Uniform convergence rates and uniform adaptive estimation in mixtures of regressions

Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultäten
der Philipps-Universität Marburg

vorgelegt von

Heiko Werner
Master of Science
aus Kassel

Marburg, 2019

Erstgutachter und Betreuer: Prof. Dr. Hajo Holzmann
Zweitgutachter: Prof. Dr. Pierre Vandekerkhove
Drittgutachter: Prof. Dr. Laurent Bordes
Eingereicht: 31. Juli 2018
Tag der Disputation: 26. September 2018
Erscheinungsort: Marburg
Hochschulkennziffer: 1180

Acknowledgements

First and foremost, I want to thank my doctoral supervisor Hajo Holzmann for his excellent supervision, many fruitful discussions and valuable suggestions, him helping me improve my understanding of the field significantly and support in many other regards. I am very grateful to Pierre Vandekerkhove who agreed to make the second assessment of this thesis and to travel to Germany in order to attend the thesis defense. I would also like to thank Laura Behner and Viktor Bengs who have read parts of this thesis thoroughly and provided valuable comments. Great thanks to the current and former members of our working group for the always pleasant working atmosphere and making the last few years very enjoyable. I also want to thank the members of the math department with whom I've spent a lot of time working and chatting for the pleasant times. I thank my family and friends for the regular diversion and great times overall. Special thanks to Laura for always being there for me and particularly for supporting me in the final stages of writing this thesis.

Contents

1 Introduction

Finite mixture models

A commonly used tool to model hidden heterogeneity is given by finite mixture distributions, i.e. for a fixed number of components $m \in \mathbb{N}$, the mixture of densities $f_c : \mathbb{R} \to [0, \infty)$, $c = 1, \ldots, m$ is given by

$$\sum_{c=1}^{m} \pi_c f_c \ , \tag{1.0.1}$$

where $\pi_1, \ldots, \pi_m \geq 0$, $\sum_{c=1}^{m} \pi_i = 1$ are mixing parameters. Suppose there are m different latent subpopulations present and observations from each subpopulation are distributed according to some density f_c. Then a mixture of the form (1.0.1) is a natural way to model the data.

Early work on finite mixture distributions goes back to Newcomb (1886) who recognized that the presupposition "that there must always be some one 'most probable value' of a quantity determined by observations, lacks generality" and therefore proposed the idea of "modified curves of probability" that he formalized by introducing mixture distributions. Pearson (1894) used a mixture of two normal distributions in order to describe evolutionary phenomena and was subsequently able to detect clusters within populations of crabs. Parameter estimation was done with the method of moments.

In most models, the component densities f_c are considered to come from a parametric family, i.e. there are sets $\Theta_c \subset \mathbb{R}^{d_c}$ and families of probability measures $\{\mathbb{P}^c_\theta : \theta \in \Theta_c\}$ so that $f_c \sim \mathbb{P}^c_{\theta^c_*}$ for some $\theta^c_* \in \Theta_c$, $c = 1, \ldots, m$. The most common example is given by mixtures of normal distributions, although mixtures of most well-known distributions have been studied extensively, for a broad overview cf. Titterington et al. (1985) or McLachlan and Peel (2004). Depending on the statistical application however, the usage of parametric families can be too restrictive, particularly because many estimation methods are sensitive to violation of distributional model assumptions.

When dropping the parametric assumption, identifiability of the model becomes an intricate issue. The less constraints one imposes on the component densities, the more often the model becomes non-identifiable, making consistent estimation impossible. Symmetry turns out to be a viable constraint for retaining identifiability. Bordes et al. (2006b) and Butucea and Vandekerkhove (2014) considered two-component mixture models in which both components are given by the translation of one single unknown zero-symmetric density. Bordes et al. (2006a), Bordes and Vandekerkhove (2010) and Hohmann and

Holzmann (2013a) considered two-component mixture models in which one component density is known a priori and the other is an unknown translated zero-symmetric density. All those models are identifiable and estimation methods were established. Hunter et al. (2007) even gave identifiability conditions for mixtures of three symmetric components belonging to the same location family and additionally conjectured identifiability for mixtures with more components. Multivariate two-component mixtures with independent marginal components were considered by Hall and Zhou (2003) who proved that the model is non-parametrically identifiable for three-variate mixtures. Two-variate mixtures are only identifiable up to a parametric family of two parameters. $\sqrt{n}$-consistent estimators for the univariate marginal distribution functions of the components and the mixing proportions were introduced.

Finite mixtures of regressions

The concept of finite mixtures of regressions (FMR) combines finite mixture distributions and regression models. Regression models are used to study the relationship of explanatory variables or covariates X on response variables Y. Assume one draws observations from a population (Y, X) taking values in $\mathbb{R} \times \mathbb{R}^d$ having the functional relationship

$$Y = g(X) + \varepsilon \,,$$

where ε is an error generally assumed to fulfil $\mathbb{E}[\varepsilon|X] = 0$ and g is the regression function. The most prominent regression type is the linear regression where $g(x) = \beta_1^T x + \beta_0$ is a linear function.

In the case of FMR, one assumes there are multiple explanatory relations and a latent random variable choosing the explanatory relationship for every single observation. To be more precise, let there be a functional relationship of covariates and response in the shape of

$$Y = \sum_{c=1}^{m} W_c \big[g_c(X) + \varepsilon_c \big] \,,$$

where $(W_1, \ldots, W_m)|X = x \sim \mathrm{Mult}\big(1; (\pi_1(x), \ldots, \pi_m(x))\big)$ with $\sum_{c=1}^{m} \pi_c = 1$, the mixing functions π_c may be predictor-dependent and conditionally on X, the ε_c and W_c are independent as well as again $\mathbb{E}[\varepsilon_c|X] = 0$, $c = 1, \ldots, m$. Here, for each latent subpopulation, g_c is the regression function and describes the regression relationship within the subpopulation. A slight variation of this model are mixtures of location scale regressions. They typically have the form

$$Y = \sum_{c=1}^{m} W_c \big[\mu_c(X) + \sigma_c(X)\varepsilon_c \big] \,,$$

where the locations μ_c and scales σ_c may be predictor-dependent, cf. Huang et al. (2013) or Butucea et al. (2017).

Early work on a two-component mixture of linear regressions model with normally distributed errors was done by Quandt (1958). A formalization in the context of FMR was discussed by Quandt (1972) who introduced the idea that switching occurs according to some random variable that determines from which regime observations are drawn. Estimation in FMR models was first considered by Quandt and Ramsey (1978) who examined a model with normally distributed errors and estimated parameters using the method of moments. Extensions of this model were considered by Jordan and Xu (1995), Young and Hunter (2010) and Huang and Yao (2012) resulting in the most general model of mixtures of normal regressions by Huang et al. (2013). In their model, any finite number of components is allowed and mixing, location and scaling functions are predictor-dependent and modelled non-parametrically. A local log-likelihood estimator and an EM algorithm were proposed and asymptotic normality of the estimator was derived.

There are also considerations of FMR in which the error distributions need not be normal. Hunter and Young (2012) considered mixtures of linear regressions with any finite number of components, each of which is supposed to have the same error distribution. The authors were able to prove identifiability by making use of the additional information carried by the covariates. Vandekerkhove (2013) examined a two-component mixture of linear regressions with zero-symmetric error distributions. Kasahara and Shimotsu (2009) and Hohmann and Holzmann (2013b) considered regression on the model of Hall and Zhou (2003). In a very recent paper, Butucea et al. (2017) considered a FMR model with two components in which proportions and locations are predictor-dependent and both errors are distributed according to a single unknown zero-symmetric distribution. They proposed a local contrast M-estimator for the parameter functions that turns out to be asymptotically normal under reasonable conditions. Furthermore, they proposed a kernel type estimator for the error distribution. Its asymptotic properties are still under consideration.

Organization of this thesis

In Chapter 2, we review some existing FMR models and methods from the literature. We further briefly discuss the models that are later thoroughly examined in Chapter 4. In Chapter 3, we give a general set of conditions under which local M-estimators have nonparametric uniform convergence rates and further give a uniform adaptive estimation procedure. Both are applicable to a variety of models, e.g. the models in Butucea et al. (2017) or Huang et al. (2013). In Chapter 4, we apply the methods to the model in Huang et al. (2013) and a regression model based on an alteration of the model without covariates in Bordes and Vandekerkhove (2010). Chapter 5 accumulates most of the proofs for Chapters 3 and 4. Furthermore, in Chapter 6, we conduct a simulation study of the local log-likelihood estimator for the model in Huang et al. (2013) that displays the finite sample properties of the supremum errors of the estimators.

2 Preliminaries

In this chapter, we will review some literature and theory relevant to this thesis. We first give classical asymptotic results on M-estimators as described by van der Vaart (2000) and van der Vaart and Wellner (1996) and subsequently briefly discuss the concept of local M-estimation. Furthermore, we discuss literature on mixture of regressions models, briefly recall function classes broadly used in non-parametric estimation and shortly discuss kernel density estimation.

2.1 M-estimation

A well-studied concept for parameter estimation is M-estimation. Suppose one draws observations $Y_1, Y_2, \dots$ from a distribution $\mathbb{P}_{\theta_*}$ coming from an identifiable statistical model $(\mathbb{P}_\theta)_{\theta\in\Theta}$ depending on some parameter set Θ. Further suppose one can find a contrast function, that is, a function $M : \Theta \to \mathbb{R}$ that has a unique maximum at the true parameter θ_*. Moreover, assume one can estimate M by a random function $M_n(\cdot) = M_n(\cdot; Y_1, \dots, Y_n) : \Theta \to \mathbb{R}$. Then a natural approach to estimate the model parameter θ_* is to maximize the random function M_n over the parameter set Θ if possible and to use the point of maximum as an estimator.

The first estimator of this type, the maximum likelihood estimator, was introduced by Fisher (1922) and has been studied extensively since, e.g. Huber et al. (1967). The log-likelihood estimator is a commonly used variation of this. Assume every $\mathbb{P}_\theta$ possesses a density p_θ. Given i.i.d. observations $Y_1, \dots, Y_n$, one chooses the parameter as estimate that is most likely to produce the observations at hand, if existent, i.e.

$$\hat{\theta}_n = \operatorname*{argmax}_{\theta\in\Theta} M_n(\theta) \ , \quad M_n(\theta) = \frac{1}{n}\sum_{j=1}^{n} \log\left(p_\theta(Y_j)\right) \ . \tag{2.1.1}$$

Other examples are least square estimators or estimators that use model specific properties like symmetry to construct contrast functions, cf. Hall and Zhou (2003), Bordes et al. (2006b), Hunter et al. (2007), Bordes and Vandekerkhove (2010), Hohmann and Holzmann (2013a) as well as Butucea and Vandekerkhove (2014).

A fundamental criterion for the quality of an estimator is the concept of consistency. We call a sequence of estimators $\hat{\theta}_n$ consistent for (every possible model parameter) $\theta_* \in \Theta$ if

$$\|\hat{\theta}_n - \theta_*\| = o_{\mathbb{P}}(1) \ ,$$

where the estimators $\hat{\theta}_n$ depend on $Y_1, \dots, Y_n$, which are distributed according to $\mathbb{P}_{\theta_*}$.

In the context of M-estimators, consistency is often achieved by the contrast function M taking a unique and well-separated global maximum at the true parameter and the random functions M_n estimating the contrast M uniformly consistently, so that points of global maxima of M_n need to converge to points of global maxima of M. Uniform consistency of M_n often turns out easy to prove when the parameter space Θ is compact. This result is summarized in the following theorem, cf. van der Vaart (2000, Theorem 5.7).

Theorem 2.1.1 (van der Vaart (2000))**.** *Let d be a metric on Θ, M_n be random functions and let M be a fixed function of θ such that for every $\varepsilon > 0$*

$$\sup_{\theta \in \Theta} \left| M_n(\theta) - M(\theta) \right| = o_{\mathbb{P}}(1) \ , \qquad \sup_{\theta \in \Theta : d(\theta, \theta_*) \geq \varepsilon} M(\theta) < M(\theta_*) \ .$$

Then any sequence of estimators $\hat{\theta}_n$ with $M_n(\hat{\theta}_n) \geq M_n(\theta_*) - o_{\mathbb{P}}(1)$ converges in probability to θ_*.

After consistency of an estimator $\hat{\theta}_n$ is established, the question arises at which rate the estimator converges, that is, one wishes to identify a deterministic sequence $r_n \to \infty$ so that

$$r_n \|\hat{\theta}_n - \theta_*\| = O_{\mathbb{P}}(1)$$

for every possible model parameter $\theta_* \in \Theta$. The sequence r_n is then an upper bound of the convergence rate.

A general convergence rate result can be found in van der Vaart and Wellner (1996, Theorem 3.2.5). Suppose that $\theta_* \in \Theta \subset \mathbb{R}^m$. Under suitable differentiability conditions, the gradient of M at θ_* has to be zero and its Hessian matrix $V(\theta) = \partial^2_{\theta^2} M(\theta)$ evaluated at θ_* is always negative semidefinite. If one assumes that it is in fact negative definite, a Taylor expansion around θ_* gives for some $\tilde{\theta} \in [\theta, \theta_*]$ that

$$M(\theta) - M(\theta_*) = (\theta - \theta_*)^T \underbrace{\partial_\theta M(\theta_*)}_{=0} \ + \ \frac{1}{2} (\theta - \theta_*)^T V(\tilde{\theta}) (\theta - \theta_*) \ .$$

By additional regularity conditions, such like Lipschitz continuity of the Hessian matrix, one can often achieve that the distance of $M(\theta)$ and $M(\theta_*)$ behaves like the squared distance of θ and θ_*. Combining this with a condition on the expectation of the continuity modulus of $M_n - M$ can give upper bounds on the convergence rate of the maximizers of M_n. The latter condition often happens to be Lipschitz continuity with Lipschitz-constant decreasing at rate r_n^{-1}.

Theorem 2.1.2 (van der Vaart and Wellner (1996))**.** *Let M_n be stochastic processes indexed by a semimetric space (Θ, d) and $M : \Theta \to \mathbb{R}$ a deterministic function, such that for every θ in a neighbourhood of θ_*,*

$$M(\theta) - M(\theta_*) \lesssim -d^2(\theta, \theta_*) \ .$$

Suppose that, for every n and sufficiently small δ, the centred process $M_n - M$ satisfies

$$\mathbb{E}^*\left[\sup_{d(\theta,\theta_*)<\delta} \left|(M_n - M)(\theta) - (M_n - M)(\theta_*)\right|\right] \lesssim \frac{\phi(\delta)}{\sqrt{n}} ,$$

for functions ϕ_n such that $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$ (not depending on n). Let

$$r_n^2 \phi_n\left(\frac{1}{r_n}\right) \leq \sqrt{n} , \qquad \text{for every } n.$$

If the sequence $\hat{\theta}_n$ satisfies $M_n(\hat{\theta}_n) \geq M_n(\theta_*) - O_{\mathbb{P}}(r_n^{-2})$ and converges in outer probability to θ_*, then $r_n d(\hat{\theta}_n, \theta_*) = O_{\mathbb{P}^*}(1)$.

