constant $\varphi_0 > 0$ and a subset $\Theta^* \subseteq \Theta$. Then, we consider two data sets $\{(X_i, Y_i)\}_{i=1}^n$ and $\{(X_i, \mathcal{Y}_i)\}_{i=1}^n$, where

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

In the rest of this section, we will write $\mathcal{Y}$ and ϑ whenever we restrict to the special case of a homogeneous setting. Else, we will write Y and θ, explicitly allowing locally varying parameter functions which satisfy the variability bound in Notation 4.1. Now we look for a description of the homogeneous propagation condition which enables an extension to the inhomogeneous setting. For this purpose, we introduce some auxiliary functions.

Notation 4.2. Let the functions $p_\theta^{(l)} : (0, \infty) \to [0, 1]$, $l = 1, 2, 3$ and $\theta \in \Theta$, be given as

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

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

Lemma 4.3. *The functions $p_\theta^{(l)}$, $l = 1, 2, 3$, in Notation 4.2 are invariant with respect to the parameter $\theta \in \Theta$ for the Gaussian, log-normal, Gamma, Erlang, scaled chi-squared, exponential, Rayleigh, Weibull, and Pareto distributions.*

The study in [Becker and Mathé, 2013, §4.1] suggests the invariance of the homogeneous propagation condition with respect to the parameter θ for the Gaussian, log-normal, exponential, Rayleigh, Weibull, and Pareto distribution. In the following lemma, we take advantage of this invariance. There, we completely determine the corresponding function $\mathfrak{Z}_\lambda$ via the distribution of the positions of the observations around the respective parameter $\vartheta \in \Theta$ given by the functions $p_\vartheta^{(l)}$, $l = 1, 2, 3$.

Lemma 4.4. *Assume the setting of Notation 4.1. If the homogeneous propagation condition is invariant with respect to the parameter $\vartheta \in \Theta$, then the corresponding function $\mathfrak{Z}_\lambda$ is uniquely determined by the functions $p_\vartheta^{(l)}$, $l = 1, 2, 3$, for every $\vartheta \in \Theta$.*

In other words, the homogeneous propagation condition is determined by the probability distributions of $\mathcal{KL}(T(\mathcal{Y}), \vartheta)$ on $\{T(\mathcal{Y}) > \vartheta\}$ and on $\{T(\mathcal{Y}) \leq \vartheta\}$. Under inhomogeneity, we have to additionally compensate for the local variability of the parameter function. We investigate the interplay of the observations via the distribution of $\mathcal{KL}(Y_i, Y_j)$, which we compare with its homogeneous counterpart $\mathcal{KL}(\mathcal{Y}_i, \mathcal{Y}_j)$. For simplicity, we presume the sufficient statistic T in Assumption 1 to equal the identity. Instead of that, we could replace in the following all observations Y_i and $\mathcal{Y}_i$ by the transformed observations $T(Y_i)$ and $T(\mathcal{Y}_i)$, leading, for every strictly monotonic T, to the same results but more tedious terms. We restrict to the favorable realizations, where the corresponding event M_0 is related to the event $\Omega_{\varkappa}$ in Corollary A.11.

Proposition 4.5. *Suppose the setting of Notation 4.1, $T = \mathrm{Id}$, and let the functions $p_\theta^{(l)}$, $l = 1, 2, 3$, be invariant with respect to the parameter $\theta \in \Theta^*$. Additionally, recall Notation A.2, and let $\varkappa \geq 1$ satisfy $\{\vartheta\} \cup \{\theta_i\}_{i=1}^n \in (\Theta_{\varkappa})^{n+1}$, where $\Theta_{\varkappa} \subseteq \Theta$ maximizes the probability of the event*

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

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

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

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

Claim 4.6. *Let Assumption 1 be satisfied with $T : \mathcal{Y} \to \mathbb{R}$ strictly monotonic, and fix a subset $\Theta^* \subseteq \Theta$ and some constant $\varphi_0 := \varphi / \max_i (\overline{N}_i^{(k^*)})^{1/2}$ with $\varphi > 0$. Additionally, let the homogeneous propagation condition and the functions $p_\theta^{(l)}$, $l = 1, 2, 3$, be invariant with respect to the parameter $\theta \in \Theta$. Finally, we presume the adaptation bandwidth $\lambda > 0$ to be in accordance with the homogeneous propagation condition at level $\epsilon > 0$. Then, the choice*

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

is in accordance with the inhomogeneous propagation condition at probability level $\epsilon(\lambda_\varphi) \leq \epsilon + 2p_{\varkappa}$ and variability level φ_0 for the parameter set Θ^*, where $p_{\varkappa}$ is as in Notation A.4.

Admittedly, the iterative approach of the algorithm impedes a definite proof. Instead, we will present a justification of Claim 4.6 in Appendix B, where we will follow an inductive argumentation in order to overcome the remaining gap at least to a certain extent.

Remark 4.7. Recall that the parameter set $\Theta^* \subseteq \Theta$ and the variability level φ_0 influence the sizes of the constant $\varkappa \geq 1$ and the corresponding probability $p_\varkappa$. Moreover, we point out that the probability level $\epsilon + 2p_\varkappa$ is an upper bound for the actual probability level $\epsilon(\lambda_\varphi)$, but this bound does not have to be sharp. Similarly, the proposed choice of λ_φ is based on a rough estimation, where the effectively required values of $\varkappa$ and φ may be much smaller than supposed. Hence, in practice, one can always use a smaller bandwidth $\lambda^* < \lambda_\varphi$ with an unknown probability level if this seems to be advantageous. This may increase the risk of adaptation to noise as usually $\epsilon(\lambda^*) \geq \epsilon(\lambda_\varphi)$, but the main property, propagation with probability $1 - \epsilon(\lambda^*)$, remains valid. In any case, one should use the homogeneous adaptation bandwidth λ as a lower bound, $\lambda \leq \lambda^*$. Recalling Example A.5 concerning the trade-off between $\varkappa$ and $p_\varkappa$, we conclude the following.

- Claim 4.6 provides a reasonable choice of the adaptation bandwidth if $\varkappa = 1$ and $p_\varkappa = 0$, such as for Gaussian and log-normal distributed observations.
- For Gamma, Erlang, scaled chi-squared, exponential, Rayleigh, Weibull, and Pareto distribution, $\varkappa$ and $p_\varkappa$ are large. For these distributions, Claim 4.6 justifies the *existence* of an adaptation bandwidth λ_φ which is in accordance with the inhomogeneous propagation condition at level $\epsilon(\lambda_\varphi) \leq \epsilon + 2p_\varkappa$, but its practical use is questionable due to the sizes of $\varkappa$ and $p_\varkappa$.

5. SIMULATIONS

In [Becker and Mathé, 2013] and Section 3, we established several theoretical properties of the simplified Propagation-Separation Approach in Notation 2.3. Here, we will illustrate these properties by simulated examples with Gaussian and exponentially distributed observations. In particular, we will compare the results of the simplified algorithm with the original procedure in Algorithm 1. We will present several example plots for the realization `seed=1`, the corresponding weighting schemes, and boxplots of the mean absolute error (MAE) over 1000 realizations, where `seed=l` and $l \in \{1, ..., 1000\}$.

5.1. **Test functions and methods.** Here, we present all test functions that we will consider in the following numerical study. Usually, we simulated data with $n = 1000$ observations. In some examples, we changed the sample size $n \in \mathbb{N}$. Then, we increased the cardinality of each region of the introduced parameter functions by the same factor such that the design portions remained unchanged.

We used the implementation of the Propagation-Separation Approach in the R-package **aws** by Polzehl [2012]. Here, the memory step is omitted by default. If desired, it can be included in the procedure setting `memory=TRUE`. The memory step is implemented for two different memory kernels, which can be specified by `aggkern="Triangle"` or `aggkern="Uniform"`. If not mentioned differently, we applied the default parameters using the command

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

where `tnoise` denotes the simulated observations and `t` is the corresponding expectation. For the sake of simplicity, we show only univariate examples where $\mathcal{X} \subseteq \mathbb{R}$.

By means of an additionally included function `awsweights`, we visualized the weighting schemes of the resulting non-adaptive weights $\{\overline{w}_{ij}^{(k)}\}_{i,j}$, the adaptive weights $\{\tilde{w}_{ij}^{(k)}\}_{i,j}$, and the adaptation kernel $\{K_{\mathrm{ad}}(s_{ij}^{(k)}/\lambda)\}_{i,j}$, which equals the quotient $\tilde{w}_{ij}^{(k)}/\overline{w}_{ij}^{(k)} = K_{\mathrm{ad}}(s_{ij}^{(k)}/\lambda)$ if $\overline{w}_{ij}^{(k)} > 0$. The values of these quantities are shown in grey scales, where zero corresponds to black and one to white, respectively. Moreover, we included a scaling factor `tadjust`, which allows a manipulation of the memory bandwidth τ. In the package **aws**, this is given as

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

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

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

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

We applied three test functions where the structural Assumption 2 is violated. First, we used a piecewise smooth function, given as

(5.1)
$$\theta(x) := \begin{cases} 7 + x/250 & \text{if } x \in \{1, ..., 250\}, \\ 11 + ((x-450)/100)^2/2 & \text{if } x \in \{251, ..., 750\}, \\ 6 - (x-750)/200 & \text{if } x \in \{751, ..., 1000\}. \end{cases}$$

Second, we used a piecewise constant function with small discontinuities and three different regions of monotonicity,

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

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

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

Here, the parameter values change slowly.