Note that we use outer expectation $\mathbb{E}^*$ and outer probability $\mathbb{P}^*$ in case there are problems with measurability. This theorem may be generalized in order to allow for convergence rates differing from $\sqrt{n}$ and be applied to models with covariates, cf. Theorem 3.1.3.

2.2 Local M-estimation

When one aims to estimate the influence of covariates X on a response Y, M-estimation is typically not directly applicable because one cannot draw observations from the conditional distribution $Y|X$. However, one can often use local M-estimation, which uses a localization strategy like weighting observations with the help of a kernel function.

To be precise, suppose one draws i.i.d. observations $(Y_1, X_1), \ldots, (Y_n, X_n)$ from a population (Y, X), where X is supported on $I \subset \mathbb{R}^d$ and Y is real-valued. For every $x \in I$, one suspects that the conditional distribution of $Y|X = x$ given by $\mathbb{P}_{\theta_*(x)}$ is determined by some unknown $\theta_*(x) \in \Theta$ where Θ is a parameter set. Then the objective is to estimate the parameter function $\theta_*(\cdot)$.

For now, consider the respective model without covariates, i.e. one has observations $Z_1, Z_2, \ldots$ with distribution $\mathbb{P}_{\theta_*}$ and like in Section 2.1, there is a function

$$\widetilde{M}_n(\cdot; Z_1, \ldots, Z_n) : \Theta \to \mathbb{R}$$

estimating a function $\widetilde{M} : \Theta \to \mathbb{R}$ that has a unique maximum at the true parameter θ_*. Further assume that $\widetilde{M}_n$ is a linear estimator in the sense that it is of the shape

$$\widetilde{M}_n(\theta; Z_1, \ldots, Z_n) = \frac{1}{n} \sum_{j=1}^{n} m_\theta(Z_j)$$

for some function m_θ, like a log-likelihood estimator defined in (2.1.1).

Then, returning to the model with covariates, the localization strategy can be applied by using a kernel function $K_h = K(\cdot/h)/h^d$, cf. Section 2.5. That is, define the local M-estimator as

$$M_n(\theta, x; h) := M_n\big(\theta, x; h; (Y_1, X_1), \ldots, (Y_1, X_1)\big) := \frac{1}{n}\sum_{j=1}^{n} m_\theta(Y_j) K_h(X_j - x) \; .$$

The most prominent estimator of this kind is the local log-likelihood estimator that was first described in a broad sense by Fan et al. (1998). Assume that all conditional distributions $\mathbb{P}_\theta$ possess densities p_θ, $\theta \in \Theta$. Given i.i.d. observations $(Y_1, X_1), \ldots, (Y_n, X_n)$, one chooses for every x the parameter that maximizes the local log-likelihood function

$$\hat{\theta}_n(x) = \operatorname*{argmax}_{\theta\in\Theta} M_n(\theta, x; h) = \operatorname*{argmax}_{\theta\in\Theta} \frac{1}{n}\sum_{j=1}^{n} \log\big(p_\theta(Y_j)\big) K_h(X_j - x) \; . \tag{2.2.1}$$

2.3 Mixture of regressions models

In this section, we outline mixture of regressions models from the literature that are related to the models examined in the main part of this thesis.

Let us start by considering the mixture of linear regressions models with normally distributed errors. Quandt (1972) considered such a switching linear regressions model. Assume one observes the $n \times (k+1)$ random matrix (Y, X), where the vector Y consists of response variables and the rows of X are k independent explanatory variables so that there is a subset $I \subset \{1, \ldots, n\}$ so that

$$\begin{aligned} Y_i &= X_i\beta_1 + \varepsilon_{1,i} \; , \; i \in I \; , \\ Y_i &= X_i\beta_2 + \varepsilon_{2,i} \; , \; i \in I^c \; , \end{aligned}$$

where X_i is the i-th row vector, $\varepsilon_{k,i} \sim \mathcal{N}(0, \sigma_k^2)$ and $\beta_i = (\beta_{i1}, \ldots, \beta_{ik})^T$, $i = 1, 2$ are parameters of interest so that $(\beta_1, \sigma_1^2) \neq (\beta_2, \sigma_2^2)$. Quandt (1972) was the first to model this by postulating that "there is an unknown probability λ that nature will choose Regime 1 for generating observations and a probability $1 - \lambda$ that it will choose Regime 2". In a more rigorous fashion, one could postulate that there is a random variable W_i that, conditionally on the covariates, is independent of the $\varepsilon_{k,i}$ and has conditional distribution $\mathrm{Ber}(\lambda)$ so that

$$Y_i = W_i\big(X_i\beta_1 + \varepsilon_{1,i}\big) + (1 - W_i)\big(X_i\beta_2 + \varepsilon_{2,i}\big) \; , \quad i \in \{1, \ldots, n\} \; .$$

Quandt (1972) directly concluded that, conditionally on $X_i = (x_{1i}, \ldots, x_{ki})$, the response Y_i has a conditional mixture of two normal densities, i.e.

$$f\big(y_i|(x_{1i}, \ldots, x_{ki}); \theta\big) = \lambda\phi\big(y_i\big|x_i^T\beta_1, \sigma_1^2\big) + (1 - \lambda)\phi\big(y_i\big|x_i^T\beta_2, \sigma_2^2\big) \; , \tag{2.3.1}$$

where $\theta = (\beta_1, \beta_2, \sigma_1, \sigma_2, \lambda)^T$, giving the log-likelihood function

$$\theta \mapsto \sum_{i=1}^{n} \log\big(f(Y_i|X_i; \theta)\big) \; ,$$

which is to be maximized under the constraints $\sigma_1, \sigma_2 > 0$, $0 \leq \lambda \leq 1$. The estimation method was tested empirically. In a later paper, Quandt and Ramsey (1978) gave an estimation procedure based on the moment generating function.

This model is commonly applied in economics, see for example, DeSarbo et al. (1992), Ramaswamy et al. (1993) or Helsen et al. (1993).

Generalizations

There are several ways to generalize the model given by the conditional model density (2.3.1). We distinguish between two possibilities, namely (1) retaining normal errors and (2) dropping the assumption of normally distributed errors. In both cases, natural ways to generalize the model are

- allowing a higher or even unspecified finite number of components;
- allowing the mixing proportions and the scaling parameters to depend on the covariates as well;
- allowing for different regression function structures like polynomials or even dropping parametric shape constraints entirely by letting parameter functions come from non-parametric function classes.

2.3.1 Models with normally distributed errors

Goldfeld and Quandt (1973) proposed a Hidden-Markov approach in order to model regime switches and proposed likelihood estimators in the framework of Quandt (1972). Bayesian approaches for models with more than two components were given by Viele and Tong (2002) and Hurn et al. (2003).

Young and Hunter (2010) extended the model of mixtures of linear regressions with normal errors in the sense that mixing proportions are allowed to be predictor-dependent. That is, in the framework of mixtures of linear normal regressions with $m \geq 2$ components, where the conditional densities are given by

$$f(y|x;\theta) = \sum_{c=1}^{m} \pi_c \phi\big(y\big|x^T\beta_c, \sigma_c^2\big) \ ,$$

the mixing proportions π_c are assumed to depend non-parametrically on the covariate values. Let $\pi_c : I \to (0,1)$, $c = 1, \ldots, m$, $\sum_{c=1}^{m} \pi_c = 1$ be functions giving the mixing proportions for fixed covariate values x. Then, the conditional model densities are given by

$$f(y|x;\theta) = \sum_{c=1}^{m} \pi_c(x) \phi\big(y\big|x^T\beta_c, \sigma_c^2\big) \ .$$

Young and Hunter (2010) gave an iterative global/local estimation (IGLE) algorithm for estimating the mixing proportion functions $\pi_c(\cdot)$ along with the other global parameters β_c, σ_c.

Huang and Yao (2012) proposed a slight variation of the model by Young and Hunter (2010). The mixing proportions are assumed to depend on another explanatory variable Z. The log-likelihood is given by

$$\sum_{i=1}^{n} \log\left(\sum_{c=1}^{m} \pi_c(Z_i)\phi\big(Y_i\big|x_i^T\beta_c, \sigma_c^2\big)\right), \tag{2.3.2}$$

so that the conditional model densities are given by

$$f(y|x,z;\theta) = \sum_{c=1}^{m} \pi_c(z)\phi\big(y\big|x^T\beta_c, \sigma_c^2\big)\ .$$

Huang and Yao (2012) proved identifiability of the model when the proportion functions are continuous, the support of the covariate X contains an open subset and the support of the other covariate Z has no isolated points.

In order to estimate the model parameters, the authors first proposed a local log-likelihood estimation approach in the sense of (2.2.1) by localizing around z for pre-estimation. That is, for fixed z, maximize the function

$$\sum_{i=1}^{n} \log\left(\sum_{c=1}^{m} \pi_c\phi\big(y\big|x_i^T\beta_c, \sigma_c^2\big)\right) K_h(Z_i - z)\ . \tag{2.3.3}$$

The global parameters β_c, σ_c, $c = 1, \ldots, m$ are estimated with this procedure as well but cannot have parametric convergence rate because they are estimated locally. Plugging in the non-parametric estimates for the mixing proportions into the log-likelihood function (2.3.2) and maximizing ensures $\sqrt{n}$-consistency of the parametric part of the estimator. Plugging in the global parametric estimates into (2.3.3) and maximizing improves the estimate of the proportion functions. This estimator is pointwise asymptotically normal with non-parametric convergence rate. Furthermore, they gave an EM-type algorithm for practical approximation of the maxima.

Huang et al. (2013) extended the model by letting all parameters be predictor-dependent and modelling them non-parametrically by sufficiently smooth functions. That is, for any number of components $m \geq 2$, any compact covariate support $I \subset \mathbb{R}^d$ containing an open subset and Hölder-α-smooth functions $\pi_c : I \to (0,1)$, $\sum_{c=1}^{m} \pi_c = 1$, $\sigma_c : I \to (0,\infty)$, $\mu_c : I \to \mathbb{R}$, the conditional model densities are given by

$$f(y|x;\theta(\cdot)) = \sum_{c=1}^{m} \pi_c(x)\phi\big(y\big|\mu_c(x), \sigma_c(x)^2\big)\ ,$$

where

$$\theta(\cdot) = \big(\pi_1(\cdot), \dots, \pi_{m-1}(\cdot), \mu_1(\cdot), \dots, \mu_m(\cdot), \sigma_1(\cdot), \dots, \sigma_m(\cdot)\big)^T$$

is the model parameter function.

Huang et al. (2013) proved non-parametric identifiability under the assumption that the covariates are compactly supported in $\mathbb{R}$, the proportion functions are positive and continuous, the location and scaling functions are differentiable and pairs of those functions are transversal, i.e. for any components $c \neq c'$, one has

$$\Big(\big\| \big(\mu_c(x), \sigma_c(x)\big) - \big(\mu_{c'}(x), \sigma_{c'}(x)\big) \big\| + \big\| \big(\partial\mu_c(x), \partial\sigma_c(x)\big) - \big(\partial\mu_{c'}(x), \partial\sigma_{c'}(x)\big) \big\| \Big) \neq 0 \, .$$

This means that the pairs of parameter functions may only intersect nontangentially.

The authors proposed a local log-likelihood estimation method in the sense of (2.2.1). By assuming amongst other things that the parameter functions are twice continuously differentiable, the authors proved that the estimator is pointwise consistent and asymptotically normal with non-parametric convergence rate. One should note that the authors provided no proof of positive definiteness of the Fisher information, which, in fact, cannot be true at intersection points of at least two pairs of parameter curves $\big(\mu_c(\cdot), \sigma_c(\cdot)\big)$ because the estimation problem is then locally overparametrized. Hence, the asymptotic properties of the estimator are only valid at covariate values x at which no parameter curves intersect.

This model will be discussed later in this thesis, cf. Section 4.1. We will give conditions under which the model with multivariate covariates is non-parametrically identifiable. Furthermore, we will show that for Hölder-α-smooth parameter functions as defined in Definition 2.4.1, the local log-likelihood estimator is uniformly consistent over compact sets within the interior of the compact covariate support $I \subset \mathbb{R}^d$ and has non-parametric L^∞ convergence rate. We will briefly discuss the benefits of uniform consistency in the context of curve estimation and the relabeling problem. Moreover, we will give an estimation procedure based on Lepskii (1992) that is adaptive with respect to the smoothness α.

2.3.2 Models with unspecified errors

When dropping the assumption of normally distributed errors and instead assuming the error distributions to be unknown, identifiability typically becomes harder to prove. Let us first consider mixtures of linear regressions in which the errors are unspecified.

Hunter and Young (2012) considered a finite mixture of linear regressions with unspecified errors, i.e. the conditional model density is given by

$$f(y|x;\theta) = \sum_{c=1}^{m} \pi_c f^*(y - x^T\beta_c) \, ,$$

where $x, \beta_c \in \mathbb{R}^d$, $\theta = (\pi_1, \ldots, \pi_{m-1}, \beta_1^T, \ldots, \beta_m^T, f^*)^T$ is the unknown parameter of interest. The authors provided an identifiability result that only requires the support of the covariates to contain an open subset and no two different regression hyperplanes to be parallel, i.e. $\beta_j \neq \beta_k$ for $j \neq k$. One should note that the model without covariates, i.e. mixtures

$$f(y) = \sum_{c=1}^{m} \pi_c f^*(y - \mu_c)$$

are in general not identifiable. Additional information carried by the covariates enables identifiability. The authors also provided a semi-parametric EM-type algorithm that estimates both the parametric part of the parameter and the non-parametric error density.

Bordes et al. (2013) considered a two-component mixture of linear regressions in which one component error distribution is known, call its density $\bar{f}$. They formulated an equivalent estimation problem in which the conditional model density is given by

$$f(y|x;\theta) = (1-p)\bar{f}(y) + pf^*(y - \alpha - \beta x) \ ,$$

where $\theta = (p, \alpha, \beta, f^*)^T$. They proposed an asymptotically normal estimator for the parametric part (p, α, β) and an estimator for the unknown component distribution function of f^* that is derived from the empirical distribution function of the observations $Y_i - \alpha - \beta X_i$. This estimator was shown to be asymptotically Gaussian.

Note that in the models of both Hunter and Young (2012) and Bordes et al. (2013), the shape of the component densities are not constrained, particularly, they need not be symmetric. However, the assumption of symmetry is a viable restriction often ensuring identifiability in mixtures of regressions with more complex parameter structures and in mixture models without covariates. Let us consider models of the latter kind before discussing regression on one of them.

Bordes et al. (2006b) considered a two-component mixture model in which both components come from the same location family of a symmetric density, i.e. the model density is given by

$$f(y;\theta) = pf^*(y - \mu_1) + (1-p)f^*(y - \mu_2) \ , \tag{2.3.4}$$

where the parameter of interest is given by $\theta = (p, \mu_1, \mu_2, f^*)^T$. They distinguished between two cases. In the first case, location parameters are a priori known, in the other case they are model parameters to be estimated. For both cases, they presented estimators both for the parametric part (p, μ_1, μ_2) and the non-parametric part f^* and proved asymptotic properties. Additionally, the authors constructed an M-estimator for the parametric part by representing the component distribution function F^* as a function of the mixture distribution function F and the parameter. The non-parametric part is estimated by inverting this relationship and directly using the empirical distribution function of the accordingly transformed observations in order to estimate F^* or kernel

density estimators of the transformed data in order to estimate f^*.

Hunter et al. (2007) considered two- and three-component mixtures in which all component densities come from a location family of one zero symmetric density f^*. Hence, the model density is given by

$$f(y;\theta) = \sum_{c=1}^{m} \pi_c f^*(y - \mu_c) \ ,$$

where $m \in \{2, 3\}$, $\sum_{c=1}^{m} \pi_c = 1$ and the parameter of interest is given by

$$\theta = (\pi_1, \ldots, \pi_{m-1}, \mu_1, \ldots, \mu_m, f^*)^T \ .$$

The authors provided an identifiability result for these models and further conjectured that identifiability can be achieved for models with any number of components but refrained from tackling the proof because of increasing complexity for higher number of components.

Bordes et al. (2006a) considered a two-component mixture model in which one component density is a priori known and zero-symmetric and the other one comes from a location family of an unknown zero-symmetric density. That is, for a known density $\bar{f}$, the model density is given by

$$f(y;\theta) = (1-p)\bar{f}(y) + pf^*(y - \mu) \ ,$$

where the parameter of interest is given by $\theta = (p, \mu, f^*)^T$. The authors showed that the model is identifiable basically if either the density $\bar{f}$ is compactly supported or if it is positive and is dominated by f^* in the tails.

Based on symmetry and the relation

$$f^*(y) = p^{-1}\big(f(y + \mu;\theta) - (1-p)\bar{f}(y + \mu)\big) \ , \tag{2.3.5}$$

Bordes et al. (2006a) were able to build a contrast for the location μ yielding a plug-in estimator for p. The unknown component density f^* can be estimated by estimating the right-hand side of (2.3.5) with a kernel density estimator when plugging in the estimators of the parametric part. The corresponding distribution function can be estimated once again by the empirical distribution function of the right-hand side of (2.3.5). In addition, the authors proved strong consistency of the parametric part, uniform consistency of the empirical distribution function estimator as well as L^1 convergence of the kernel density estimator.

Bordes and Vandekerkhove (2010) provided estimators for the model in Bordes et al. (2006a) that are asymptotically normal in the sense that the deviations of the estimators from the true parameters have convergence rate $\sqrt{n}$ and when scaled appropriately

converge in distribution to a Gaussian process.

Hohmann and Holzmann (2013a) extended the model in Bordes et al. (2006a) by introducing a location parameter to the component with the known density. The authors provided results on identifiability based on symmetry of the component distributions and the assumption that the Fourier transforms of the component distribution functions are distinguishable in the tails. Additionally, Hohmann and Holzmann (2013a) introduced asymptotically normal estimators both for the parametric part and the component distribution function.

Butucea and Vandekerkhove (2014) once again considered the model in Bordes et al. (2006b), cf. (2.3.4). For the parametric part, the authors proposed a smooth U-statistic estimator that is based on the characteristic functions of the component density f^*, which needs to be real-valued as f^* is symmetric. The estimator is asymptotically normal with parametric convergence rate $\sqrt{n}$. The non-parametric part is estimated with a kernel estimator. For Sobolev functions, the estimator was shown to have non-parametric convergence rate, which is in fact the minimax rate for the model.

Let us discuss the model of Butucea et al. (2017) who considered regression on the model in Butucea and Vandekerkhove (2014) in more detail. Suppose one draws observations from a population (Y, X), where X is a $\mathbb{R}^d$-valued explanatory variable with density ℓ for the scalar response variable Y. Furthermore, assume the relation

$$Y = W\big(a(X) + \varepsilon_1\big) + (1 - W)\big(b(X) + \varepsilon_2\big) \,,$$

where conditionally on $X = x$, W has distribution $\mathrm{Ber}(p(x))$ for some mixing function $p : \mathbb{R}^d \to (0, 1)$ and is independent of ε_1 and ε_2, which have common zero-symmetric conditional density f_x. Additionally, $a, b : \mathbb{R}^d \to \mathbb{R}$ are location functions to be estimated along with the mixing function p and the conditional densities f_x.

For every point in the interior of the covariates' support $x \in \mathrm{supp}(\ell)^\circ$, the conditional density of $Y|X = x$ is given by

$$f\big(y|x; \theta(\cdot)\big) = p(x) f_x\big(y - a(x)\big) + \big(1 - p(x)\big) f_x\big(y - b(x)\big) \,.$$

Due to the relabeling problem, the parameters $(p(x), a(x), b(x))$ and $(1-p(x), b(x), a(x))$ yield the same mixture density. In order to deal with this label switching problem for a fixed x, one can allow $a(x)$ and $b(x)$ to coincide only if one restricts the mixing parameter to a compact subset of either $(0, 1/2)$ or $(1/2, 1)$; or one can demand the location parameters to be ordered, i.e. $a(x) < b(x)$ or $b(x) < a(x)$.

There are also strategies to overcome label switching when examining global identifiability, that is identifiability of the parameter functions on $\mathrm{supp}(\ell)$. Butucea et al. (2017) were able to give identifiability of the parameter curves for univariate covariates by imposing transversality constraints on the location functions. In detail, they required the

location functions to be differentiable and to intersect nowhere tangentially, i.e.

$$\|a(x) - b(x)\| + \|\partial a(x) - \partial b(x)\| \neq 0 \ , \quad x \in \operatorname{supp}(\ell) \ ,$$

as well as the existence of an x_0 so that $a(x_0) < b(x_0)$ and the mixing function p to be continuous. This allows for intersections of the location functions, meaning local non-identifiability. At a point of intersection x, however, the labels are identifiable by the labels in the neighbourhood of x. That is because switching labels at x would then make the parameter curve non-differentiable due to transversality.

The estimation procedure proposed by Butucea et al. (2017) was adopted from the work on the model without covariates by Butucea and Vandekerkhove (2014). It is strongly based on symmetry of the component density f_x, which particularly implies that its characteristic function φ_{f_x} is real-valued. By linearity of the Fourier transform, the characteristic function of $f(\cdot|x)$ is then given by

$$\varphi_{f(\cdot|x)}(t) = \Big(p(x)e^{ita(x)} + \big(1-p(x)\big)e^{itb(x)}\Big)\varphi_{f_x}(t) \ , \quad t \in \mathbb{R} \ . \tag{2.3.6}$$

Denote the parameter function by $\theta(\cdot) = \big(p(\cdot), a(\cdot), b(\cdot)\big)^T$. For any $x \in \operatorname{supp}(\ell)$ and an a priori fixed density q, the function

$$S(\theta) = \int \Im\Big(\varphi_{f(\cdot|x)}(t)\Big(pe^{-ita} + (1-p)e^{-itb}\Big)\Big)^2 q(t)\,\mathrm{d}t \cdot \ell^2(x) \ ,$$

is a non-negative contrast function under the identifiability assumptions, i.e. $S(\theta) = 0$ iff θ is the true parameter $\theta(x)$ or a label switched version thereof. When inserting the true parameter, the second factor in the argument of the imaginary part is the complex conjugate of the first factor in (2.3.6), leaving the real-valued characteristic function φ_{f_x} scaled by a positive real number within the imaginary part, hence a zero.

An empirical version of S was constructed by estimating

$$\Im\big(\varphi_{f(\cdot|x)}(t)\big(pe^{-ita} + (1-p)e^{-itb}\big)\big)$$

locally and using these estimates to build a U-statistic. Therefore, a kernel $K : \mathbb{R}^d \to \mathbb{R}$ along with a bandwidth parameter h was being used to propose the estimate

$$Z_k(\theta, t, h) = \Im\bigg(e^{itY_k} \cdot \Big(pe^{-ita} + (1-p)e^{-itb}\Big)\bigg)\frac{1}{h^d}K\bigg(\frac{X_k - x}{h}\bigg)$$

for i.i.d. observations (Y_k, X_k) coming from the population (Y, X). The first factor in the argument of the imaginary part function can be interpreted as the empirical estimate of the characteristic function of $Y|X = x$ so that choosing θ such that the imaginary part is close to zero should translate to θ being close to $\theta(x)$ with probability approaching one under usual assumptions. The M-estimator in the form of a U-statistic type empirical contrast was then defined by

$$S_n(\theta) = -\frac{1}{n(n-1)}\sum_{\substack{j,k=1\\ j\neq k}}^{n}\int Z_k(\theta, t, h)Z_j(\theta, t, h)q(t)\,\mathrm{d}t \ .$$

A minimizer $\hat{\theta}_n$ of S_n was proposed as the estimator for the model parameter.

Butucea et al. (2017) proved that the estimator is asymptotically normal with convergence rate $\sqrt{nh^d}$ when $h \to 0$, $nh^d \to \infty$ as well as

$$h^{2\alpha+d} = o(n^{-1}) \tag{2.3.7}$$

under usual assumptions. Particularly, the parameter functions are assumed to be Hölder-α-smooth with $\alpha > 1$, so that the balanced bandwidth choice $h = n^{-\frac{1}{2\alpha+d}}$ does not fulfil (2.3.7). Hence, asymptotic normality is achieved by undersmoothing. However, in Theorem 4, the authors pointed out that the estimator has convergence rate $n^{\frac{\alpha}{2\alpha+d}}$ when balancing the bandwidth. This convergence rate also turns out to be the lower bound on the L^1 minimax risk in this model.

Two-component mixture of location scale regressions

Let us finally give a variation of the mixture of regressions model in Butucea et al. (2017) that is studied in the main part of this thesis, cf. Section 4.2. Consider regression in the framework of Bordes et al. (2006a) when adding an unknown scaling parameter to the a priori known component. The regression relationship is then given by

$$Y_i = W_i\big(\mu(X_i) + \varepsilon_{1,i}\big) + \big(1 - W_i\big)\sigma(X_i)\varepsilon_{2,i}\ , \quad i \in \mathbb{N}\ ,$$

for sequences of i.i.d. random vectors $(X_i)_{i\in\mathbb{N}}$ having support $I \subset \mathbb{R}^d$, where I is a compact cuboid containing an open subset, $d \geq 1$ and i.i.d. random variables $(Y_i)_{i\in\mathbb{N}}$, $(W_i)_{i\in\mathbb{N}}$, $(\varepsilon_{1,i})_{i\in\mathbb{N}}$ and $(\varepsilon_{2,i})_{i\in\mathbb{N}}$. The explanatory variables X_i and the response variables Y_i are observable, the latent variables W_i and the error variables $\varepsilon_{1,i}$ and $\varepsilon_{2,i}$ are not. The covariates X_i are assumed to have a Lebesgue density $\ell : I \to (0, \infty)$.

The unknown location and scaling functions $\mu : I \to \mathbb{R}$, $\sigma : I \to (0, \infty)$ are functions to be estimated as they partially determine the distributional relation between the explanatory and response variables along with the unknown mixing function $p : I \to (0, 1)$. That is because conditionally on $X_i = x$, the variables W_i are assumed to have a Bernoulli-distribution with parameter $p(x)$, i.e.

$$\mathbb{P}(W_i = 1|X_i = x) = p(x) \quad \text{and} \quad \mathbb{P}(W_i = 0|X_i = x) = 1 - p(x)\ .$$

Let us further assume that conditionally on $X_i = x$, the vectors $\varepsilon_{1,i}$ and $\varepsilon_{2,i}$ have zero-symmetric conditional densities denoted by f_x and $\bar{f}$, respectively, where we assume that $\bar{f}$ is known and f_x is not. If we furthermore have the conditional independence relations

$$\varepsilon_{1,i} \perp\!\!\!\perp W_i|X_i \quad \text{and} \quad \varepsilon_{2,i} \perp\!\!\!\perp W_i|X_i\ ,$$

then, conditional on $X_i = x$, the random variables Y_i have the conditional density

$$f\big(y|x; \vartheta(\cdot)\big) = p(x)f_x\big(y - \mu(x)\big) + \frac{1 - p(x)}{\sigma(x)}\bar{f}\Big(\frac{y}{\sigma(x)}\Big)\ , \quad y \in \mathbb{R}\ ,$$

where the parameter function of interest is given by $\vartheta(\cdot) = \big(p(\cdot), \mu(\cdot), \sigma(\cdot), f_\cdot\big)^T$.

We will prove identifiability of the model by showing that for every $x \in I$, the conditional mixture density $f\big(\,\cdot\,|x; \vartheta(\cdot)\big)$ is identifiable within all mixture densities of the postulated type while making mild assumptions.

We propose an estimation procedure that is strongly based on symmetry of the component densities and the idea of making use of a relationship analogous to (2.3.5) that was used by Bordes et al. (2006a) in their model without covariates. The M-estimator minimizes a smooth U-statistic and is shown to be uniformly consistent over compact sets within the interior of the covariate support. It especially has the typical uniform non-parametric convergence rate $\big(\frac{\log n}{n}\big)^{\frac{\alpha}{2\alpha+d}}$ for Hölder-α-smooth parameter functions. Additionally, an estimation procedure that is adaptive with respect to the smoothness parameter α is proposed which is based on Lepskii (1992).

2.4 Function classes

When decomposing estimation errors of non-parametric estimators into bias and variance terms, upper bounds on the variance term can often be obtained without shape or smoothness constraints on the model parameter function, e.g. variance of kernel density estimators, cf. Tsybakov (2008, Proposition 1.1). However, bounds on the bias term typically rely on smoothness conditions. Classically, the parameter functions are assumed to be continuous and particularly to come from a smoothness class within the set of continuous functions. The predominant smoothness classes are Sobolev- and Hölder classes. In the main part of this thesis, we will work with Hölder classes as defined below.