Additionally, we consider a shifted and scaled indicator function. This piecewise constant setting coincides with the setting that the original Propagation-Separation Approach in [Polzehl and Spokoiny, 2006] and its simplified version in Notation 2.3 assume. Let the sample size be even, $n \in 2\mathbb{N}$. Then, we split the design into two parts with coinciding cardinality, $\mathcal{X}_1 := \{X_i\}_{i=1}^{n/2}$ and $\mathcal{X}_2 := \{X_i\}_{i=n/2+1}^{n}$. We consider the test function

(5.4)
$$\theta(x) := 1 + 4 \cdot \mathbf{1}_{\mathcal{X}_2}(x), \qquad x \in \mathcal{X},$$

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

5.2. **Formation of the associated step function.** We study the formation of the associated step function, which we introduced in § 3.3. For this purpose, we visualize the resulting weighting schemes.

In Figure 1, we consider the piecewise smooth function (5.1) with Gaussian observations. In the first row, we provide the weighting schemes of the iteration step where the MAE is minimized. The product of the adaptive term $\{K_{\text{ad}}(s_{ij}^{(k)}/\lambda)\}_{i,j}$ (a) and the non-adaptive weights $\{\overline{w}_{ij}^{(k)}\}_{i,j}$ (b) results in the adaptive weights $\{\tilde{w}_{ij}^{(k)}\}_{i,j}$ (c). This illustrates the interaction of adaptation and location. For `hmax=2000`, the algorithm results in the associated step function (d). Here, the adaptive weights (f) and the weighting scheme of the corresponding adaptive term $\{K_{\text{ad}}(s_{ij}^{(k)}/\lambda)\}_{i,j}$ (not shown) were visually indistinguishable due to the large size of the considered local neighborhood, which is determined by the non-adaptive weighting scheme (e).
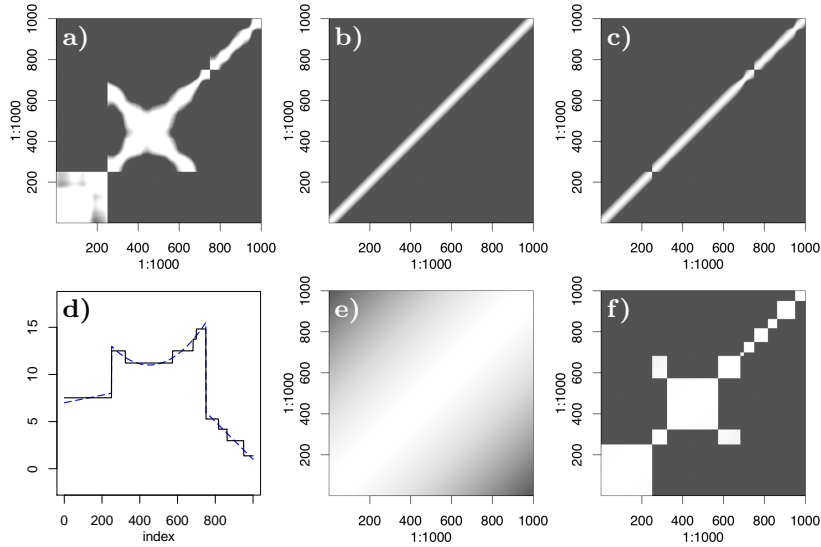
FIGURE 1. Formation of the associated step function for the piecewise smooth function (5.1) with Gaussian observations.
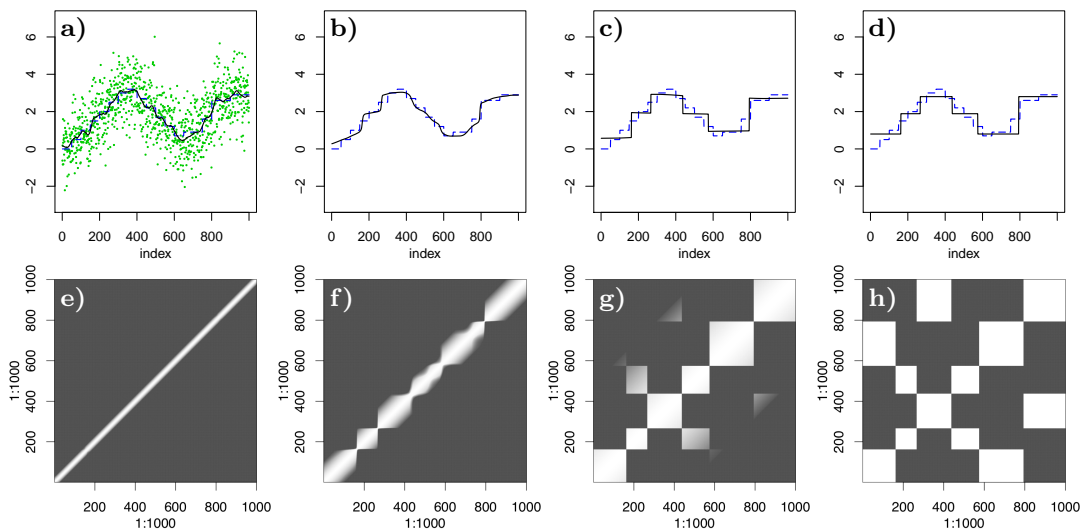

FIGURE 2. Formation of the associated step function for the step function (5.2) with Gaussian observations.

We present in Figure 2 the example plots and corresponding weighting schemes $\{\tilde{w}_{ij}^{(k)}\}_{i,j}$ of the step function (5.2). The discontinuities are too small for separation. Therefore, the algorithm forms a step function which differs from the original parameter function. The minimal MAE is provided for `hmax=30`, where the considered local neighborhood is small and separation does not yet occur (a+e). We observe, in the example plot (b) as well as in the weighting scheme (f), that the estimation function starts to form a step function for `hmax=120`. In (c), the estimation function of `hmax=600` resembles a step function, but the weighting scheme (g) already indicates that the formed steps may change with increasing location bandwidths. Indeed, in (d), several steps in different locations have been assimilated as the weighting scheme (h) points out. For the plots in the last column, we set `hmax=20000`.
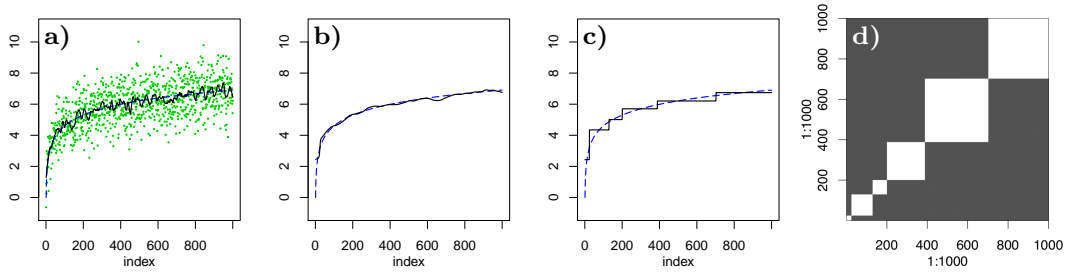
FIGURE 3. Formation of the associated step function for the logarithmic function (5.3) with Gaussian observations.
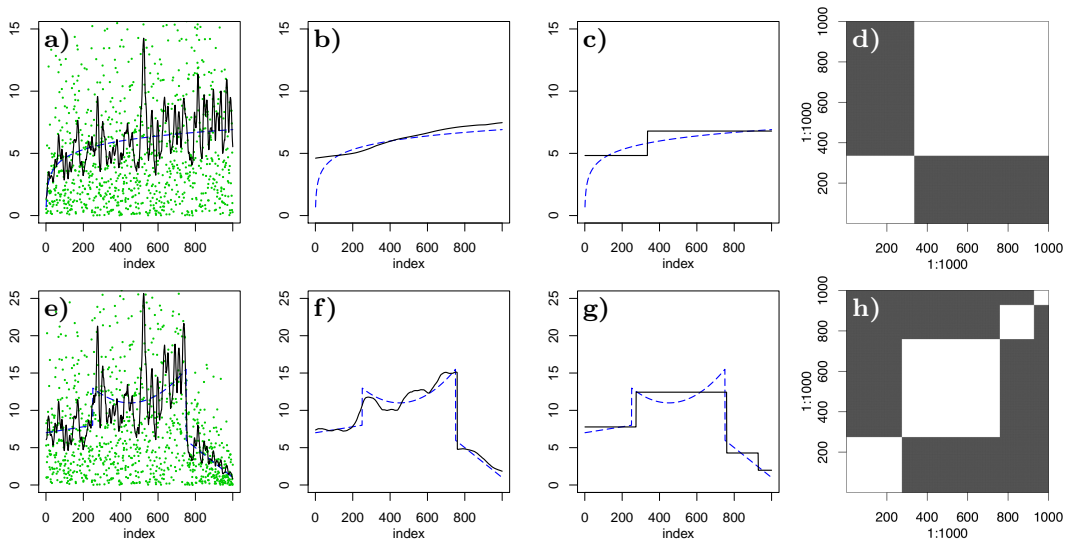

FIGURE 4. Formation of the associated step function for the piecewise smooth function (5.1) and the logarithmic function (5.3) with exponentially distributed observations.

Even for the logarithmic function (5.3), the simplified algorithm results in a step function with disjoint regions (d). In Figure 3, we show the example plots for a small location bandwidth `hmax=10` (a), at `hmax=60` (b), where the MAE is minimal, and at `hmax=2000` (c).

Additionally, we studied the formation of the associated step function for exponentially distributed observations on several test functions. In Figure 4, we provide the results for the parameter functions in Equations (5.1) (first row) and (5.3) (second row). Here again, for sufficiently large location bandwidths, the algorithm results in the associated step function with disjoint regions of non-zero adaptive weights (d+h). We show the example plots for a small location bandwidth (a+e), where `hmax=10`, an intermediate iteration step with minimized MAE (b+f), and a large location bandwidth `hmax=20000` (c+g).