Definition 2.4.1 (Hölder class)**.** *Let $\alpha, L \in (0, \infty)$ and $I \subset \mathbb{R}^d$ be compact so that $I = \overline{\mathrm{int}(I)}$. The class of Hölder-$\alpha$-smooth functions $H(\alpha, L, U)$ on $I \subset \mathbb{R}^d$ taking values in some set $U \subset \mathbb{R}$ is defined by the set of all functions $\ell : I \to U$ that are continuous, $\lfloor\alpha\rfloor$-times differentiable in $\mathrm{int}(I)$ and fulfil*

$$|\partial^k \ell(x) - \partial^k \ell(y)| \leq L\|x - y\|^{\alpha - \lfloor\alpha\rfloor}\,, \quad |k| = \lfloor\alpha\rfloor, \quad x, y \in \mathrm{int}(I)\,,$$

as well as

$$\|\partial^k \ell\|_\infty \leq L\,, \quad 1 \leq |k| \leq \lfloor\alpha\rfloor\,.$$

In the main part of this thesis, Hölder classes are used to model smoothness of parameter functions. Therefore, let us discuss some relevant properties. First, we briefly discuss Hölder classes as subsets of Banach spaces. For a more general and extensive overview, cf. Driver (2003). Driver (2003) introduces Hölder spaces on open sets and extends functions and their derivatives to the boundary. We will define functions for compact sets I so that $I = \overline{\mathrm{int}(I)}$, which leads to the same objects.

Definition 2.4.2. *For any $\alpha \in (0,1]$ and any compact $I \subset \mathbb{R}^d$ with $I = \overline{\text{int}(I)}$, a continuous function $\ell : I \to \mathbb{R}$ is called Hölder-α-continuous if $[\ell]_\alpha < \infty$, where*

$$[\ell]_\alpha = \sup_{\substack{x,y \in \text{int}(I) \\ x \neq y}} \frac{|\ell(x) - \ell(y)|}{\|x - y\|^\alpha} .$$

The map $[\cdot]_\alpha$ is a seminorm on all continuous functions ℓ with $[\ell]_\alpha < \infty$. It is not a norm because all constant functions are mapped to zero.

Definition 2.4.3 (Hölder norm and Hölder-α-continuous functions)**.**
Let $I \subset \mathbb{R}^d$ be a compact set with $I = \overline{\text{int}(I)}$, $\alpha > 0$. For any continuous function ℓ having continuous $\lfloor \alpha \rfloor$-order derivatives, define the Hölder-α-norm as

$$\begin{aligned} \|\ell\|_\alpha &= \|\ell\|_\infty + [\ell]_\alpha , \quad \alpha \in (0,1] \\ \|\ell\|_\alpha &= \sum_{0 \leq |k| \leq \lfloor \alpha \rfloor} \|\partial^k \ell\|_\infty + \sum_{|k| = \lfloor \alpha \rfloor} [\partial^k \ell]_{\alpha - \lfloor \alpha \rfloor} , \quad \alpha \in (1, \infty) . \end{aligned} \tag{2.4.1}$$

The set of Hölder-α-continuous functions $\mathcal{C}^\alpha(I)$ on I, call it Hölder-α-space, is defined as the set of all continuous functions ℓ that are $\lfloor \alpha \rfloor$-times continuously differentiable in $\text{int}(I)$ *with $\|\ell\|_\alpha < \infty$.*

Remark 2.4.4.

(i) The map $\|\cdot\|_\alpha$ is a norm since it is the sum of norms and seminorms.

(ii) For any compact $I \subset \mathbb{R}^d$ with $I = \overline{\text{int}(I)}$ and any $\alpha > 0$, $(\mathcal{C}^\alpha(I), \|\cdot\|_\alpha)$ is a Banach space, cf. Driver (2003, Theorem 5.8).

(iii) By definition of the Hölder-α-norm (2.4.1), any $\ell \in \mathcal{C}^\alpha(I)$ is bounded, i.e. $\|\ell\|_\infty < \infty$.

(iv) Regarding Hölder classes in Definition 2.4.1, whenever U is bounded, we have that

$$\sup_{\ell \in H(\alpha, L, U)} \|\ell\|_\infty \leq \max\{-\inf U, \sup U\} < \infty .$$

(v) For any compact set $I \subset \mathbb{R}^d$ with $I = \overline{\text{int}(I)}$ and any compact $U \subset \mathbb{R}$, we have $H(\alpha, L, U) \subset \mathcal{C}^\alpha(I)$. Particularly, for $\ell \in H(\alpha, L, U)$, we have that

$$\|\ell\|_\alpha \leq \max\{-\inf U, \sup U\} + \begin{Bmatrix} (\frac{d^{\lfloor \alpha \rfloor + 1} - d}{d-1} + d^{\lfloor \alpha \rfloor}) L , & d > 1 \\ (\lfloor \alpha \rfloor + 1) L , & d = 1 \end{Bmatrix} < \infty ,$$

because $\sum_{k=1}^{\lfloor \alpha \rfloor} d^k = \frac{d^{\lfloor \alpha \rfloor + 1} - d}{d-1}$.

(vi) The Hölder class $H(\alpha, L, U)$ is closed with respect to $\|\cdot\|_\alpha$. In order to see this, define the functions $\Psi_k, \widetilde{\Psi}_k : \mathcal{C}^\alpha(I) \to [0, \infty)$

$$\Psi_k(\ell) = \|\partial^k \ell\|_\infty , \ 1 \leq |k| \leq \lfloor \alpha \rfloor , \qquad \widetilde{\Psi}_k(\ell) = [\partial^k \ell]_{\alpha - \lfloor \alpha \rfloor} , \ |k| = \lfloor \alpha \rfloor$$

and observe that they are continuous with respect to $\|\cdot\|_\alpha$. Then

$$H(\alpha, L, U) = \|\cdot\|_\infty^{-1}\Big(\big[0, \max\{-\inf U, \sup U\}\big]\Big) \cap \bigcap_{1\le|k|\le\lfloor\alpha\rfloor} \Psi_k^{-1}\big([0,L]\big) \cap \bigcap_{|k|=\lfloor\alpha\rfloor} \widetilde{\Psi}_k^{-1}\big([0,L]\big) \ .$$

(vii) For compact $I \subset \mathbb{R}^d$ with $I = \overline{\mathrm{int}(I)}$, the Hölder-$\alpha$-spaces $\mathcal{C}^\alpha(I)$ are compactly nested in the sense that for $0 < \alpha < \beta < \infty$, we have

$$\mathcal{C}^\beta(I) \subset \mathcal{C}^\alpha(I)$$

and the unit disk $\{\ell \in \mathcal{C}^\beta(I) : \|\ell\|_\beta \le 1\}$ is compact with respect to $\|\cdot\|_\alpha$, cf. Driver (2003, Theorem 5.14).

Unlike Hölder-α-spaces, Hölder classes are not nested in general. However, they are whenever the diameter of the domain I is at most one. For domains with larger diameters, it can be useful to transform the domain appropriately in order to get inclusion of the Hölder classes.

Remark 2.4.5.

(i) When $\operatorname{diam} I \le 1$ the Hölder classes are nested, i.e. we have the inclusion property

$$H(\beta, L, U) \subset H(\alpha, L, U) \ , \quad \alpha \le \beta \ .$$

In order to see this, fix $\ell \in H(\beta, L, U)$. If $\lfloor\alpha\rfloor < \lfloor\beta\rfloor$, the functions $\partial^k\ell$, $|k| = \lfloor\alpha\rfloor$ are continuously differentiable with respect to any argument x_i, hence Lipschitz continuous with Lipschitz constant $\max_i \|\partial_{x_i}\partial^k\ell\|_\infty \le L$ so that

$$|\partial^k\ell(x) - \partial^k\ell(y)| \le L\|x-y\| \le L\|x-y\|^{\alpha-\lfloor\alpha\rfloor} \ .$$

Whenever $\lfloor\alpha\rfloor = \lfloor\beta\rfloor$, we directly conclude

$$|\partial^k\ell(x) - \partial^k\ell(y)| \le L\|x-y\|^{\beta-\lfloor\beta\rfloor} \le L\|x-y\|^{\alpha-\lfloor\alpha\rfloor} \ , \quad |k| = \lfloor\alpha\rfloor = \lfloor\beta\rfloor \ .$$

If $\operatorname{diam} I > 1$, we transform I into $I^* = \frac{I}{\operatorname{diam} I}$ inducing new Hölder classes on I^*, i.e.

$$\begin{aligned} H^*(\alpha, L, U) = \big\{ & \ell : I^* \to U \,\big|\, \ell \text{ is continuous and } \lfloor\alpha\rfloor\text{-times differentiable in } \mathrm{int}(I^*), \\ & |\partial^k\ell(x) - \partial^k\ell(y)| \le L\|x-y\|^{\alpha-\lfloor\alpha\rfloor} \ , \quad |k| = \lfloor\alpha\rfloor, \ x, y \in \mathrm{int}(I^*) \\ & \|\partial^k\ell\|_\infty \le L \ , \quad 1 \le |k| \le \lfloor\alpha\rfloor \ \big\} \ . \end{aligned}$$

We directly see that $\ell \in H(\alpha, L, U)$ corresponds to $\ell^* \in H^*(\alpha, (\operatorname{diam} I)^\alpha \cdot L, U)$, where $\ell^*(x) = \ell((\operatorname{diam} I)x)$. Indeed, for $1 \le |k| \le \lfloor\alpha\rfloor$, it holds that

$$\|\partial^k\ell^*\|_\infty = (\operatorname{diam} I)^{|k|}\|\partial^k\ell\|_\infty \le \big(\operatorname{diam} I\big)^\alpha L \ ,$$

and for $|k| = \lfloor \alpha \rfloor$, we have

$$\begin{aligned}\left|\partial^k \ell^*(x) - \partial^k \ell^*(y)\right| &\leq \big(\operatorname{diam} I\big)^{\lfloor \alpha \rfloor} \left|\partial^k \ell\big((\operatorname{diam} I)x\big) - \partial^k \ell\big((\operatorname{diam} I)y\big)\right| \\ &\leq \big(\operatorname{diam} I\big)^{\alpha} L \|x-y\|^{\alpha - \lfloor \alpha \rfloor} .\end{aligned}$$

Furthermore, we see that for any $0 < a < b < \infty$, we have that every $\ell \in \bigcup_{\alpha \in [a,b]} H(\alpha, L, U)$ corresponds to some

$$\begin{aligned}\ell^* \in \bigcup_{\alpha \in [a,b]} H^*(\alpha, (\operatorname{diam} I)^{\alpha} \cdot L, U) &\subset \bigcup_{\alpha \in [a,b]} H^*(\alpha, (\operatorname{diam} I)^{b} \cdot L, U) \\ &\subset H^*(a, (\operatorname{diam} I)^{b} \cdot L, U) ,\end{aligned}$$

so that results stated uniformly over $H^*(a, (\operatorname{diam} I)^b \cdot L, U)$ especially give the corresponding results uniformly over $\bigcup_{\alpha \in [a,b]} H(\alpha, L, U)$.

(ii) According to Remark 2.4.4 (v) - (vii), $H(\alpha, L, U)$, $\alpha \in [a, b]$, are compact with respect to $\|\cdot\|_{a/2}$, defined in (2.4.1).

(iii) Whenever continuity of a functional $\Psi : H(\alpha, L, U) \to \mathbb{R}^k$, $k \in \mathbb{N}$, $\alpha \in [a, b]$ with respect to $\|\cdot\|_{a/2}$ is to be proven, it is enough to show continuity with respect to the sup-norm $\|\cdot\|_\infty$ since the sup-norm, as a norm on the domain space, is weaker than $\|\cdot\|_{a/2}$.

(iv) Assume $\operatorname{diam} I \leq 1$. For any $\alpha > 0$ and sequences $\alpha_n \nearrow \alpha$, we have

$$\bigcap_{n \in \mathbb{N}} H(\alpha_n, L, U) = H(\alpha, L, U)$$

as one can simply assume that $\alpha_n > \lfloor \alpha \rfloor$ for all n and observe that the map

$$\beta \mapsto [\ell]_{\beta - \lfloor \alpha \rfloor} = \sup_{\substack{x,y \in \operatorname{int}(I) \\ x \neq y}} \frac{|\ell(x) - \ell(y)|}{\|x-y\|^{\beta - \lfloor \alpha \rfloor}} , \quad \beta \in \big(\lfloor \alpha \rfloor, \alpha\big]$$

is continuous.

2.5 Kernel density estimation

Kernel density estimators are broadly used in order to estimate probability densities. For i.i.d. observations $X_1, X_2, \ldots$ with common probability density function $\ell : \mathbb{R}^d \to [0, \infty)$, take some probability density $K : \mathbb{R}^d \to [0, \infty)$ as kernel function and define the kernel density estimator of ℓ as

$$\hat{\ell}_n(x) := \frac{1}{n} \sum_{k=1}^{n} K_h(X_k - x) := \frac{1}{nh^d} \sum_{k=1}^{n} K\left(\frac{X_k - x}{h}\right) ,$$

where h is a bandwidth parameter typically converging to 0 in n that crucially influences the asymptotic properties of the method.

Kernel density estimators have been first studied by Rosenblatt (1956) for indicator kernels and in a more general setting by Parzen (1962) who proved consistency, uniform consistency, asymptotic normality for the mode and studied the mean squared error of kernel density estimators. Silverman (1978) proved strong uniform consistency of kernel density estimators and their derivatives. Silverman (1981) used kernel density estimators in order to detect multimodality in a distribution. Lepskii (1992) gave a method for choosing bandwidth parameters adaptive to the smoothness of Hölder classes. Giné and Guillou (2002) gave uniform convergence rates for kernel density estimators. For extensive overviews cf. Tsybakov (2008), Scott (2015) or Silverman (2018).

In order to exploit smoothness of higher order, e.g. $\alpha > 2$ for Hölder classes, one needs to drop the assumption of non-negativity of the kernel function.

Definition 2.5.1 (Higher order kernels)**.**
Let $\alpha > 0$, $K : \mathbb{R}^d \to \mathbb{R}$ be a kernel. K is said to be of order α if for all $k = (k_1, \dots, k_d) \in \mathbb{N}_0^d$ with $|k| \in \{1, \dots, \lfloor \alpha \rfloor\}$, we have

$$\int K(z)\,\mathrm{d}z = 1\ , \quad \int z^k K(z)\,\mathrm{d}z = 0\ , \quad \int \|z\|^\alpha |K(z)|\,\mathrm{d}z < \infty\ . \tag{2.5.1}$$

The mixed moments of order 2 in (2.5.1) can only be zero if K takes negative values. Hence, kernels of order $\alpha > 2$ always take positive and negative values.

Univariate kernels of order α having support $[-1, 1]$ can be constructed from orthogonal polynomials, such as Legendre polynomials, cf. Tsybakov (2008, Section 1.2.2). d-variate kernels of order α can then easily be obtained by multiplying d univariate kernels of order α, i.e. for univariate kernels $K_1, \dots, K_d$ of order α define $K : \mathbb{R}^d \to \mathbb{R}$ by

$$K = \prod_{j=1}^{n} K_j\ .$$

3 M-estimation and supremum distance

3.1 General methods for deriving uniform convergence rates

In order to judge the quality of non-parametric estimation methods, global risk measures are of particular interest. Typical examples are mean integrated squared error, i.e. L^2-risk or the sup-error, i.e. L^∞-risk.

The main results in this thesis regard local M-estimation in mixture of regressions models. Particularly, asymptotic results are given for the L^∞-error of the estimators as well as uniformly over the whole parameter class. In one of the models, cf. Section 4.1, we have to deal with a model that is not fully identifiable. In fact, it is only identifiable up to label switching of the mixture components. In the following, we give variations of Theorems 2.1.1 and 2.1.2 that are tailored to those problems as well as a set of assumptions allowing assertions regarding convergence rates of estimators in those models.

Let Γ, I be arbitrary sets and let Θ be a normed space with norm $\|\cdot\|$. For every $\gamma \in \Gamma$ introduce a deterministic function $M(\cdot,\cdot;\gamma) : \Theta \times I \to \mathbb{R}$ that, for any x, γ, is minimized in its first argument by $\theta \in \Theta$ iff $\theta \in \mathfrak{S}_{x;\gamma}$, where $\mathfrak{S}_{x;\gamma} \subset \Theta$ is a non-empty finite set. If the model is identifiable, the sets $\mathfrak{S}_{x;\gamma}$ only contain one element each, namely $\theta(x;\gamma)$, the minimizer of $M(\cdot, x;\gamma)$, i.e.

$$\theta(x;\gamma) \in \operatorname*{argmin}_{\theta\in\Theta} M(\theta, x;\gamma) \ . \tag{3.1.1}$$

Those parameter sets $\mathfrak{S}_{x;\gamma}$ are to be estimated.

Assume we have a statistical model $\big(\Omega, \mathcal{A}, (\mathbb{P}_\gamma)_{\gamma\in\Gamma}\big)$ and a sequence of random functions $M_n : \Theta \times I \times \Gamma \times \Omega \to \mathbb{R}$, where we use the abbreviation

$$M_n(\theta, x;\gamma) = M_n(\theta, x;\gamma;\cdot) \ , \quad \theta \in \Theta,\ x \in I,\ \gamma \in \Gamma \ .$$

Further suppose that, for any $x \in I$, the random functions $M_n(\cdot, x;\gamma)$ are minimized by some $\hat\theta_n(x;\gamma)$, i.e.

$$\hat\theta_n(x;\gamma) \in \operatorname*{argmin}_{\theta\in\Theta} M_n(\theta, x;\gamma) \ .$$

We will give conditions sufficient for uniform consistency of $\hat\theta_n(\cdot;\gamma)$ as well as for $\hat\theta_n(\cdot;\gamma)$ having uniform convergence rate $r_{n,\gamma}$ both uniformly over the whole class Γ. That is, we

give conditions under which we have

$$\lim_{n\to\infty} \sup_{\gamma\in\Gamma} \mathbb{P}_\gamma\Big(\sup_{x\in I} \min_{\theta_*\in\mathfrak{S}_{x;\gamma}} \big\|\hat\theta_n(x;\gamma) - \theta_*\big\| \geq \varepsilon\Big) = 0 \,, \quad \varepsilon > 0 \,,$$

$$\lim_{\delta\to\infty} \limsup_{n\to\infty} \sup_{\gamma\in\Gamma} \mathbb{P}_\gamma\Big(r_{n,\gamma} \sup_{x\in I} \min_{\theta_*\in\mathfrak{S}_{x;\gamma}} \big\|\hat\theta_n(x;\gamma) - \theta_*\big\| \geq \delta\Big) = 0 \,.$$

Remark 3.1.1.

(i) The set Γ typically consists of all model parameters. The random functions M_n are generally independent of γ as proper estimators do not depend on unknown model parameters. However, we will later discuss estimators that depend on some nuisance parameter like the smoothness α in regard of Hölder-α-smooth functions and include α in the parameter γ. We will formulate asymptotic results for families of estimators indexed by the nuisance and subsequently propose a method to make a data driven choice for the nuisance parameter so that the resulting estimator is in fact again independent of the model and nuisance parameters. Note that this is why convergence rates $r_{n,\gamma}$ also depend on γ. For notational simplicity in the following two general results, we use this unusual notation in which random functions take model parameters as arguments and interpret $\gamma \in \Gamma$ to consist of model parameters of which the random functions are independent and nuisance parameters on which they may depend.

(ii) In regression models for example, Γ might consist of tuples $(\theta(\cdot), \ell, \alpha)$, where $\theta(\cdot)$ are parameter functions coming from some Hölder classes, cf. Section 2.4, α is the Hölder smoothness and ℓ are covariate densities also coming from some set of functions.

(iii) Whenever a model is not identifiable, the set $\mathfrak{S}_{x;\gamma}$ helps formalizing asymptotic properties of estimators that cannot be consistent due to non-identifiability. Consider for example mixture of regressions models that often are only identifiable up to label switching. In those models $\mathfrak{S}_{x;\gamma}$ usually consists of the model parameter and all label switched versions thereof.

Let us start with a uniform consistency result. The following result is a rather direct extension of Theorem 2.1.1 in the sense that uniform consistency of the random functions M_n and the minima of M being well-separated also yield uniform consistency in this context.

Theorem 3.1.2 (Uniform consistency)**.** *Let Θ be a normed space with norm $\|\cdot\|$ and assume that*

$$\lim_{n\to\infty} \sup_{\gamma\in\Gamma} \mathbb{P}_\gamma \Big(\sup_{\theta\in\Theta} \sup_{x\in I} \big|M_n(\theta, x;\gamma) - M(\theta, x;\gamma)\big| \geq \eta\Big) = 0 \,, \quad \eta > 0$$

as well as that

(∗) *for all* $\varepsilon > 0$ *there is an* $\eta > 0$ *so that for every* $\theta \in \Theta$, $x \in I$, $\gamma \in \Gamma$ *with* $M(\theta, x; \gamma) - M(\theta_*, x; \gamma) < \eta$ *for some* $\theta_* \in \mathfrak{S}_{x;\gamma}$, *we have* $\min_{\theta_* \in \mathfrak{S}_{x;\gamma}} \|\theta - \theta_*\| < \varepsilon$.

Then the estimator $\hat{\theta}_n(\cdot;\gamma)$ *is uniformly consistent for* $\mathfrak{S}_{x;\gamma}$, *i.e. for all* $\varepsilon > 0$, *we have*

$$\lim_{n\to\infty} \sup_{\gamma\in\Gamma} \mathbb{P}_\gamma\Big(\sup_{x\in I} \min_{\theta_*\in\mathfrak{S}_{x;\gamma}} \big\|\hat{\theta}_n(x;\gamma) - \theta_*\big\| \geq \varepsilon\Big) = 0 \, .$$

The following theorem is a generalization of Theorem 2.1.2 that gives conditions for uniform convergence rates uniformly over the model parameters γ for possibly unidentifiable models.

Theorem 3.1.3. *Let the following assumptions be true:*

(i) There is an $\eta > 0$ *and constants* $C_1, C_2 > 0$ *so that for every* $\varepsilon \leq \eta$,

$$\inf_{\gamma\in\Gamma} \inf_{x\in I} \inf_{*} \Big[M(\theta, x; \gamma) - M(\theta_*(x;\gamma), x; \gamma)\Big] \quad \geq \quad C_1\varepsilon^2$$

$$\limsup_{n\to\infty} \sup_{\gamma\in\Gamma} \sup_{\varepsilon\leq\eta} \frac{t_{n,\gamma}}{\phi_n(\varepsilon)} \mathbb{E}_\gamma\Big[\sup_{x\in I} \sup_{*} \min_{\theta_*\in\mathfrak{S}_{x;\gamma}} \big\|W_n(\theta, x;\gamma) - W_n(\theta_*, x;\gamma)\big\|\Big] \quad \leq \quad C_2 \, ,$$

where the third infimum is taken over $\{\theta \in \Theta : \min_{\theta_*\in\mathfrak{S}_{x;\gamma}} \|\theta - \theta_*\| = \varepsilon\}$, *the fourth supremum is taken over* $\{\theta \in \Theta : \min_{\theta_*\in\mathfrak{S}_{x;\gamma}} \|\theta - \theta_*\| \leq \varepsilon\}$ *and* $\theta_*(x;\gamma) \in \mathfrak{S}_{x;\gamma}$ *is any minimizer of* $M(\cdot, x;\gamma)$. *Furthermore,* $W_n(\theta, x;\gamma) := M_n(\theta, x;\gamma) - M(\theta, x;\gamma)$, $\phi_n : (0,\infty) \to (0,\infty)$ *are functions so that* $\phi_n(\cdot)/\cdot^\alpha$ *is decreasing for some* $\alpha < 2$ *and* $t_{n,\gamma} \to \infty$ *for every* $\gamma \in \Gamma$.

(ii) For all $\delta > 0$, *we have* $\sup_{\gamma\in\Gamma} \mathbb{P}_\gamma\big(\sup_{x\in I} \min_{\theta_*\in\mathfrak{S}_{x;\gamma}} \big\|\hat{\theta}_n(x;\gamma) - \theta_*\big\| \geq \delta\big) = o(1)$.

If sequences $(r_{n,\gamma})$ *satisfy* $r_{n,\gamma}^2 \phi(1/r_{n;\gamma}) \leq t_{n,\gamma}$ *for all* n, γ *as well as* $\inf_\gamma r_{n,\gamma} \to \infty$, *then*

$$\lim_{\delta\to\infty} \limsup_{n\to\infty} \sup_{\gamma\in\Gamma} \mathbb{P}_\gamma\Big(r_{n,\gamma} \sup_{x\in I} \min_{\theta_*\in\mathfrak{S}_{x;\gamma}} \big\|\hat{\theta}_n(x;\gamma) - \theta_*\big\| \geq \delta\Big) = 0 \, .$$

The proof of both results can be found in Section 5.1.

Application to local M-estimation

Let us give a more specific set of conditions that yield assertions on M-estimators having uniform convergence rates r_n. The first assumption in (i) of Theorem 3.1.3 can be provided by a Taylor expansion of M of order two around its minimizer $\theta_*(x;\gamma)$. Hence, differentiability assumptions need to be made. The second assumption in (i) holds whenever the gradients of the random functions M_n have uniform L^1 convergence rate r_n. Uniform consistency follows from compactness of all sets, continuity of M and the uniform consistency of M_n by Theorem 2.1.1. Note that the set of assumptions simplifies when the model is identified, cf. Assumption A.2.1. Moreover, note that we will drop dependence of the random functions M_n and the convergence rates r_n on γ from now on.

Assumption 3.1.4. Let $\Theta \subset \Xi \subset \mathbb{R}^m$, $I \subset \mathbb{R}^d$, $(\Gamma, \|\cdot\|)$ be a normed space, $M : \Xi \times I \times \Gamma \to \mathbb{R}$ be a deterministic function, $M_n : \Xi \times I \to \mathbb{R}$ be random functions, $(r_n)_{n\in\mathbb{N}} \subset (0,\infty)$, be a sequence with $r_n \to \infty$.

(A1) Assume that Θ is compact and convex with $\Theta = \overline{\operatorname{int}(\Theta)}$, Ξ is open and convex, I is compact and $(\Gamma, \|\cdot\|)$ is a compact normed space. Furthermore, there is a constant $\bar{C} < \infty$ so that for all $\theta, \theta' \in \Theta$, there is an $l \in \mathbb{N}_0$ and $\bar{\theta}_1, \dots, \bar{\theta}_l \in \Theta$ so that with $\theta = \bar{\theta}_0$, $\theta' = \bar{\theta}_{l+1}$, we have

$$\bar{\theta}_{k+1} - \bar{\theta}_k = c_k e_{j_k} \ , \quad k = 0, \dots, l$$

for some unit vectors e_{j_k} and some coefficients $c_k \in \mathbb{R}$ with $\sum_{k=0}^{l} |c_k| \le \bar{C}\|\theta - \theta'\|$.

(A2) There is a set of permutations

$$\mathfrak{Z} \subset \{\zeta : \{1, \dots, m\} \to \{1, \dots, m\} : \zeta \text{ bijective}\} \ ,$$

so that M and M_n are invariant under permuting the first argument by $\zeta \in \mathfrak{Z}$.

(A3) The function M is continuous, i.e. the map

$$(\theta, x; \gamma) \mapsto M(\theta, x; \gamma)$$

is continuous. For every $x \in I$, $\gamma \in \Gamma$, the contrast $M(\cdot, x; \gamma)$ attains a minimum at $\theta_*(x;\gamma)$ iff

$$\theta_*(x;\gamma) \in \mathfrak{S}_{x;\gamma} \ ,$$

where $\mathfrak{S}_{x;\gamma} = \zeta(\mathfrak{S}_{x;\gamma}) \subset \Theta$, $\zeta \in \mathfrak{Z}$ and $\#\mathfrak{S}_{x;\gamma} = \#\mathfrak{Z}$. Furthermore, for any $x \in I$, $\gamma \in \Gamma$, $\theta_*(x;\gamma) \in \mathfrak{S}_{x;\gamma}$, there is a permutation $\zeta_{x;\gamma} \in \mathfrak{Z}$ so that the maps $(x;\gamma) \mapsto \zeta_{x;\gamma}(\theta_*(x;\gamma))$ are continuous.

(A4) For all $x \in I, \gamma \in \Gamma$, the function $M(\cdot, x; \gamma)$ is twice continuously differentiable in its first argument and the Hessian matrix

$$V_x\big(\theta_*(x;\gamma);\gamma\big) := \partial^2_{\theta^2} M\big(\theta_*(x;\gamma), x; \gamma\big)$$

is positive definite for all $\theta_*(x;\gamma) \in \mathfrak{S}_{x;\gamma}$. Particularly, the eigenvalues $\lambda^1_{x,\gamma} \ge \dots \ge \lambda^m_{x,\gamma}$ of $V_x(\theta_*(x;\gamma);\gamma)$ are positive. Furthermore, the map $(x,\gamma) \mapsto V_x\big(\theta_*(x;\gamma);\gamma\big)$ is continuous.