5.3. **Impact of the memory step.** In [Polzehl and Spokoiny, 2006, Thm. 5.7], the memory step provided a general result on the stability of estimates, up to some constant. However, its practical use is questionable. No situation has been reported to date where the memory step considerably improved the results of the Propagation-Separation Approach. Therefore, we aim for a better understanding of its impact on the resulting estimates. For this purpose, we compared the results of the original and the simplified algorithm on the test functions in § 5.1 for Gaussian and exponentially distributed observations.
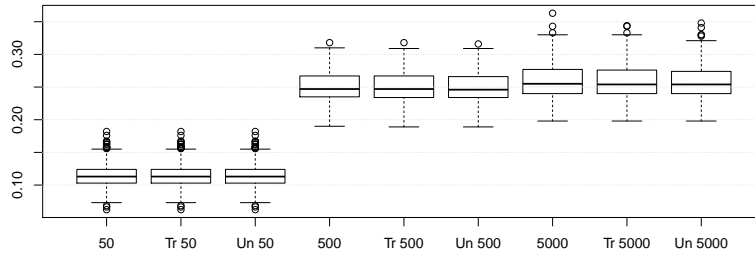
FIGURE 5. MAE-boxplots for `hmax=50,500,5000` with and without memory step, setting `aggkern= "Triangle"` (Tr) and `aggkern="Uniform"` (Un).
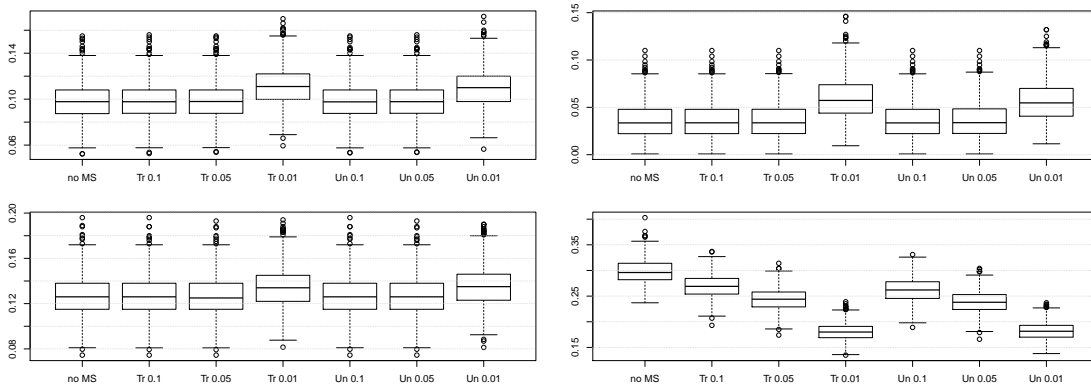

FIGURE 6. MAE-boxplots for `hmax=50` (left) and `hmax=10000` (right) for the indicator function (top) and on the piecewise smooth function (bottom) with Gaussian observations. We applied the algorithm without (no MS) and with memory step, setting `aggkern="Triangle"` (Tr) and `aggkern="Uniform"` (Un), where `tadjust=0.1,0.05,0.01` increases the amount of aggregation.

In Figure 5, we show the results for the piecewise smooth function (5.1) with Gaussian distributed observations. Here, we applied three location bandwidths, `hmax=50,500,5000`, each of them without memory step (`memory=FALSE`) and with memory step (`memory=TRUE`), using a triangular kernel (`aggkern="Triangle"`) and a uniform kernel (`aggkern="Uniform"`). As for all other test functions with Gaussian or exponentially distributed observations, there is (almost) no difference between the resulting boxplots with and without memory step and for the two memory kernels. This raises the question whether the memory step itself does not have any effect, or whether the default parameter choices in the **R**-package **aws** are unfavorable.

In order to provide a deeper insight into the mode of action of the memory step, we increased the amount of aggregation by means of the additionally implemented scaling factor `tadjust` of the memory bandwidth $\tau > 0$. In Figure 6, we present, for the indicator function (5.4) (top) and for the piecewise smooth function (5.1) (bottom), the MAE-boxplots at some early iteration step with `hmax=50` (left) and for `hmax=10000` (right), assuming Gaussian observations. The amount of aggregation increases due to the choices `tadjust=0.1,0.05,0.01` for `memory=TRUE` with `aggkern="Triangle"` and `aggkern="Uniform"`. For comparison, we show the result of the simplified procedure as well, where `memory=FALSE`. For `hmax=50`, we observe an increase of the MAE for both test functions at `tadjust=0.01`, while the MAE without memory step coincides with the results for `tadjust=0.1`
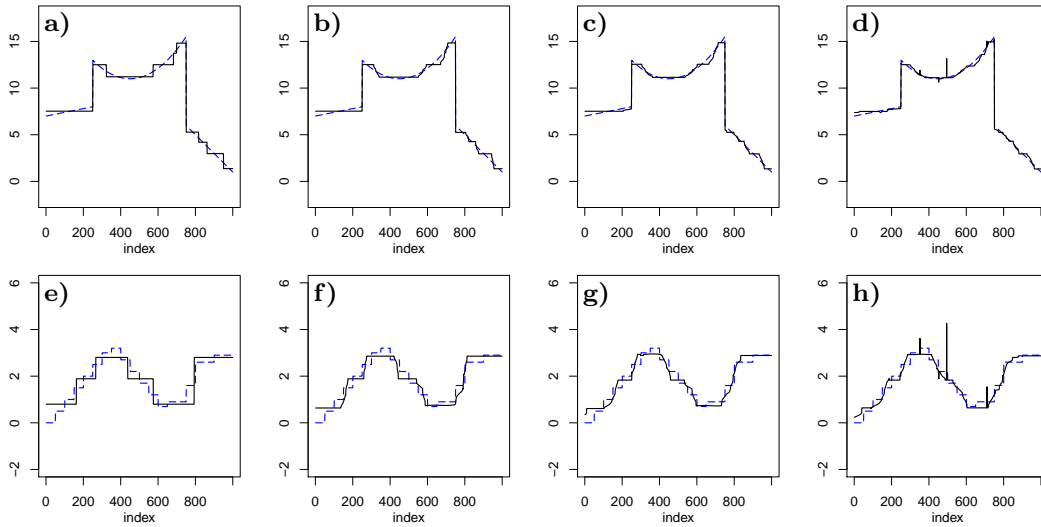
FIGURE 7. Example plots for the piecewise smooth function (top) and the step function (bottom) at `hmax=100000`. We applied the algorithm without (a+e) and with memory step, setting `aggkern="Triangle"` and (from left to right) `tadjust= 0.05,0.02,0.01`.

and `tadjust=0.05`. For `hmax=10000`, this observation remains valid for the indicator function (top right), where the locally constant model of the Propagation-Separation Approach is satisfied.

In contrast, for the piecewise smooth function, we know from § 5.2 that the estimation function approaches the associated step function, which leads to an increase of the MAE. As demonstrated in the bottom right of Figure 6, the MAE decreases with increasing amount of aggregation, that is with decreasing `tadjust`. Unfortunately, this increases the risk of adaptation to noise as we illustrate on some example plots in Figure 7 for the piecewise smooth function (top) and for the step function (5.2) (bottom), both with Gaussian observations. Without memory step as well as with memory step and `tadjust=1`, the algorithm results in the associated step function (a+e). To some extent, this effect can be attenuated by increasing the amount of aggregation, setting `tadjust=0.05` (b+f) or even `tadjust=0.02` (c+g). For `tadjust=0.01` (d+h), we observe adaptation to noise, which indicates the increased risk of adaptation to outliers due to the decreased memory bandwidth. Naturally, for other realizations, larger sample sizes or different test functions, this could happen for larger values of `tadjust` as well. We got similar results for the other test functions in § 5.1 with Gaussian and as well with exponentially distributed observations (not shown).

5.4. **Stability of estimates.** The numerical results in § 5.2 suggest that the simplified Propagation Separation Approach provides a certain stability of estimates, where the associated step function acts as an intrinsic stopping criterion. In § 6.1, we will discuss the reasons which impede a theoretical proof of this heuristic property. Here, we present some boxplots which indicate the immutability of the MAE for sufficiently large location bandwidths.

We show results for Gaussian and exponentially distributed observations (Figure 8). We set `n=1000` and `hmax=20,50,200,500,1000,10000,20000` for the former, and `n=4000` and `hmax= 50,200,500,2000,15000,20000` for the latter. Here again, we consider the indicator function (5.4), where the structural assumption of the Propagation-Separation Approach is satisfied. This leads to a decreasing MAE during iteration. As an example for the case of a misspecified model, we again apply the piecewise smooth function (5.1). Here, the MAE increases for larger location bandwidth
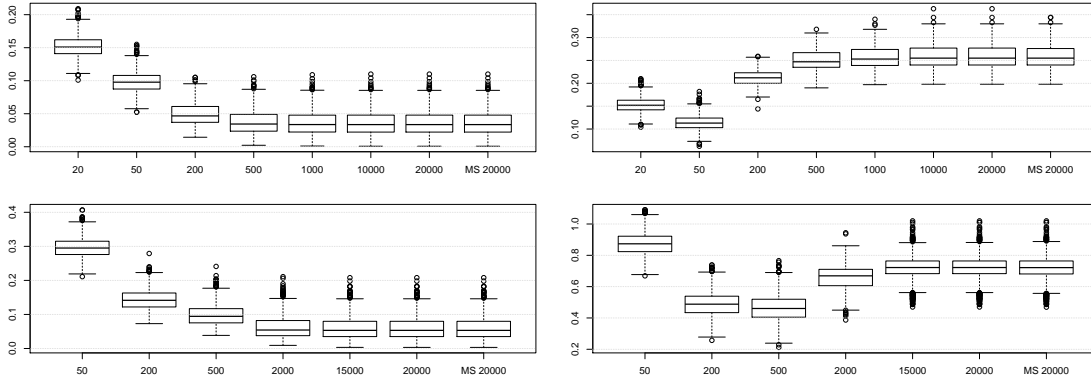
FIGURE 8. Stability of estimates for Gaussian (top) and exponentially (bottom) distributed observations for the indicator function (left) and for the piecewise smooth function (right).

as the estimator is forced into a step function. Nevertheless, for both test functions and both probability distributions, the MAE stabilizes for sufficiently large location bandwidths. For comparison, we show the MAE which results from the choices `memory=TRUE, aggkern="Triangle", tadjust=1`.