(A5) The Hessian matrices $V_x(\cdot;\gamma)$ are uniformly Lipschitz continuous in θ, i.e. for all $\theta, \theta' \in \Xi$, we have

$$\sup_{\gamma\in\Gamma} \sup_{x\in I} \|V_x(\theta;\gamma) - V_x(\theta';\gamma)\| \le L_{\text{Hess}} \|\theta - \theta'\| \ ,$$

where $L_{\text{Hess}} < \infty$ depends only on Ξ, I and Γ.

(A6) There is an $\varepsilon^* > 0$ so that for all $0 < \varepsilon < \varepsilon^*$, $x \in I$, $\gamma \in \Gamma$, the balls $\{\theta \in \Xi : \|\theta - \theta_*\| \leq \varepsilon\}$, $\theta_* \in \mathfrak{S}_{x;\gamma}$ are disjoint.

(A7) The empirical contrast M_n is continuously differentiable in its first argument and for the gradients

$$S_n(\theta, x) := \partial_\theta M_n(\theta, x) \,, \quad S(\theta, x; \gamma) := \partial_\theta M(\theta, x; \gamma) \tag{3.1.2}$$

it holds that

$$\limsup_{n\to\infty} \sup_{\gamma\in\Gamma} r_n \mathbb{E}_\gamma \Big[\sup_{x\in I, \theta\in\Theta} \big\| S_n(\theta, x) - S(\theta, x; \gamma) \big\| \Big] < \infty \,.$$

(A8) The empirical contrast M_n is uniformly consistent for M, i.e.

$$\lim_{n\to\infty} \sup_{\gamma\in\Gamma} \mathbb{P}_\gamma \Big(\sup_{x\in I, \theta\in\Theta} \big| M_n(\theta, x) - M(\theta, x; \gamma) \big| \geq \varepsilon \Big) = 0 \,, \quad \varepsilon > 0 \,.$$

Remark 3.1.5.

(i) The latter part of Assumption **(A1)** lets us bound increments $|f(\theta) - f(\tilde{\theta})|$ of functions f on Θ by sums of increments for which arguments only differ in one component. Note that for those vectors $\bar{\theta}_k$, we have $[\bar{\theta}_k, \bar{\theta}_{k+1}] \subset \Theta$ as Θ is convex. Also note that the assumption is fulfilled by every compact cuboid $\Theta \subset \mathbb{R}^m$ containing an open subset because for any $\theta' = (\vartheta'_1, \ldots, \vartheta'_m)^T$, $\theta = (\vartheta_1, \ldots, \vartheta_m)^T \in \Theta$, the vectors

$$\bar{\theta}^j = (\vartheta_1, \ldots, \vartheta_j, \vartheta'_{j+1}, \ldots, \vartheta'_m)^T \in \Theta \,, \quad j = 0, \ldots, m$$

fulfil the postulated properties. Particularly, one can choose $\bar{C} = m$.

(ii) Whenever the model is identifiable, the set of Assumptions 3.1.4 shrinks drastically. **(A2)** and **(A6)** can be dropped entirely. **(A3)** reduces to M being continuous, the minima being unique and the parameter functions being continuous. A complete list of altered assumptions can be found in the Appendix, cf. Assumption A.2.1.

(iii) We will consider estimation only for parameters in Θ but introduced Ξ so that differentiation of the contrast functions on the boundary of Θ is well-defined and so that particularly $S(\theta_*(x;\gamma), x; \gamma) = 0$ when $\theta_*(x;\gamma) \in \mathfrak{S}_{x;\gamma}$ is a boundary point of Θ. The latter is true because points of minima of M_n on the boundary of Θ are especially local minima on the open set Ξ.

Theorem 3.1.6. *Under Assumption 3.1.4, any sequence* $\hat{\theta}_n(x) = \operatorname{argmin} M_n(\cdot, x)$ *has uniform convergence rate* r_n*, i.e.*

$$\lim_{\delta\to\infty} \limsup_{n\to\infty} \sup_{\gamma\in\Gamma} \mathbb{P}_\gamma \Big(r_n \sup_{x\in I} \min_{\theta_*\in\mathfrak{S}_{x;\gamma}} \big\| \hat{\theta}_n(x) - \theta_* \big\| \geq \delta \Big) = 0 \,.$$

Proof of Theorem 3.1.6. We need to check the assumptions of Theorem 3.1.3 for

$$\phi_n = \mathrm{id} \qquad \text{and} \qquad t_{n,\gamma} = r_{n,\gamma} = r_n \ .$$

We obviously have that $t_{n,\gamma} \to \infty$, $t \mapsto \phi_n(t)/t^{\frac{3}{2}} = t^{-\frac{1}{2}}$ is decreasing on $(0,\infty)$, $r_{n,\gamma}^2 \phi_n(1/r_{n,\gamma}) = r_{n,\gamma} = t_{n,\gamma}$.

First, observe that there is a bounded open set $\Theta \subset \widetilde{\Xi} \subset \Xi$ so that

$$\mathrm{dist}(\Theta, \partial\widetilde{\Xi}) =: \bar{\varepsilon} > 0 \ .$$

Indeed, assume $\bar{\varepsilon} = 0$, then there is a sequence $(\theta_n)_{n\in\mathbb{N}} \subset \Theta$ so that $\mathrm{dist}(\theta_n, \partial\widetilde{\Xi}) \to 0$. As Θ is compact, there is a subsequence $(\theta_{n_k})_{k\in\mathbb{N}}$ of $(\theta_n)_{n\in\mathbb{N}}$ so that $\theta_{n_k} \to \bar{\theta} \in \Theta$. Since $\theta \mapsto \mathrm{dist}(\theta, \partial\widetilde{\Xi})$ is continuous, $\partial\widetilde{\Xi}$ is closed and $\mathrm{dist}(\theta_{n_k}, \partial\widetilde{\Xi}) \to 0$, we deduce $\bar{\theta} \in \partial\widetilde{\Xi}$, a contradiction. We can without loss of generality assume that $\widetilde{\Xi}$ is convex as the convex hull of $\widetilde{\Xi}$ is bounded and a subset of Ξ.

Fix some $\varepsilon < \min\{\varepsilon^*, \bar{\varepsilon}\}$. Then, this implies that for any $\gamma \in \Gamma$, $x \in I$,

$$\big\{\theta \in \Xi : \|\theta - \theta_*(x;\gamma)\| \le \varepsilon\big\} \subset \Xi \ , \quad \theta_*(x;\gamma) \in \mathfrak{S}_{x;\gamma}$$

and according to **(A6)**, the balls are in particular disjoint.

Let us prove the first point of (i). For any $\gamma \in \Gamma$, $x \in I$, our considerations above and **(A2)** give some $\theta_*(x;\gamma) \in \mathfrak{S}_{x;\gamma}$ so that a second-order Taylor approximation around $\theta_*(x;\gamma)$ yields for every $\theta \in \Xi$ with $\|\theta - \theta_*(x;\gamma)\| = \varepsilon$ the existence of a $\xi_{x,\theta,\gamma} \in [\theta, \theta_*(x;\gamma)]$ so that

$$\begin{aligned}
&\inf_{\gamma\in\Gamma}\inf_{x\in I}\min_{\theta_*\in\mathfrak{S}_{x;\gamma}}\inf_{\substack{\theta\in\Theta:\\ \|\theta-\theta_*\|=\varepsilon}}\Big[M(\theta,x;\gamma) - M(\theta_*,x;\gamma)\Big] \\
&\ge \inf_{\gamma\in\Gamma}\inf_{x\in I}\min_{\theta_*\in\mathfrak{S}_{x;\gamma}}\inf_{\substack{\theta\in\Xi:\\ \|\theta-\theta_*\|=\varepsilon}}\Big[M(\theta,x;\gamma) - M(\theta_*,x;\gamma)\Big] \\
&= \inf_{\gamma\in\Gamma}\inf_{x\in I}\inf_{\substack{\theta\in\Xi:\\ \|\theta-\theta_*(x;\gamma)\|=\varepsilon}}\Big[M(\theta,x;\gamma) - M\big(\theta_*(x;\gamma),x;\gamma\big)\Big] \\
&\ge \inf_{\gamma\in\Gamma}\inf_{x\in I}\inf_{\substack{\theta\in\Xi:\\ \|\theta-\theta_*(x;\gamma)\|=\varepsilon}}\frac{1}{2}\big(\theta-\theta_*(x;\gamma)\big)^T V_x\big(\theta_*(x;\gamma);\gamma\big)\big(\theta-\theta_*(x;\gamma)\big) \\
&\quad - \sup_{\gamma\in\Gamma}\sup_{x\in I}\sup_{\substack{\theta\in\Xi:\\ \|\theta-\theta_*(x;\gamma)\|=\varepsilon}}\left|\frac{1}{2}\big(\theta-\theta_*(x;\gamma)\big)^T\Big(V_x\big(\theta_*(x;\gamma);\gamma\big) - V_x\big(\xi_{x,\theta,\gamma};\gamma\big)\Big)\big(\theta-\theta_*(x;\gamma)\big)\right| \\
&\ge \inf_{\gamma\in\Gamma}\inf_{x\in I}\frac{1}{2}\varepsilon^2\lambda_{x,\gamma}^m - \frac{L_{\mathrm{Hess}}}{2}\varepsilon^3 \ ,
\end{aligned}$$

according to **(A4)** and **(A5)**, where $\lambda_{x,\gamma}^m$ is the smallest eigenvalue of $V_x\big(\theta_*(x;\gamma);\gamma\big)$. Since eigenvalues of a matrix depend continuously on its entries, the entries of the Hessian matrices $V_x\big(\theta_*(x;\gamma);\gamma\big)$ depend continuously on $(x;\gamma)$ by **(A4)**, so that we can deduce

$$\inf_{\gamma\in\Gamma}\inf_{x\in I}\lambda_{x,\gamma}^m > 0$$

by compactness of $\Gamma \times I$, cf. **(A1)**. Conclude by choosing $\eta \leq \min\{\varepsilon^*, \bar{\varepsilon}\}$ small enough.

Prove the second part of (i). Because of the invariance of the contrast functions under permutations $\zeta \in \mathfrak{Z}$, cf. **(A2)**, we only need to show that for some $\eta < \min\{\varepsilon^*, \bar{\varepsilon}\}$ and for any $\theta_*(x;\gamma) \in \mathfrak{S}_{x;\gamma}$,

$$\limsup_{n\to\infty} \sup_{0<\varepsilon\leq\eta} \sup_{\gamma\in\Gamma} \frac{r_n}{\varepsilon} \mathbb{E}_\gamma \Big[\sup_{x\in I} \sup_{\substack{\theta\in\Theta: \\ \|\theta-\theta_*(x;\gamma)\|\leq\varepsilon}} \big| W_n(\theta,x;\gamma) - W_n(\theta_*(x;\gamma),x;\gamma) \big| \Big] \leq C_2 \ ,$$

for some constant $C_2 > 0$, where $W_n(\theta,x;\gamma) := M_n(\theta,x) - M(\theta,x;\gamma)$.

For any $\theta \in \Theta$, there is some $l_{\theta,x;\gamma} \in \mathbb{N}_0$, $\bar{\theta}_1(\theta,x;\gamma), \ldots, \bar{\theta}_{l_{\theta,x;\gamma}}(\theta,x;\gamma) \in \Theta$ with the properties described in **(A1)**. By using the notation

$$\theta = \bar{\theta}_0(\theta,x;\gamma) \ , \quad \theta_*(x;\gamma) = \bar{\theta}_{l_{\theta,x;\gamma}+1}(\theta,x;\gamma) \ ,$$
$$\bar{\theta}_k(\theta,x;\gamma) = \big(\bar{\vartheta}^k_1(\theta,x;\gamma), \ldots, \bar{\vartheta}^k_m(\theta,x;\gamma)\big)^T \ , \quad k = 0, \ldots, l_{\theta,x;\gamma}+1$$

as well as $t = (t_1, \ldots, t_m)$, the fundamental theorem of calculus gives for any n, γ that

$$\begin{aligned}
&\sup_{x\in I} \sup_{\substack{\theta\in\Theta: \\ \|\theta-\theta_*(x;\gamma)\|\leq\varepsilon}} \big| W_n(\theta,x;\gamma) - W_n\big(\theta_*(x;\gamma),x;\gamma\big) \big| \\
&\leq \sup_{x\in I} \sup_{\substack{\theta\in\Theta: \\ \|\theta-\theta_*(x;\gamma)\|\leq\varepsilon}} \sum_{k=0}^{l_{\theta,x;\gamma}} \big| W_n\big(\bar{\theta}_{k+1}(\theta,x;\gamma),x;\gamma\big) - W_n\big(\bar{\theta}_k(\theta,x;\gamma),x;\gamma\big) \big| \\
&= \sup_{x\in I} \sup_{\substack{\theta\in\Theta: \\ \|\theta-\theta_*(x;\gamma)\|\leq\varepsilon}} \sum_{k=0}^{l_{\theta,x;\gamma}} \Big| \int_{\bar{\vartheta}^k_{j_k}(\theta,x;\gamma)}^{\bar{\vartheta}^{k+1}_{j_k}(\theta,x;\gamma)} \partial_{t_j} W_n(t,x;\gamma) \, \mathrm{d}t_j \Big| \\
&\leq \bar{C}\varepsilon \sup_{x\in I} \sup_{\theta\in\Theta} \big\| S_n(\theta,x) - S(\theta,x;\gamma) \big\| \ ,
\end{aligned}$$

so that the second part of (i) is given directly by **(A7)**.

We will prove (ii), i.e. the uniform consistency of $\hat{\theta}_n(\cdot)$, by using Theorem 3.1.2. As uniform consistency of the contrast M_n is given by **(A8)**, only $(*)$ in the assumptions of Theorem 3.1.2 needs to be proved.

Assume $(*)$ does not hold. Then there is an $\varepsilon > 0$ so that for any sequence $\eta_n \to 0$, we find $x_n \in I$, $\theta_n \in \Theta$, $\gamma_n \in \Gamma$ so that for every $n \in \mathbb{N}$

$$M(\theta_n,x_n;\gamma_n) - M\big(\theta_*(x_n;\gamma_n),x_n;\gamma_n\big) < \eta_n \ , \quad \min_{\theta_*(x_n;\gamma_n)\in\mathfrak{S}_{x_n,\gamma_n}} \big\|\theta_n - \theta_*(x_n;\gamma_n)\big\| \geq \varepsilon \ . \tag{3.1.3}$$

As $\Theta \times I \times \Gamma$ is compact according to **(A1)**, there is a subsequence $\big((\theta_{n_k},x_{n_k},\gamma_{n_k})\big)_{k\in\mathbb{N}}$ of $\big((\theta_n,x_n,\gamma_n)\big)_{n\in\mathbb{N}}$ converging to a point $(\theta',x',\gamma') \in \Theta\times I\times\Gamma$. By continuity of $(\theta,x;\gamma) \mapsto$

$M(\theta, x; \gamma)$ that is given by **(A3)** as well as the continuity of $(x;\gamma) \mapsto \zeta_{x;\gamma}(\theta_*(x;\gamma))$, we have $M(\theta', x'; \gamma') = M(\theta_*(x';\gamma'), x'; \gamma')$, which implies that $\theta' \in \mathfrak{S}_{x',\gamma'}$ Now, according to the right-hand side of (3.1.3) and the continuity of the functions $(x;\gamma) \mapsto \zeta_{x;\gamma}(\theta_*(x;\gamma))$, we have

$$\min_{\theta_*(x';\gamma') \in \mathfrak{S}_{x',\gamma'}} \left\| \theta' - \theta_*(x';\gamma') \right\| \geq \liminf_{k\to\infty} \min_{\theta_*(x_{n_k};\gamma_{n_k}) \in \mathfrak{S}_{x_{n_k},\gamma_{n_k}}} \left\| \theta_{n_k} - \theta_*(x_{n_k};\gamma_{n_k}) \right\| \geq \varepsilon \,,$$

a contradiction as $\theta \mapsto M(\theta, x'; \gamma')$ is only minimized by elements of $\mathfrak{S}_{x',\gamma'}$, cf. **(A3)**. □

3.1.1 Uniform adaptive estimation methods

As Lepskii (1991) pointed out, there is no optimally adaptive pointwise estimator for estimating signals identified by a function coming from univariate Hölder classes in the Gaussian white noise model whenever the smoothness parameter is only known to come from a set A with at least two elements. That is, there is no estimator estimating all functions within $H(\alpha, L, U)$, $\alpha \in A$ at the respective minimax rates. The author further gives a weaker notion of quality of adaptive estimation.

Lepskii (1992) gave general conditions under which estimators are adaptive according to the notion in Lepskii (1991). This theory is applicable to a variety of estimation problems, e.g. kernel density estimation. Under those conditions, he particularly gave a method for constructing adaptive estimators from estimators that each estimate functions coming from one nuisance class with minimax rate, the Lepski method.

Assume one has estimators corresponding to every possible nuisance parameter fulfilling certain conditions, cf. Lepskii (1992). Proceed by laying a finite ordered net on the set of nuisance parameters and examine how estimators corresponding to the net points improve with respect to a certain loss function when choosing net points of higher order. Then choose the last estimator that still improves. As he further shows, inserting a multiplicative logarithmic punitive term into the convergence rate suffices in many models in order to achieve adaptivity, e.g. kernel density estimation with densities coming from Hölder classes.

Uniform adaptive estimation in local M-estimation

Proving adaptivity for specific estimators according to the Lepski method requires establishing exponential deviation inequalities for the estimators, cf. Lepskii (1992, Lemma 1). While inequalities of this type are available for a variety of estimators under mild conditions, e.g. estimators in the form of sums of i.i.d. random variables like the kernel density estimator, this is not the case for all estimators. We will give an approach on uniform adaptive M-estimation when there is no exponential deviation inequality for the estimators at hand.

Assume that in the general M-estimation setting described at the beginning of this chapter, the parameter function $\theta(\cdot;\gamma)$ defined in (3.1.1) depends on some nuisance parameter α like the smoothness α in regard of Hölder classes. If this nuisance is unknown a priori, a pointwise optimal adaptive estimator in the sense described by Lepskii (1991) should be unattainable as mentioned above. However, it can be possible to obtain estimators of this type when examining the L^∞-errors because a logarithmic punitive term typically is already present.

Assume one knows estimators

$$\hat{\theta}_n(x;\alpha) = \operatorname*{argmin}_{\theta\in\Theta} M_n(\theta, x;\alpha) \ ,$$

where M_n now depends on the nuisance parameter α, to have uniform convergence rates $\left(\frac{n}{\log n}\right)^{\frac{\alpha}{2\alpha+d}}$, i.e.

$$\lim_{\eta\to\infty} \limsup_{n\to\infty} \sup_{\gamma\in\Gamma(\alpha)} \mathbb{P}_\gamma\left(\left(\frac{n}{\log n}\right)^{\frac{\alpha}{2\alpha+d}} \sup_{x\in I} \min_{\theta_*\in\mathfrak{S}_{x;\gamma;\alpha}} \left\|\hat{\theta}_n(x;\alpha) - \theta_*\right\| \geq \eta\right) = 0 \ , \qquad \alpha > 0 \ ,$$

where $\Gamma(\alpha)$ as well as $\mathfrak{S}_{x;\gamma;\alpha}$ now depend on α as well. Note that this is the most common uniform non-parametric convergence rate depending on nuisance parameters. Consider $\gamma \in \Gamma(\alpha)$ to also model dependency of the parameter $\theta_*(\cdot;\gamma)$ and the asymptotic criterion function M on α. Nonetheless, let the functions M and S depend on α in notation to make it more obvious to which nuisance class the respective γ belongs, i.e.

$$M(\cdot,\cdot;\gamma;\alpha) := M(\cdot,\cdot;\gamma) \ , \quad S(\cdot,\cdot;\gamma;\alpha) := S(\cdot,\cdot;\gamma) \ , \quad \gamma \in \Gamma(\alpha) \ .$$

Further assume one knows the nuisance parameter to come from a compact interval $[a,b] \subset (0,\infty)$, the contrast functions to be differentiable and that **(A7)** holds uniformly over all $\alpha \in [a,b]$, i.e.

$$\limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \left(\frac{n}{\log n}\right)^{\frac{\alpha}{2\alpha+d}} \mathbb{E}_\gamma\left[\sup_{x\in I,\theta\in\Theta} \left\|S_n(\theta,x;\alpha) - S(\theta,x;\gamma;\alpha)\right\|\right] \leq C^* < \infty \ . \tag{3.1.4}$$

Remember that S_n and S are the gradients of M_n and M, cf. (3.1.2). The general idea is to use the Lepski method first described in Lepskii (1992) for the gradients S_n in order to obtain a data driven nuisance parameter $\hat{\alpha}_n \in [a,b]$ so that

$$\limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \left(\frac{n}{\log n}\right)^{\frac{\alpha}{2\alpha+d}} \mathbb{E}_\gamma\left[\sup_{x\in I,\theta\in\Theta} \left\|S_n(\theta,x;\hat{\alpha}_n) - S(\theta,x;\gamma;\alpha)\right\|\right] < \infty \ , \tag{3.1.5}$$

that is, $S_n(\cdot,\cdot;\hat{\alpha}_n)$ being an adaptive estimator for the asymptotic gradients S with nuisance coming from $[a,b]$.

Under Assumptions similar to Assumption 3.1.4, it is reasonable to assume that the adaptivity extends from the gradient to the estimator $\hat{\theta}_n(\cdot;\hat{\alpha}_n)$ itself by using Theorem 3.1.3. Note that the convergence rates should typically remain identical because a logarithmic punitive term is already present due to examining L^∞-errors.

The application of the Lepski method now works as follows. Let us first lay a grid over $[a,b]$ that grows logarithmically in n, i.e.

$$\beta_k = a + k\frac{b-a}{N}, \quad k = 0,\dots,N, \quad N = \lceil \log n \rceil \,, \tag{3.1.6}$$

where $\lceil x \rceil$ is the smallest integer strictly larger than x. Let us further use the notation

$$r(\alpha) = \left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+d}}, \qquad r_k = r(\beta_k)\,. \tag{3.1.7}$$

Subsequently, define the adaptive data driven grid point by

$$\hat{\alpha}_n = \beta_{\hat{k}},$$

where

$$\hat{k}_n = \hat{k} = \max\left\{0 \le k \le N : \sup_{x\in I,\theta\in\Theta} \left\|S_n(\theta,x;\beta_k) - S_n(\theta,x;\beta_l)\right\| \le C_{\mathrm{Lep}} r_l \quad \forall\, 0 \le l \le k\right\}, \tag{3.1.8}$$

where the Lepski constant $C_{\mathrm{Lep}} < \infty$ is to be chosen large enough. It will be model-dependent and gets specified within the proof of the adaptivity of the corresponding empirical gradient

$$S_n(\cdot,\cdot;\hat{\alpha}_n)\,, \quad \hat{\alpha}_n = \beta_{\hat{k}}\,, \tag{3.1.9}$$

i.e. (3.1.5). All parameters above actually depend on n. However, we choose to drop the dependence in notation for convenience.

Remark 3.1.7.

(i) Note that $\hat{k}$ is a random variable with values in $\{0,\dots,N\}$.

(ii) Also note that we implicitly expanded the domain of the functions M_n and S_n to $\Xi\times I\times[a,b]$ although $\Theta\times I\times[a,b]$ would suffice throughout this section. Additionally, we expanded the domain of the functions M and S to $\Xi\times I\times\bigcup_{\alpha\in[a,b]}\big(\Gamma(\alpha)\times\{\alpha\}\big)$.

This construction yields the estimator

$$\widehat{\theta}_n^{\mathrm{adap}}(x) = \hat{\theta}_n(x;\hat{\alpha}_n) = \operatorname*{argmin}_{\theta\in\Theta} M_n(\theta,x;\hat{\alpha}_n)\,, \tag{3.1.10}$$

that, under mild conditions, attains adaptivity from the empirical gradients by Theorem 3.1.3, i.e.

$$\lim_{\eta\to\infty}\limsup_{n\to\infty}\sup_{\alpha\in[a,b]}\sup_{\gamma\in\Gamma(\alpha)}\mathbb{P}_\gamma\left(\left(\frac{n}{\log n}\right)^{\frac{\alpha}{2\alpha+d}}\sup_{x\in I}\min_{\theta_*\in\mathfrak{S}_{x;\gamma;\alpha}}\left\|\widehat{\theta}_n^{\mathrm{adap}}(x)-\theta_*\right\| \ge \eta\right) = 0\,.$$

Let us discuss the intuition behind this construction in further detail. Under the assumption that $\Gamma(\alpha) \subset \Gamma(\tilde{\alpha})$ for $\alpha > \tilde{\alpha}$, which particularly implies

$$\bigcup_{\alpha \in [a,b]} \Gamma(\alpha) = \Gamma(a) \ ,$$

we make the following observations.

Remark 3.1.8. For some $\alpha \in [a,b]$ and $0 \leq k_n(\alpha) \leq N$ so that $\beta_{k_n(\alpha)} \leq \alpha \leq \beta_{k_n(\alpha)+1}$, we have:

(i) For a sequence (l_n) of grid points so that $l_n < k_n(\alpha)$ for all $n \in \mathbb{N}$, we have

$$\Gamma(\alpha) \subset \Gamma(\beta_{k_n(\alpha)}) \subset \Gamma(\beta_{l_n})$$

and thus, according to (3.1.4),

$$\sup_{\gamma \in \Gamma(\alpha)} \mathbb{E}_\gamma \Big[\sup_{x \in I, \theta \in \Theta} \big\| S_n(\theta, x; \beta_{l_n}) - S(\theta, x; \gamma; \alpha) \big\| \Big] \lesssim C^* r_{l_n} \ ,$$
$$\sup_{\gamma \in \Gamma(\alpha)} \mathbb{E}_\gamma \Big[\sup_{x \in I, \theta \in \Theta} \big\| S_n(\theta, x; \beta_{k_n(\alpha)}) - S(\theta, x; \gamma; \alpha) \big\| \Big] \lesssim C^* r_{k_n(\alpha)}$$

and hence

$$\begin{aligned} \sup_{\gamma \in \Gamma(\alpha)} \mathbb{E}_\gamma \Big[\sup_{x \in I, \theta \in \Theta} \big\| S_n(\theta, x; \beta_{l_n}) - S_n(\theta, x; \beta_{k_n(\alpha)}) \big\| \Big] &\lesssim C^* (r_{l_n} + r_{k_n(\alpha)}) \\ &\lesssim 2C^* r_{l_n} \ . \end{aligned}$$

This means that the sequence $k_n(\alpha)$ behaves like a deterministic asymptotic Lepski choice for $C_{\mathrm{Lep}} = C^*$ and the actual Lepski choice $\hat{k}$ should be at least $k_n(\alpha)$ with probability approaching one as $n \to \infty$. The detailed formalization and formal proof of this turns out to be rather difficult. It involves the aforementioned exponential deviation inequalities.

(ii) If the Lepski parameter is too big, i.e. for any $\gamma \in \Gamma(\alpha)$, we have $\hat{k}_n > k_n(\alpha)$ for $n \in \mathcal{J} \subset \mathbb{N}$, $\#\mathcal{J} = \infty$, then, by definition of $\hat{k}_n$, we have

$$\sup_{x \in I, \theta \in \Theta} \big\| S_n(\theta, x; \beta_{\hat{k}_n}) - S_n(\theta, x; \beta_{k_n(\alpha)}) \big\| \leq C_{\mathrm{Lep}} r_{k_n(\alpha)} \ , \quad n \in \mathcal{J}$$

as well as

$$\mathbb{E}_\gamma \Big[\sup_{x \in I, \theta \in \Theta} \big\| S_n(\theta, x; \beta_{k_n(\alpha)}) - S(\theta, x; \gamma; \alpha) \big\| \Big] \lesssim C^* r_{k_n(\alpha)} \ ,$$

because $\Gamma(\alpha) \subset \Gamma(\beta_{k_n(\alpha)})$, giving

$$\mathbb{E}_\gamma \Big[\sup_{x \in I, \theta \in \Theta} \big\| S_n(\theta, x; \beta_{\hat{k}_n}) - S(\theta, x; \gamma; \alpha) \big\| \Big]$$

$$\begin{aligned}
&\leq \mathbb{E}_\gamma\Big[\sup_{x\in I,\theta\in\Theta}\big\|S_n(\theta,x;\beta_{\hat{k}_n})-S_n(\theta,x;\beta_{k_n(\alpha)})\big\|\Big]\\
&\quad+\mathbb{E}_\gamma\big[\sup_{x\in I,\theta\in\Theta}\big\|S_n(\theta,x;\beta_{k_n(\alpha)})-S(\theta,x;\gamma;\alpha)\big\|\Big]\\
&\lesssim (C_{\mathrm{Lep}}+C)^* r_{k_n(\alpha)}
\end{aligned}$$

and as we will see in the proof of the following lemma, the sequences $r_{k_n(\alpha)}$ and $r(\alpha)$ are equivalent. This basically means that a too large Lepski parameter results in an alteration of the convergence rate that is asymptotically negligible.

Let us formalize our observations.