6. DISCUSSION OF THE SIMPLIFIED PROPAGATION-SEPARATION APPROACH

Our study provides theoretical and numerical results for the simplified Propagation-Separation Approach in Notation 2.3, where the memory step is omitted. This helps for a better understanding of the procedure as the impact and interaction of the involved components is clarified. Furthermore, the presented results substantiate the reasons for omitting the memory step and provide, for the first time, a detailed study of its impact.

Next we will discuss the following two questions.

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

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

6.1. **Does the Propagation-Separation Approach converge?** In § 3.3, we introduced a specific step function, which approximates the estimation function from the Propagation-Separation Approach. The formation of this *associated step function* can be explained as follows.

Due to the support $\text{supp}(K_{ad}) = [0, 1)$, the statistical penalty $s_{ij}^{(k)}$ defined in Algorithm 1 (page 3) ensures zero weights $\tilde{w}_{ij}^{(k)} = 0$ if the Kullback-Leibler divergence of the estimators from the last iteration step $\mathcal{KL}(\tilde{\theta}_i^{(k-1)}, \tilde{\theta}_j^{(k-1)})$ exceeds some lower bound $\lambda/\tilde{N}_i^{(k-1)}$. Let us consider the case where $\overline{w}_{ij}^{(k^*)} > 0$ implies $\tilde{w}_{ij}^{(k^*)} > 0$ for all $i, j \in \{1, ..., n\}$. This means that separation did not occur. The monotonicity of the sequence of location bandwidths $\{h^{(k)}\}_{k=0}^{k^*}$ ensures that the non-adaptive weights increase during iteration, and, without separation, all estimators approach each other. With $h^{(k^*)}$ sufficiently large, this results in an almost constant estimation function. However, in many cases, there are $X_i, X_j \in \mathcal{X}$ such that $\overline{w}_{ij}^{(k^*)} > 0$, but $\tilde{w}_{ij}^{(k^*)} = 0$.

We know from Proposition A.13 that separation occurs if $\max\{\mathcal{KL}(\theta_i, \theta_j) : X_i, X_j \in \mathcal{X}\}$ is sufficiently large, or if the algorithm adapts to outliers. The latter leads to separation of single observations, probably

together with some local neighborhood. If separation happens due to the variability of the true parameter function $\theta(.)$, then it starts in regions with a large absolute value of the first derivative $|\theta'(.)|$. Beginning either at the boundaries of the design space $\mathcal{X}$ or close to discontinuities and local extrema of $\theta(.)$ and $|\theta'(.)|$, this leads, by subsequent attraction and repulsion of the estimators, to the formation of a step function which approximates the associated step function in Definition 3.11.

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

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

- Each realization may yield another associated step function with slightly shifted steps.
- The improvement of the estimation quality during iteration is not ensured to be monotonic, neither for the non-adaptive nor for the adaptive estimates. Several other iterative methods, such as the expectation-maximization algorithm or the conjugate gradient method, rely on the minimization or maximization of a certain criterion. This provides a monotonic improvement of some quality criterion, which ensures the convergence of the algorithm. In contrast, the Propagation-Separation Approach considers an increasing local neighborhood, where unfavorable, newly included or stronger weighted observations may worsen the estimation quality in comparison to a previous iteration step.
- The immutability of the associated step function for sufficiently large iteration steps requires the existence of some iteration step $k_0 < \infty$ such that the considered neighborhood equals the complete design, that is $\overline{w}_{ij}^{(k_0)} > 0$ for all $X_i, X_j \in \mathcal{X}$, and $\{X_j \in \mathcal{X} : s_{ij}^{(k_0)} \leq \lambda\} = \{X_j \in \mathcal{X} : s_{ij}^{(k)} \leq \lambda\}$ for all $k > k_0$ and every $X_i \in \mathcal{X}$.

Let us consider the last reason in more detail. We know from the definition of the statistical penalty, see Algorithm 1 (page 3), that a violation of the above condition can arise from

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

We discuss these events case by case.

Recall that the non-adaptive sequence $\{\overline{N}_i^{(k)}\}_{k=0}^{k^*}$ is monotonically increasing, whereas its adaptive counterpart $\{\tilde{N}_i^{(k)}\}_{k=0}^{k^*}$ does not need to be monotonic. Therefore, we propose a slight modification of the statistical penalty in Algorithm 1, setting

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

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

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

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

and a lower bound of

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

where $\mathcal{H}_i^{(k)}$ is as in Equation (3.7) (page 8). Due to the factor $\varkappa$ in Lemma A.1, the corresponding discussion in Appendix A, and the missing monotonicity of the Kullback-Leibler divergences $\mathcal{KL}(\tilde{\theta}_i^{(k-1)}, \tilde{\theta}_j^{(k-1)})$ in $k > k_0$, this may lead to a criterion which is too restrictive to be satisfiable with $k_0 < \infty$.

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

Hence, we prefer to consider $\tilde{N}_i^{(k)}$ as the achieved improvement of the estimation quality in terms of the Kullback-Leibler divergence. This is motivated by the Theorems A.8 and A.12 and the propagation condition, which yield with high probability and in case of sharp discontinuities for $\mathcal{KL}(\tilde{\theta}_i^{(k)}, \mathcal{E}\tilde{\theta}_i^{(k)})$ the rate of convergence $\tilde{N}_i^{(k)}$, see Proposition 3.5. Nevertheless, there remains a certain probability that the intensification of the statistical penalty is not justified. Furthermore, the mentioned propagation results do not generalize to the case of model misspecification. All of these results are based on the propagation condition and this requires well separated regions. If the corresponding structural assumptions are violated, the impact of the adaptivity may change such that propagation cannot be ensured any more. In fact, model misspecification leads to a decrease of the probability for propagation. Therefore, we may still observe propagation in practice, but its probability cannot be quantified as the established exponential bounds do not hold under model misspecification. As a consequence, we cannot ensure neither the immutability of the associated step function nor the convergence of the simplified Propagation Separation Approach.

6.2. **(Where) do we need the memory step?** As discussed in [Becker and Mathé, 2013, §5], our approach is not constructed to provide asymptotic results, and the theoretical analysis is mainly restricted to piecewise bounded parameter functions. Hence, we lose the general stability of estimates by Polzehl and Spokoiny [2006, Thm 5.7]. Nevertheless, we justified the essential properties for the simplified procedure, that is propagation and separation. This emphasizes that both properties result from the adaptivity of the estimator, but not from the memory step.

From a practical point of view, the benefits of the memory step are still questionable. In § 5.3, we illustrated the impact of the memory step for several test functions. Using the default parameter choices of the **R**-package **aws** by Polzehl [2012], we could not observe any effect of the memory step. However, these choices are not arbitrary. The memory bandwidth was chosen in accordance with a former version of the propagation condition, and, indeed, we observed an increased risk of adaptation to noise for considerably smaller bandwidths. On the one hand, this emphasizes the importance of a sufficiently large memory bandwidth to avoid adaptation to outliers. On the other hand, we got a smaller MAE by increasing the amount of aggregation, which slightly attenuated the formation of a step function during iteration. In any case, we found the best results by restricting the maximal location bandwidths appropriately. The

omittance of the memory step provides a better interpretability of the procedure and, as a consequence, of the results since the memory step introduces additional interactions between the involved components, which are not fully understood yet.

6.3. **Future research.** There are several topics for future research that arise from this article. For instance, one could study the impact of the Kullback-Leibler divergence. Especially for the justification of the homogeneous propagation condition in [Becker and Mathé, 2013, §4.1] and of the inhomogeneous propagation condition in Section 4, we took advantage of its properties. Are there other (possibly asymmetric) distance functions or f-divergences which provide similar results?

Here, we would like to concentrate on another question concerning the consequences of a violated structural assumption. As indicated by our numerical results in Section 5, an appropriate stopping of the iterative procedure may reduce the resulting estimation bias considerably by avoiding the formation of a step function. For the presented univariate examples, a choice by visual inspection seems to be promising. In all observed cases, the iteration step where the formation of the step function started to dominate the smoothing result could be easily identified. Additionally, we always observed a certain range of iteration steps where the estimation quality is very similar.

However, on more complicated test functions or for higher dimensional design spaces, an automatic choice of the maximal number of iterations is desired or even required. In the context of local polynomial regression and locally weighted maximum likelihood estimation, there is a large amount of literature concerning the choice of the location bandwidth. For instance, the maximal location bandwidth $h^{(k^*)}$ could be chosen such that the non-adaptive estimator in Notation 2.2 behaves well within regions without discontinuities. Then, assuming an appropriate choice of the adaptation bandwidth λ, the simplified algorithm in Notation 1 would yield similar results as non-adaptive smoothing within these regions, while smoothing among distinct regions would be avoided as sharp discontinuities could be detected by the adaptive weights. To evaluate the appropriateness of the different approaches for the Propagation-Separation Algorithm would form a promising research project for the future. Alternatively, one could search for a criterion which takes advantage of the involved components of the method. The evaluation of the behavior of the statistical penalty or of the sum of the adaptive weights could provide useful information about the iteration step where the formation of a step function negatively affects the smoothing results.

APPENDIX A. SOME AUXILIARY RESULTS

We recall some previous results by Polzehl and Spokoiny [2006] and Becker and Mathé [2013]. The following lemma has been stated in Polzehl and Spokoiny [2006, Pages 339 & 352], but see also Becker and Mathé [2013, Lem. 2.3].

Lemma A.1. *Under Assumption 1 it holds the following.*

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

$$(A.1) \qquad \frac{I(\theta_1)}{I(\theta_2)} \leq \varkappa^2 \qquad \text{for all } \theta_1, \theta_2 \in \Theta'.$$

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

$$(A.2) \qquad \mathcal{KL}\left(\theta, \theta'\right) = \theta\left[C(\theta) - C(\theta')\right] - \left[B(\theta) - B(\theta')\right], \quad \theta, \theta' \in \Theta.$$

Equation (A.1) allows the following notations.

Notation A.2. For every compact and convex subset $\Theta' \subseteq \Theta$, we set

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

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

In this study, we usually require the set $\Theta_\varkappa$ to be sufficiently large such that $\theta(X_i) \in \Theta_\varkappa$ for all $i \in \{1, ..., n\}$. Its precise choice will be specified where necessary.

Lemma A.3. *Suppose Assumption 1 and let $\Theta_\varkappa \subseteq \Theta$ and $\varkappa \geq 1$ be as in Notation A.2. For any sequence $\theta_0, \theta_1, ..., \theta_m \in \Theta_\varkappa$, it holds*

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

For the proof, we refer the reader to Polzehl and Spokoiny [2006, Lem. 5.2]. Here, we recall some details concerning the applicability of the technical Lemma A.3 and the related Equation (A.1), see [Becker and Mathé, 2013, App. A]. For several results, we apply Equation (A.1) and Lemma A.3 not only with respect to the true parameters $\{\theta_i\}_i$, but as well with respect to the transformed observations $\{T(Y_i)\}_i$ or the associated estimates $\{\tilde{\theta}_i^{(k)}\}_i$, $k \in \{0, ..., k^*\}$. Therefore, we restrict our analysis to the favorable realizations $\{T(Y_i) \in \Theta_\varkappa$ for all $i\}$, and we quantify the probability of its complementary set. For every $\varkappa$, we use the most convenient choice of the set $\Theta_\varkappa$. Furthermore, we restrict the range of $\theta(.)$ by a subset $\Theta^* \subseteq \Theta$.

Notation A.4. We fix a subset $\Theta^* \subseteq \Theta$ and a constant $\varphi_0 \geq 0$. Then, we recall Notation A.2, and we choose $\varkappa \geq 1$ sufficiently large such that $\Theta^* \subseteq \Theta_\varkappa$. We define the function $\mathfrak{p}_\varkappa : (\Theta^*)^n \to [0, 1]$ by

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

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

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

The probability $p_\varkappa$ decreases with increasing $\varkappa \geq 1$.

Example A.5*.*

- For Gaussian and log-normal distributed observations with $\mathcal{P} = \{\mathcal{N}(\theta, \sigma^2)\}_{\theta \in \Theta}$ and $\mathcal{P} = \{\log \mathcal{N}(\theta, \sigma^2)\}_{\theta \in \Theta}$, respectively, it holds $I(\theta) = 1/\sigma^2$, leading to $\varkappa = 1$. In this case, Equation (A.1) and Lemma A.3 hold for every subset $\Theta' \subseteq \Theta$ without the restriction to compact sets, and we get $p_\varkappa = 0$. This is the optimal scenario.
- For the Gamma, Erlang, scaled chi-squared, exponential, Rayleigh, Weibull, and Pareto distributions, it holds after reparametrization $I(\theta) = 1/\theta^2$. In this case $\varkappa$ and $p_\varkappa$ become large. However, the effective values of $\varkappa$ and $p_\varkappa$ may be much smaller than the global ones, which attenuates the consequences in practice.

Next we recall an exponential bound by Polzehl and Spokoiny [2006, Thm. 6.1], from which the Theorems A.8 and A.12 follow as special cases.

Theorem A.6. *Suppose Assumption 1, and reparametrize $v := C(\theta)$ and $D(v) := B(\theta)$. Furthermore, let $\overline{W}_i := \{\overline{w}_{ij}\}_{j=1}^n \in [0, 1]^n$ be a weighting scheme, and consider the corresponding non-adaptive estimator $\overline{\theta}_i$ in Notation 2.2 and its expectation $\breve{\theta}_i := \mathbb{E}\overline{\theta}_i = \sum_j \overline{w}_{ij}\theta_j/\overline{N}_i$. We set $q(u|v) := \mathcal{KL}(v, v + u)$ and define, for a given constant $z \geq 0$ and $\breve{v}_i = C(\breve{\theta}_i)$, the set*

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

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

(A.3) $$q(\mu u \overline{w}_{ij} | v_j) \leq (1 + \alpha) \mu^2 \overline{w}_{ij} q(u | \breve{v}_i), \qquad j = 1, ..., n,$$

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

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

Remark A.7. Polzehl and Spokoiny [2006] assumed the sufficient statistic T in Assumption 1 to be the identity map. Fortunately, Theorem A.6 depends on the probability distribution and consequently on T via the Kullback-Leibler divergence only. This ensures with Lemma A.1 (3) that the choice of T does not have any effect, and the original result remains valid.

The next result can be found in [Polzehl and Spokoiny, 2006, Thm. 2.1].

Theorem A.8. *Let Assumption 1 be satisfied, and presume a parametric model,* $\theta(.) \equiv \theta$. *Then, for each* $i \in \{1, ..., n\}$ *and every weighting scheme* $\overline{W}_i := \{\overline{w}_{ij}\}_{j=1}^n \in [0, 1]^n$, *we get*

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

with $\overline{N}_i$ *and* $\overline{\theta}_i$ *as in Notation 2.2.*

Next we extend Theorem A.8 to parameter functions with bounded variability. In the corresponding proof, Polzehl and Spokoiny [2006] used Equation (A.1) (page 20). Although not stated in [Polzehl and Spokoiny, 2006, Thm. 2.2], this requires a restriction to the favorable realizations, which may worsen the result. In order to quantify the probability of the complementary set, we proceed in an analogous manner as in Notation A.4. We consider a different set of realizations, whose definition will be motivated in the proof of Theorem A.12 (page 28).

Notation A.9. Recall Notation A.2. We fix a subset $\Theta^* \subseteq \Theta$ and a constant $\varphi_0 \geq 0$. Let $\varkappa \geq 1$ be sufficiently large such that $\Theta^* \subseteq \Theta_\varkappa$. Then, for every $i \in \{1, ..., n\}$, we consider the non-adaptive estimator $\overline{\theta}_i$ in Notation 2.2 with weighting scheme $\overline{W}_i := \{\overline{w}_{ij}\}_{j=1}^n \in [0, 1]^n$. The function $\breve{\mathfrak{p}}_\varkappa : (\Theta^*)^n \to [0, 1]$ is given by

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

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

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

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

Example A.10. For Gaussian and log-normal distributed observations, it holds $\breve{p}_\varkappa = 0$ since $\varkappa = 1$ for every set $\Theta_\varkappa \subseteq \Theta$. For the Gamma and its related distributions, the probability $\breve{p}_\varkappa$ may be large, and it increases with decreasing values of $\varkappa$ as well as with increasing sample sizes.

The probabilities $\breve{p}_{\varkappa,0}$ in Notation A.9 and $p_\varkappa$ in Notation A.4 are closely related.

Corollary A.11. Suppose Assumption 1 and the setting of Notation A.9. Then, it holds

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

and, as a consequence, we get $\breve{p}_{\varkappa,0} \geq p_\varkappa$, where $p_\varkappa$ is as in Notation A.4.

Theorem A.12. *Suppose Assumption 1, and fix a subset $\Theta^* \subseteq \Theta$ and a constant $\varphi_0 \geq 0$ such that $\{\theta_i\}_{i=1}^n \in (\Theta^*)^n$ and $\max_{i,j} \mathcal{KL}(\theta_i, \theta_j) \leq \varphi_0^2$. Moreover, recall Notation A.2 and let $\varkappa \geq 1$ be sufficiently large such that $\Theta^* \subseteq \Theta_\kappa$. Finally, let $\overline{W}_i := \{\overline{w}_{ij}\}_{j=1}^n \in [0,1]^n$ denote a weighting scheme, and recall the corresponding quantities $\overline{\theta}_i$ and $\overline{N}_i$ in Notation 2.2. Then, for each $i \in \{1, ..., n\}$ and every $z > 0$, it holds*

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

where $\breve{p}_\varkappa$ is as in Notation A.9.

We give the proof in Appendix B in order to clarify the appearance of the probability $\breve{p}_\varkappa$. Finally, we recall the separation property, which can be found in [Polzehl and Spokoiny, 2006, Thm. 5.9] and [Becker and Mathé, 2013, Prop. 3.2].

Proposition A.13. *Suppose Assumption 1, and consider two design points $X_{i_1}, X_{i_2} \in \mathcal{X}$. Assume that the realization at hand satisfies at these points in iteration step k the estimation accuracy $\mathcal{KL}(\tilde{\theta}_{i_m}^{(k)}, \theta_{i_m}) \leq z_m^{(k)} := z/\overline{N}_{i_m}^{(k)}$ with some constant $z > 0$ and $\theta_{i_m}, \tilde{\theta}_{i_m}^{(k)} \in \Theta_\varkappa$, $m = 1, 2$, for $\varkappa \geq 1$ fixed and $\Theta_\varkappa$ as in Notation A.2. If additionally*

$$(A.4) \qquad \mathcal{KL}^{1/2}(\theta_{i_1}, \theta_{i_2}) > \varkappa\left(\sqrt{\lambda/\tilde{N}_{i_1}^{(k)}} + \sqrt{z_1^{(k)}} + \sqrt{z_2^{(k)}}\right),$$

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

APPENDIX B. PROOFS

Proof of Proposition 3.5. The inhomogeneous propagation condition yields the monotonicity of the function $\hat{\mathfrak{z}}_\lambda\left(k, p; \{\theta_j\}_{j=1}^n, i\right)$ in $k \leq k_0$ for all $p \in (\epsilon, 1)$ and every $i \in \{1, ..., n\}$. This implies Equation (3.3). We turn to Equation (3.2), and we consider the event $\breve{\Omega}_\varkappa$ in Corollary A.11. The adaptive estimator is defined as a weighted mean of the transformed observations. Therefore, for all $k \in \{0, ..., k^*\}$, we get $\breve{\Omega}_\varkappa \subseteq \{\tilde{\theta}_i^{(k)} \in \Theta_\varkappa, i \in \{1, ..., n\}\}$. Then, we use the convexity of the Kullback-Leibler divergence with respect to the first argument, see Lemma A.1. Denoting the complement of the set M by M^c, it follows from the initialization of Algorithm 1 (page 3) with $\tilde{\theta}_i^{(0)} = \overline{\theta}_i^{(0)}$ that

$$\mathbb{P}\left(\overline{N}_i^{(k)} \mathcal{KL}\left(\tilde{\theta}_i^{(k)}, \theta_i\right) > z\right)$$

$$\overset{\text{Lem. A.3}}{\leq} \mathbb{P}\left(\left\{\varkappa^2 \overline{N}_i^{(k)}\left[\mathcal{KL}^{1/2}\left(\tilde{\theta}_i^{(k)}, \mathcal{E}\tilde{\theta}_i^{(k)}\right) + \mathcal{KL}^{1/2}\left(\mathcal{E}\tilde{\theta}_i^{(k)}, \theta_i\right)\right]^2 > z\right\} \cap \breve{\Omega}_\varkappa\right) + \mathbb{P}\left(\breve{\Omega}_\varkappa^c\right)$$

$$\overset{\text{Lem. A.1}}{\leq} \mathbb{P}\left(\left\{\overline{N}_i^{(k)} \mathcal{KL}\left(\tilde{\theta}_i^{(k)}, \mathcal{E}\tilde{\theta}_i^{(k)}\right) > \left[\sqrt{z}/\varkappa - \varphi\right]^2\right\} \cap \breve{\Omega}_\varkappa\right) + \breve{p}_{\varkappa,0}$$

$$\overset{\text{Eq. (3.3)}}{\leq} \max\left\{\mathbb{P}\left(\left\{\overline{N}_i^{(0)} \mathcal{KL}\left(\overline{\theta}_i^{(0)}, \mathbb{E}\overline{\theta}_i^{(0)}\right) > \left[\sqrt{z}/\varkappa - \varphi\right]^2\right\} \cap \breve{\Omega}_\varkappa\right), \epsilon\right\} + \breve{p}_{\varkappa,0}$$

$$\overset{\text{Thm. A.12}}{\leq} \max\left\{2e^{-\left[\sqrt{z}/\varkappa - \varphi\right]^2/\varkappa^2}, \epsilon\right\} + \breve{p}_{\varkappa,0}$$

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

Proof of Theorem 3.9. Recall the event $\breve{\Omega}_{\varkappa}$ in Corollary A.11, and let M^c denote the complement of the set M. We construct a disjoint union

$$\left[\mathcal{B}^{(k_0)}(z)\right]^c = \bigcup_{k=0}^{k_0} \left(\left[\mathcal{B}^{(k)}(z)\right]^c \cap \left[\bigcap_{k'=0}^{k-1} \mathcal{B}^{(k')}(z)\right]\right),$$

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

(B.1) $\qquad \mathbb{P}\left(\mathcal{B}^{(k_0)}(z)|M^{(k_0)}(z)\right)$

$$\geq 1 - \frac{\left[\breve{p}_{\varkappa,0} + \sum_{k=0}^{k_0} \mathbb{P}\left(M^{(k)}(z) \cap \breve{\Omega}_{\varkappa} \cap \left[\mathcal{B}^{(k)}(z)\right]^c \cap \left[\bigcap_{k'=0}^{k-1} \mathcal{B}^{(k')}(z)\right]\right)\right]}{\mathbb{P}\left(M^{(k_0)}(z)\right)},$$

where we used that $M^{(k_0)}(z) \subseteq M^{(k)}(z)$ for $k \leq k_0$. The choice of $h^{(0)}$ ensures, for every $i \in \{1,...,n\}$, that $U_i^{(0)} \setminus \mathcal{V}_i = \emptyset$. Moreover, it holds $\tilde{\theta}_i^{(0)} = \overline{\theta}_i^{(0)}$ by the initialization of Algorithm 1 (page 3), and it follows in the same manner as in the proof of Proposition 3.5 that

(B.2) $\qquad \mathbb{P}\left(M^{(0)}(z) \cap \breve{\Omega}_{\varkappa} \cap \left[\mathcal{B}^{(0)}(z)\right]^c\right) \overset{\overline{n}_i^{(0)}=\overline{N}_i^{(0)}}{\leq} n \cdot \mathbb{P}\left(\overline{N}_i^{(0)} \mathcal{KL}(\overline{\theta}_i^{(0)}, \theta_i) > z\right)$

$$\leq 2\,ne^{-\left[\sqrt{z}/\varkappa - \varphi\right]^2/\varkappa^2}.$$

By definition of the events $\mathcal{B}^{(k)}(z)$, $M^{(k)}(z)$, and $\breve{\Omega}_{\varkappa}$, the conditions of Proposition A.13 are satisfied on the intersection

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

for all $k \in \{1,...,k_0\}$. There, it follows that $\tilde{w}_{ij}^{(k)} = 0$ for all $X_j \notin U_i^{(k)} \cap \mathcal{V}_i$. Hence, smoothing is restricted to the neighborhood $\mathcal{V}_i$, and $\mathbb{E}\left[T(Y_j)\right] = \theta_i$ for every X_j with $\tilde{w}_{ij}^{(k)} > 0$. Then, we get with Proposition 3.5 that

$$\mathbb{P}\left(\{\overline{n}_i^{(k)} \mathcal{KL}\left(\tilde{\theta}_i^{(k)}, \theta_i\right) > z\} \cap M^{(k)}(z) \cap \breve{\Omega}_{\varkappa} \cap \left[\bigcap_{k'=0}^{k-1} \mathcal{B}^{(k')}(z)\right]\right)$$

(B.3) $$\leq \max\left\{2e^{-\left[\sqrt{z}/\varkappa - \varphi\right]^2/\varkappa^2}, \epsilon\right\}$$

for all $k \in \{1,...,k_0\}$. Finally, Equations (B.1), (B.2), and (B.3) lead to

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

This terminates the proof. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ $\square$

Proof of Lemma 4.3. It holds

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

Hence, in each of these cases, the non-adaptive estimator follows the same distribution as for Gaussian or exponentially distributed observations. Additionally, it holds $\mathrm{Exp}(1/\theta) = \Gamma(1, \theta)$, $\mathrm{Erlang}(k, \theta) = \Gamma(k, \theta)$, and $Y \sim \Gamma(k/2, 2\theta/k)$ if $kY/\theta \sim \chi^2(k) = \Gamma(k/2, 2)$, where $k \in \mathbb{N}$. Since the associated

Kullback-Leibler divergences coincide, it suffices to show the assertion for the Gaussian and the Gamma distribution, which satisfy Assumption 1 with $T = Id$. By Lemma A.1, the function $g_\theta(y) = \mathcal{KL}(y, \theta)$ fulfills, for every $y \in \mathcal{Y}$, that

(B.4)
$$\frac{dg_\theta}{dy}(y) = [C(y) - C(\theta)],$$

where we used that $B'(y) = yC'(y)$. Due to the strict monotonicity of the function C, $\{Y > \theta\}$ and $\{Y \le \theta\}$ restrict the random variable $\mathcal{KL}(y, \theta)$ to its regions of monotonicity. On each of these regions the assertion follows from [Becker and Mathé, 2013, Ex. 4.3]. □

Proof of Lemma 4.4. The parametric family of probability distributions $\mathcal{P}$ is known, and it characterizes, for every $\vartheta \in \Theta$, the function $g_\vartheta : \mathcal{Y} \to [0, \infty)$ given by $g_\vartheta(y) = \mathcal{KL}(T(y), \vartheta)$. Moreover, the probability distribution of the observations $\mathcal{Y}_i \overset{iid}{\sim} \mathbb{P}_\vartheta \in \mathcal{P}$ determines the probability distribution of the random quantities $\tilde{N}_i^{(k)}$ and $\tilde{\vartheta}_i^{(k)}$ and consequently of the random variables $\{\tilde{N}_i^{(k)} \mathcal{KL}(\tilde{\vartheta}_i^{(k)}, \vartheta)\}_i$, on which the function $\mathfrak{Z}_\lambda$ is based. The function $\mathfrak{Z}_\lambda$ is invariant with respect to the parameter $\vartheta \in \Theta$ if and only if the homogeneous propagation condition is invariant with respect to ϑ. Therefore, it suffices to show that, for every $\vartheta \in \Theta$, the functions $p_\vartheta^{(l)}$ with $l = 1, 2, 3$ allow exact reconstruction of $\mathbb{P}_\vartheta$ via the inverse of g_ϑ.

If the sufficient statistic in Assumption 1 satisfies $T = Id$, then it follows from Equation (B.4) that the inverse g_ϑ^{-1} has exactly one solution on $\{y < \vartheta\}$ and $\{y > \vartheta\}$, respectively, and it holds $g_\vartheta^{-1}(0) = \vartheta$. If $T \ne Id$, then we get

$$\frac{dg_\vartheta}{dy}(y) = T'(y) [C(T(y)) - C(\vartheta)],$$

where the assumed strict monotonicity of T leads again to the regions of monotonicity $\{T(y) > \vartheta\}$ and $\{T(y) \le \vartheta\}$. Furthermore, knowledge of $p_\vartheta^{(l)}$ with $l = 1, 2, 3$ yields knowledge of $\mathbb{P}(\{T(\mathcal{Y}_i) > \vartheta\} \cap \{\mathcal{KL}(T(\mathcal{Y}_i), \vartheta) \le z\})$, $i \in \{1, ..., n\}$. Therefore, we can reconstruct $\mathbb{P}_\vartheta$ for every ϑ from $p_\vartheta^{(l)}$, $l = 1, 2, 3$, which leads to the assertion. □

For the proof of Proposition 4.5, we recall the following basic result.

Corollary B.1. Let $a_1, a_2, b_1, b_2 \in \mathbb{R}$ satisfy $(a_1 - b_1) \cdot (a_2 - b_2) \ge 0$. Then it holds

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

Proof of Proposition 4.5. Lemma A.3 provides on M_0, for all $i, j \in \{1, ..., n\}$, that

(B.5)
$$\mathcal{KL}(Y_i, Y_j) \le \varkappa^2 \left[\mathcal{KL}^{1/2}(Y_i, \theta_i) + \mathcal{KL}^{1/2}(\theta_i, \theta_j) + \mathcal{KL}^{1/2}(Y_j, \theta_j) \right]^2.$$

On a certain set of realizations this upper bound can be improved by Corollary B.1. For this purpose, we distinguish the following sets

(B.6)
$$M_1 := \{(Y_i - \theta_i) \cdot (Y_j - \theta_j) \ge 0\} \text{ and } M_2 := \{(Y_i - \theta_i) \cdot (Y_j - \theta_j) < 0\}$$

and analogously in the homogeneous setting

(B.7)
$$M_3 := \{(\mathcal{Y}_i - \vartheta) \cdot (\mathcal{Y}_j - \vartheta) \ge 0\} \text{ and } M_4 := \{(\mathcal{Y}_i - \vartheta) \cdot (\mathcal{Y}_j - \vartheta) < 0\}.$$

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

and on M_2, the resulting formulas can be reduced to the divergence $\mathcal{KL}(\mathcal{Y}_i, \mathcal{Y}_j)$, which will lead to the assertion.

On the set M_2, we just use the upper bound (B.5). For the set M_1, we get from Taylor's Theorem for all $\theta_1, \theta_2 \in \Theta$ the existence of a parameter $\theta^* \in \Theta$ between θ_1 and θ_2 which satisfies

$$\text{(B.8)} \qquad \mathcal{KL}(\theta_1, \theta_2) = \tfrac{1}{2I(\theta^*)} \left[C(\theta_1) - C(\theta_2)\right]^2.$$

Therefore, on $M_1 \cap M_0$ it follows from Corollary B.1 and the monotonicity of the function C that

$$\mathcal{KL}(Y_i, Y_j) \quad \leq \quad \frac{1}{2I(\theta^*)} \left[||C(Y_i) - C(\theta_i)| - |C(Y_j) - C(\theta_j)|| + |C(\theta_i) - C(\theta_j)|\right]^2$$

$$\text{(B.9)} \qquad \overset{\text{Eq. (A.1)}}{\leq} \quad \varkappa^2 \left[\left|\mathcal{KL}^{1/2}(Y_i, \theta_i) - \mathcal{KL}^{1/2}(Y_j, \theta_j)\right| + \mathcal{KL}^{1/2}(\theta_i, \theta_j)\right]^2.$$

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

$$\mathbb{P}\left(\{\mathcal{KL}(Y_i, Y_j) > z\} \cap M_0\right)$$

$$\leq \quad \mathbb{P}\left(M_1 \cap M_0 \cap \left\{\left|\mathcal{KL}^{1/2}(Y_i, \theta_i) - \mathcal{KL}^{1/2}(Y_j, \theta_j)\right|^2 > \left[\sqrt{z}/\varkappa - \varphi_0\right]^2\right\}\right)$$

$$+ \mathbb{P}\left(M_2 \cap M_0 \cap \left\{\left[\mathcal{KL}^{1/2}(Y_i, \theta_i) + \mathcal{KL}^{1/2}(Y_j, \theta_j)\right]^2 > \left[\sqrt{z}/\varkappa - \varphi_0\right]^2\right\}\right)$$

$$= \quad \mathbb{P}\left(M_3 \cap M_0 \cap \left\{\left[\mathcal{KL}^{1/2}(\mathcal{Y}_i, \vartheta) - \mathcal{KL}^{1/2}(\mathcal{Y}_j, \vartheta)\right]^2 > \left[\sqrt{z}/\varkappa - \varphi_0\right]^2\right\}\right)$$

$$+ \mathbb{P}\left(M_4 \cap M_0 \cap \left\{\left[\mathcal{KL}^{1/2}(\mathcal{Y}_i, \vartheta) + \mathcal{KL}^{1/2}(\mathcal{Y}_j, \vartheta)\right]^2 > \left[\sqrt{z}/\varkappa - \varphi_0\right]^2\right\}\right).$$

Equation (B.8) leads on $M_3 \cap M_0$ with appropriate parameters $\vartheta_1^*, \vartheta_2^* \in \Theta_\varkappa$ to

$$\left[\mathcal{KL}^{1/2}(\mathcal{Y}_i, \vartheta) - \mathcal{KL}^{1/2}(\mathcal{Y}_j, \vartheta)\right]^2$$

$$\overset{M_3}{\leq} \quad \max\left\{\frac{1}{2I(\vartheta_1^*)}, \frac{1}{2I(\vartheta_2^*)}\right\} \left[(C(\mathcal{Y}_i) - C(\vartheta)) - (C(\mathcal{Y}_j) - C(\vartheta))\right]^2$$

$$\overset{\text{Eq. (A.1)}}{\leq} \quad \varkappa^2 \mathcal{KL}(\mathcal{Y}_i, \mathcal{Y}_j).$$

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

$$\left[\mathcal{KL}^{1/2}(\mathcal{Y}_i, \vartheta) + \mathcal{KL}^{1/2}(\mathcal{Y}_j, \vartheta)\right]^2 \overset{\text{Eq. (A.1)}}{\leq} \varkappa^2 \mathcal{KL}(\mathcal{Y}_i, \mathcal{Y}_j).$$

Hence, we conclude that

$$\mathbb{P}\left(\{\mathcal{KL}(Y_i, Y_j) > z\} \cap M_0\right) \quad \leq \quad \mathbb{P}\left(\left\{\varkappa^2 \mathcal{KL}(\mathcal{Y}_i, \mathcal{Y}_j) > \left[\sqrt{z}/\varkappa - \varphi_0\right]^2\right\} \cap M_0\right)$$

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

This terminates the proof. $\qquad \square$

Justification of Claim 4.6. Recall Notation 4.1. For simplicity, we concentrate on the case $T = \mathrm{Id}$ as presumed in Proposition 4.5. For $T \neq Id$, the assertion follows just as for $T = Id$, replacing in the proof of Proposition 4.5 and in the following formulas the observations Y_i and $\mathcal{Y}_i$ by the transformed observations $T(Y_i)$ and $T(\mathcal{Y}_i)$ for all $i \in \{1, ..., n\}$. Recall from the proof of Lemma 4.4 that the assumed strict monotonicity of T ensures that the regions of monotonicity in Equation (B.4) remain valid.

We know from Theorems A.8 and A.12 that the Kullback-Leibler divergence between the non-adaptive estimator and its expectation converges, in probability, at least with rate $\overline{N}_i^{(k)}$ under homogeneity and under inhomogeneity. The choice of λ is in accordance with the homogeneous propagation condition by assumption. Hence, it compensates with probability $1 - \epsilon$ the impact of the adaptivity under homogeneity, and it only depends on the functions $p_\vartheta^{(l)}$, $l = 1, 2, 3$, see Lemma 4.4. Due to the assumed invariance of $p_\vartheta^{(l)}$ with respect to ϑ, it holds $p_{\theta_i}^{(l)} = p_\vartheta^{(l)}$ for every $l = 1, 2, 3$ and all $i \in \{1, ..., n\}$. Therefore, it suffices to increase the homogeneous bandwidth λ pursuant to the maximal impact of the local variability of $\theta(.)$, but independent of the precise definition of $\theta(.)$.

The latter effects the interplay of the observations and hence the adaptive weights, where we consider the random variables $\{s_{ij}^{(k)}\}_{i,j}$, see Algorithm 1 (page 3). Proposition 4.5 provides, on the set M_0, an upper bound for the augmentation of the random variable $\mathcal{KL}(Y_i, Y_j)$ compared to $\mathcal{KL}(\mathcal{Y}_i, \mathcal{Y}_j)$. It justifies the given choice of λ_φ for the iteration step $k = 1$ if $h^{(0)}$ satisfies $\overline{w}_{ij}^{(0)} = 0$ for all $X_i \neq X_j$. We seek for a generalization to other choices of $h^{(0)}$ and the subsequent iteration steps.

The adaptive estimator is defined as a weighted mean of the observations. Therefore, for all $k \in \{0, ..., k^*\}$, it holds $M_0 \subseteq \{\tilde{\theta}_i^{(k)}, \tilde{\vartheta}_i^{(k)} \in \Theta_\varkappa, i \in \{1, ..., n\}\}\}$, where M_0 is as in Proposition 4.5. This enables on M_0 the application of Equation (A.1) (page 20) and Lemma A.3 with respect to the adaptive estimates. We distinguish the same cases as in the proof of Proposition 4.5, recall Equations (B.6) and (B.7) and the corresponding upper bounds in Equations (B.5) and (B.9). For the sake of brevity, we summarize both cases in one equation, using the operation $\pm$. Then, we get on the set M_0 in a uniform manner as in the proof of Proposition 4.5 that

$$
\begin{aligned}
s_{ij}^{(k)} \quad \leq \quad & \varkappa^2 \tilde{N}_i^{(k-1)} \left[\left| \mathcal{KL}^{1/2}\left(\tilde{\theta}_i^{(k-1)}, \mathcal{E}\tilde{\theta}_i^{(k-1)}\right) \pm \mathcal{KL}^{1/2}\left(\tilde{\theta}_j^{(k-1)}, \mathcal{E}\tilde{\theta}_j^{(k-1)}\right) \right| \right. \\
& \left. \quad + \mathcal{KL}^{1/2}\left(\mathcal{E}\tilde{\theta}_i^{(k-1)}, \mathcal{E}\tilde{\theta}_j^{(k-1)}\right) \right]^2,
\end{aligned}
$$
(B.10)

where $\mathcal{E}\tilde{\theta}_i^{(k)}$ is as in Notation 3.1. The variability of the parameter function $\theta(.)$ effects the third summand, which satisfies by Equation (A.1) (page 20) and the convexity of the Kullback-Leibler divergence with respect to the first argument that

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

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

$$
\sqrt{\tilde{N}_i^{(k-1)}} \left| \mathcal{KL}^{1/2}\left(\tilde{\theta}_i^{(k-1)}, \mathcal{E}\tilde{\theta}_i^{(k-1)}\right) \pm \mathcal{KL}^{1/2}\left(\tilde{\theta}_j^{(k-1)}, \mathcal{E}\tilde{\theta}_j^{(k-1)}\right) \right|
$$
(B.11)

forms the inhomogeneous analog of

$$
\sqrt{\tilde{N}_i^{(k-1)}} \left| \mathcal{KL}^{1/2}\left(\tilde{\vartheta}_i^{(k-1)}, \vartheta\right) \pm \mathcal{KL}^{1/2}\left(\tilde{\vartheta}_j^{(k-1)}, \vartheta\right) \right|.
$$
(B.12)

However, the corresponding probability distributions cannot be compared as for the single observations since the probability distributions of $\tilde{\vartheta}_l^{(k-1)}$ and $\tilde{\theta}_l^{(k-1)}$, $l = i, j$, may differ considerably. Nevertheless, it follows in the same lines as at the end of the proof of Proposition 4.5 that

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

Hence, Equation (B.12) is controlled by $\sqrt{\lambda}$, up to the factor $\varkappa$. Similarly, Equation (B.11) mainly depends on the randomness of the observations. Admittedly, this cannot be proven due to the impact of the adaptive weights which are influenced by the variability of the inhomogeneous parameter function.

Instead, we follow an inductive argumentation, considering the relation to the non-adaptive estimator. The initialization of the algorithm by the non-adaptive estimator serves as the base clause. Assuming that the adaptive weights in iteration step k are, with high probability, similar to the non-adaptive ones, we get that the divergence $\mathcal{KL}(\tilde{\theta}_i^{(k)}, \mathcal{E}\tilde{\theta}_i^{(k)})$ behaves similar to $\mathcal{KL}(\overline{\theta}_i^{(k)}, \mathbb{E}\overline{\theta}_i^{(k)})$. Additionally, we know from Theorems A.8 and A.12 that $\mathcal{KL}(\tilde{\vartheta}_i^{(k)}, \vartheta)$ and $\mathcal{KL}(\overline{\theta}_i^{(k)}, \mathbb{E}\overline{\theta}_i^{(k)})$ satisfy, in probability, the same rate of convergence. The divergence $\mathcal{KL}(\overline{\vartheta}_i^{(k)}, \vartheta)$ relates via the homogeneous propagation condition to the divergence $\mathcal{KL}(\tilde{\vartheta}_i^{(k)}, \vartheta)$ and, as a consequence, to Equation (B.12), which we controlled by the constant $\varkappa\sqrt{\lambda}$. This motivates together with Proposition 4.5 and the invariance of the functions $p_\vartheta^{(l)}$, $l = 1, 2, 3$, with respect to $\theta \in \Theta$, the supposition that the impact of the variability of the parameter function on Equation (B.11) is sufficiently small such that $\varkappa\sqrt{\lambda}$ can still control it. Then, we may conclude that the choice

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

ensures in the next iteration step $k+1$ the similarity of the adaptive and the non-adaptive weights, yielding on M_0 the desired behavior of $\hat{3}_{\lambda_\varphi}$. The restriction to the set M_0 leads to an increased probability level of $\epsilon + 2p_\varkappa$ since $\mathbb{P}(M_0^c) \leq 2p_\varkappa$. $\qquad\square$

Proof of Theorem A.12. Let $q(u|v) = \mathcal{KL}(v, v + u)$ be as in Theorem A.6. The reparametrization $v = C(\theta)$ and $D(v) = B(\theta)$ yields $D'(v) = \theta$, $D''(v) = 1/I(\theta)$, and

(B.13) $$\mathcal{KL}\left(\mathbb{P}_{v_1}, \mathbb{P}_{v_2}\right) = D'(v_1)\left[v_1 - v_2\right] - \left[D(v_1) - D(v_2)\right].$$

This provides with the Taylor expansion that

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

where the remainder is in Lagrange form, and $c \in [0, 1]$ is chosen appropriately. We set $\alpha := \varkappa^2 - 1$, and recall that $\overline{w}_{ij}^2 \leq \overline{w}_{ij}$ since $\overline{w}_{ij} \in [0, 1]$. For $c_1, c_2 \in [0, 1]$ appropriate and all $i, j \in \{1, ..., n\}$, this yields condition (A.3) in Theorem A.6 via

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

if $C^{-1}(v_j + c_1 \overline{w}_{ij}\mu u), C^{-1}(\check{v}_i + c_2 u) \in \Theta_\varkappa$. The function C is strictly monotonic increasing, and the expectation satisfies $\check{v}_i \in [\min_j v_j, \max_j v_j]$. It holds by assumption that $C^{-1}(v_i) \in \Theta_\varkappa$ for all $i \in \{1, ..., n\}$ and, as a consequence, $C^{-1}(\check{v}_i) \in \Theta_\varkappa$. Therefore, it suffices to ensure that $C^{-1}(v_j + u) \in \Theta_\varkappa$ for all $j \in \{1, ..., n\}$ and $u \in \mathcal{U}(\overline{W}_i, z)$. The assertion of Theorem A.6 remains valid if condition (A.3) is only satisfied for $u \in \mathcal{U}(\overline{W}_i, z)$ with $u := \overline{v}_i - \check{v}_i$. Hence, we restrict our analysis to the favorable realizations, where $C^{-1}(v_j + \overline{v}_i - \check{v}_i) \in \Theta_\varkappa$ for all $i, j \in \{1, ..., n\}$ and some most favorable subset $\Theta_\varkappa \subseteq \Theta$. The probability of the complementary set of realizations is bounded by the probability $\check{p}_\varkappa$ in Notation A.9, and we get by Theorem A.6 that

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

which leads to the assertion. $\qquad\square$

REFERENCES

S.M.A. Becker and P. Mathé. A different perspective on the Propagation-Separation Approach. *Electronic Journal of Statistics*, 7:2702–2736, 2013. doi: 10.1214/13-EJS860.

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

A. Buades, B. Coll, and J.-M. Morel. A review of image denoising algorithms, with a new one. *Multiscale Modeling & Simulation*, 4(2):490–530, 2005.

D.V. Divine, J. Polzehl, and F. Godtliebsen. A propagation-separation approach to estimate the autocorrelation in a time-series. *Nonlinear processes in geophysics*, 15(4):591–599, 2008.

J. Fan and I. Gijbels. *Local polynomial modelling and its applications*, volume 66 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1996. ISBN 0-412-98321-4.

O.V. Lepskiĭ. A problem of adaptive estimation in Gaussian white noise. *Akademiya Nauk SSSR. Teoriya Veroyatnosteĭ i ee Primeneniya*, 35(3):459–470, 1990. URL `http://dx.doi.org/10.1137/1135065`.

Y. Li, H. Zhu, D. Shen, W. Lin, J.H. Gilmore, and J.G. Ibrahim. Multiscale adaptive regression models for neuroimaging data. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 73(4):559–578, 2011. URL `http://dx.doi.org/10.1111/j.1467-9868.2010.00767.x`.

Y. Li, J.H. Gilmore, J. Wang, M. Styner, W. Lin, and H. Zhu. TwinMARM: two-stage multiscale adaptive regression methods for twin neuroimaging data. *IEEE Transaction on Medical Imaging*, 31(5):1100–1112, 2012. doi: 10.1109/TMI.2012.2185830.

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

J. Polzehl. *aws: Adaptive Weights Smoothing*, 2012. URL `http://cran.r-project.org/package=aws`. **R**-package version 1.9-3.

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

J. Polzehl and V. Spokoiny. Propagation-separation approach for local likelihood estimation. *Probability Theory and Related Fields*, 135:335–362, 2006.

J. Polzehl and V. Spokoiny. Structural adaptive smoothing by propagation-separation methods. In *Handbook of Data Visualization*, Springer Handbooks of Computational Statistics, pages 471–492. Springer-Verlag, Berlin Heidelberg, 2008. ISBN 978-3-540-33036-3. URL `http://dx.doi.org/10.1007/978-3-540-33037-0_19`.

J. S. Simonoff. *Smoothing methods in statistics*. Springer Series in Statistics. Springer-Verlag, New York, 1996. ISBN 0-387-94716-7. URL `http://dx.doi.org/10.1007/978-1-4612-4026-6`.

K. Tabelow, J. Polzehl, V. Spokoiny, and H.U. Voss. Diffusion tensor imaging: Structural adaptive smoothing. *NeuroImage*, 39(4):1763 – 1773, 2008. doi: 10.1016/j.neuroimage.2007.10.024.