Lemma 3.1.9. *Let $0<a<b<\infty$, $\Gamma(\alpha)$, $\alpha\in[a,b]$ be sets, $\Gamma(\alpha)\subset\Gamma(\tilde{\alpha})$ whenever $\tilde{\alpha}<\alpha$; $\Theta\subset\mathbb{R}^m$, $I\subset\mathbb{R}^d$ and $S_n:\Theta\times I\times[a,b]$ be random functions so that* (3.1.4) *holds. Then, for any $C_{\mathrm{Lep}}>0$,*

$$\begin{aligned}
&\limsup_{n\to\infty}\sup_{\alpha\in[a,b]}\sup_{\gamma\in\Gamma(\alpha)} r(\alpha)^{-1}\mathbb{E}_\gamma\Big[\sup_{x\in I,\theta\in\Theta}\big\|S_n(\theta,x;\hat{\alpha}_n)-S(\theta,x;\gamma;\alpha)\big\|\Big]\\
&\leq (C_{\mathrm{Lep}}+C^*)\exp(d(b-a))\\
&\quad+C^*\limsup_{n\to\infty}\sup_{\alpha\in[a,b]}\sup_{\gamma\in\Gamma(\alpha)}\\
&\quad\sup_{a\leq\beta\leq\alpha} r(\alpha)^{-1}\Big(\mathbb{E}_\gamma\Big[\Big(\sup_{x\in I,\theta\in\Theta}\big\|S_n(\theta,x;\beta)-S(\theta,x;\gamma;\alpha)\big\|\Big)^2\Big]\Big)^{\frac{1}{2}}\cdot\log(n)^{\frac{3}{2}}\sup_{0\leq l\leq j\leq k_n(\alpha)} p_{lj}^{\frac{1}{2}},
\end{aligned}\tag{3.1.11}$$

where

$$p_{lj}=2\,\mathbb{P}_\gamma\Big(\sup_{x\in I,\theta\in\Theta}\big\|S_n(\theta,x;\beta_j)-\mathbb{E}_\gamma[S_n(\theta,x;\beta_j)]\big\|>\tfrac{C_{\mathrm{Lep}}-3C^{**}}{2}r_l\Big)\,.\tag{3.1.12}$$

In order to prove adaptivity, one needs to show that the second summand in (3.1.11) is finite. Whenever one has an appropriate exponential deviation inequality for the gradients S_n at hand and the sets Θ, I are compact, the terms p_{lj} converge to zero at polynomial rate with exponent tuned by the Lepski constant C_{Lep}. Two estimation problems in which this is true can be found in Sections 4.1 and 4.2. Techniques for deriving this can be found in Section 3.2.3. Note that the exponent will converge to ∞ for $C_{\mathrm{Lep}}\to\infty$. This means that it is enough to assume that the L^2-error above converges to ∞ with at most polynomial rate, if at all, i.e.

$$n^{-T}\sup_{\alpha\in[a,b]}\sup_{\gamma\in\Gamma(\alpha)}\mathbb{E}_\gamma\Big[\sup_{x\in I,\theta\in\Theta}\big\|S_n(\theta,x;\alpha)-S(\theta,x;\gamma;\alpha)\big\|^2\Big]<\infty$$

for some $T>0$. As the uniform L^1-error converges uniformly to zero at rate $\big(\frac{\log n}{n}\big)^{\frac{\alpha}{2\alpha+d}}$, it is reasonable to expect the L^2 error to converge to zero at the same rate, which is the assumption in the following lemma.

Lemma 3.1.10. *Let $0 < a < b < \infty$, $\Gamma(\alpha)$, $\alpha \in [a,b]$ be sets, $\Gamma(\alpha) \subset \Gamma(\tilde{\alpha})$ whenever $\tilde{\alpha} < \alpha$; $\Theta \subset \mathbb{R}^m$, $I \subset \mathbb{R}^d$ be compact and $S_n : \Theta \times I \times [a,b]$ be random functions so that the L^2 version of (3.1.4) holds, i.e.*

$$\limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} r(\alpha)^{-2}\, \mathbb{E}_\gamma\Big[\sup_{x\in I,\theta\in\Theta} \big\|S_n(\theta,x;\alpha) - S(\theta,x;\gamma;\alpha)\big\|^2\Big] \leq C^{**} < \infty\ . \tag{3.1.13}$$

Further, let there be a monotone function $u : [C_-,\infty) \to [0,\infty)$ with $u(t) \to \infty$, $t \to \infty$ and a constant $C_- > 0$ so that for every $C_{\mathrm{Lep}} \geq C_-$,

$$\limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \sup_{0\leq l\leq j\leq k_n(\alpha)} n^{u(C_{\mathrm{Lep}})} p_{lj} < \infty\ , \tag{3.1.14}$$

where p_{lj} is defined in (3.1.12) and depends on C_{Lep}. Then, for any $C_{\mathrm{Lep}} \geq C_-$ with $u(C_{\mathrm{Lep}})/2 > \frac{b}{2b+d} - \frac{a}{2a+d}$,

$$\begin{aligned}&\limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} r(\alpha)^{-1}\mathbb{E}_\gamma\Big[\sup_{x\in I,\theta\in\Theta} \big\|S_n(\theta,x;\hat{\alpha}_n) - S(\theta,x;\gamma;\alpha)\big\|\Big]\\ \leq&(C_{\mathrm{Lep}} + C^{**})\exp(d(b-a))\ .\end{aligned}$$

The function u typically depends on C^{**}, Θ, I, a, b, the exponential deviation inequality and possibly more model parameters. By Jensen's inequality, (3.1.13) implies (3.1.4) with $C^* = C^{**}$. The proofs of both lemmata are straightforward and given here. In Sections 3.2.1 and 3.2.2, we will give tools that help determining the uniform L^2 convergence rate for specific types of contrast functions. Techniques for treating the gradient deviation probabilities (3.1.14) are given in Section 3.2.3.

Proof of Lemma 3.1.9. Let for all $\alpha \in [a,b]$, $0 \leq k_n(\alpha) \leq N-1$ so that $\beta_{k_n(\alpha)} \leq \alpha \leq \beta_{k_n(\alpha)+1}$. Then we have for any $\alpha \in [a,b]$, $\gamma \in \Gamma(\alpha)$

$$\begin{aligned}&\mathbb{E}_\gamma\Big[\sup_{x\in I,\theta\in\Theta} \big\|S_n(\theta,x;\beta_{\hat{k}}) - S(\theta,x;\gamma;\alpha)\big\|\Big]\\ \leq&\mathbb{E}_\gamma\Big[\sup_{x\in I,\theta\in\Theta} \big\|S_n(\theta,x;\beta_{\hat{k}}) - S(\theta,x;\gamma;\alpha)\big\| \mathbb{1}_{\hat{k}\leq k_n(\alpha)-1}\Big] &(3.1.15)\\ &+ \mathbb{E}_\gamma\Big[\sup_{x\in I,\theta\in\Theta} \big\|S_n(\theta,x;\beta_{\hat{k}}) - S(\theta,x;\gamma;\alpha)\big\| \mathbb{1}_{\hat{k}\geq k_n(\alpha)}\Big]\ . &(3.1.16)\end{aligned}$$

The term (3.1.16) can be handled by a zero-addition of the term $S_n(\theta,x;\beta_{k_n(\alpha)})$ within the supremum according to our prior observations in Remark 3.1.8, i.e.

$$\begin{aligned}&\mathbb{E}_\gamma\Big[\sup_{x\in I,\theta\in\Theta} \big\|S_n(\theta,x;\beta_{\hat{k}}) - S(\theta,x;\gamma;\alpha)\big\| \mathbb{1}_{\hat{k}\geq k_n(\alpha)}\Big]\\ \leq&\,\mathbb{E}_\gamma\Big[\sup_{x\in I,\theta\in\Theta} \big\|S_n(\theta,x;\beta_{\hat{k}}) - S_n(\theta,x;\beta_{k_n(\alpha)})\big\| \mathbb{1}_{\hat{k}\geq k_n(\alpha)}\Big]\\ &+ \mathbb{E}_\gamma\Big[\sup_{x\in I,\theta\in\Theta} \big\|S_n(\theta,x;\beta_{k_n(\alpha)}) - S(\theta,x;\gamma;\alpha)\big\| \mathbb{1}_{\hat{k}\geq k_n(\alpha)}\Big]\end{aligned}$$

$$\begin{aligned}&\lesssim C_{\mathrm{Lep}} r_{k_n(\alpha)} + C^* r_{k_n(\alpha)} \\ &= (C_{\mathrm{Lep}} + C^*) r_{k_n(\alpha)} \ ,\end{aligned}$$

where we used that $\Gamma(\alpha) \subset \Gamma(\beta_{k_n(\alpha)})$.

Now let us show that the convergence rates $r_{k_n(\alpha)}$ and $r(\alpha)$ are asymptotically equivalent by deriving that for all n, α,

$$\begin{aligned}1 \leq \frac{r_{k_n(\alpha)}}{r(\alpha)} &= \left(\frac{n}{\log n}\right)^{\frac{\alpha}{2\alpha+d} - \frac{\beta_{k_n(\alpha)}}{2\beta_{k_n(\alpha)}+d}} \\ &= \left(\frac{n}{\log n}\right)^{\frac{\alpha d - \beta_{k_n(\alpha)} d}{(2\alpha+d)(2\beta_{k_n(\alpha)}+d)}} \\ &\leq \left(\frac{n}{\log n}\right)^{d(\alpha - \beta_{k_n(\alpha)})} \\ &\leq n^{d(\alpha - \beta_{k_n(\alpha)})} \\ &\leq n^{d\frac{b-a}{\log n}} \\ &= \exp(d(b-a)) < \infty \ ,\end{aligned}$$

where we used that the net over $[a, b]$ grows logarithmically. We get the desired bound on (3.1.16), i.e.

$$\begin{aligned}&\limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} r(\alpha)^{-1} \mathbb{E}_\gamma \Big[\sup_{x\in I, \theta\in\Theta} \big\| S_n(\theta, x; \beta_{\hat{k}}) - S(\theta, x; \gamma; \alpha) \big\| \mathbb{1}_{\hat{k} \geq k_n(\alpha)} \Big] \\ \leq &(C_{\mathrm{Lep}} + C^*) \exp(d(b-a)) \, .\end{aligned}$$

So let us examine (3.1.15). By using Cauchy-Schwarz' inequality, we get for all $\alpha \in [a, b]$, $\gamma \in \Gamma(\alpha)$ that

$$\begin{aligned}&\mathbb{E}_\gamma \Big[\sup_{x\in I, \theta\in\Theta} \big\| S_n(\theta, x; \beta_{\hat{k}}) - S(\theta, x; \gamma; \alpha) \big\| \mathbb{1}_{\hat{k} \leq k_n(\alpha)-1} \Big] \\ = &\sum_{j=0}^{k_n(\alpha)-1} \mathbb{E}_\gamma \Big[\sup_{x\in I, \theta\in\Theta} \big\| S_n(\theta, x; \beta_j) - S(\theta, x; \gamma; \alpha) \big\| \mathbb{1}_{\hat{k}=j} \Big] \\ \leq &\sum_{j=0}^{k_n(\alpha)-1} \Big(\mathbb{E}_\gamma \Big[\Big(\sup_{x\in I, \theta\in\Theta} \big\| S_n(\theta, x; \beta_j) - S(\theta, x; \gamma; \alpha) \big\| \Big)^2 \Big] \Big)^{\frac{1}{2}} \mathbb{P}_\gamma(\hat{k} = j)^{\frac{1}{2}} \\ \leq &\sup_{a\leq\beta\leq\alpha} \Big(\mathbb{E}_\gamma \Big[\Big(\sup_{x\in I, \theta\in\Theta} \big\| S_n(\theta, x; \beta) - S(\theta, x; \gamma; \alpha) \big\| \Big)^2 \Big] \Big)^{\frac{1}{2}} \cdot \sum_{j=0}^{k_n(\alpha)-1} \mathbb{P}_\gamma(\hat{k} = j)^{\frac{1}{2}} \ .\end{aligned}$$

By definition of $\hat{k}$, we have

$$\mathbb{P}_\gamma(\hat{k} = j) \leq \sum_{l=0}^{j} \mathbb{P}_\gamma \Big(\sup_{x\in I, \theta\in\Theta} \big\| S_n(\theta, x; \beta_{j+1}) - S_n(\theta, x; \beta_l) \big\| > C_{\mathrm{Lep}} r_l \Big)$$

$$\begin{aligned}
&\leq (j+1) \max_{l=0,\dots,j} \mathbb{P}_\gamma\Big(\sup_{x\in I,\theta\in\Theta} \big\|S_n(\theta,x;\beta_{j+1}) - S_n(\theta,x;\beta_l)\big\| > C_{\mathrm{Lep}} r_l \Big) \\
&\lesssim \log(n) \max_{l=0,\dots,j} \mathbb{P}_\gamma\Big(\sup_{x\in I,\theta\in\Theta} \big\|S_n(\theta,x;\beta_{j+1}) - S_n(\theta,x;\beta_l)\big\| > C_{\mathrm{Lep}} r_l \Big) ,
\end{aligned}$$

as the set of grid points grows logarithmically in n. Hence, we further deduce by index shifting that

$$\begin{aligned}
&\mathbb{E}_\gamma\Big[\sup_{x\in I,\theta\in\Theta} \big\|S_n(\theta,x;\beta_{\hat{k}}) - S(\theta,x;\gamma;\alpha)\big\| \mathbb{1}_{\hat{k}\leq k_n(\alpha)-1} \Big] \\
&\leq \sup_{a\leq\beta<\alpha} \Big(\mathbb{E}_\gamma\Big[\Big(\sup_{x\in I,\theta\in\Theta} \big\|S_n(\theta,x;\beta) - S(\theta,x;\gamma;\alpha)\big\| \Big)^2 \Big] \Big)^{\frac{1}{2}} \\
&\quad \cdot \log(n)^{\frac{3}{2}} \sup_{0\leq l<j\leq k_n(\alpha)} \mathbb{P}_\gamma\Big(\sup_{x\in I,\theta\in\Theta} \big\|S_n(\theta,x;\beta_j) - S_n(\theta,x;\beta_l)\big\| > C_{\mathrm{Lep}} r_l \Big)^{\frac{1}{2}}
\end{aligned}$$

In order to treat the last factor, we first observe that for $l<j$ we have $r_j \leq r_l$, yielding

$$\begin{aligned}
&\limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \sup_{0\leq l<j\leq k_n(\alpha)} r_l^{-1} \sup_{x\in I,\theta\in\Theta} \big\|\mathbb{E}_\gamma\big[S_n(\theta,x;\beta_l)\big] - \mathbb{E}_\gamma\big[S_n(\theta,x;\beta_j)\big]\big\| \\
&\leq \limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} r_l^{-1} \sup_{0\leq l<k_n(\alpha)} \sup_{x\in I,\theta\in\Theta} \big\|\mathbb{E}_\gamma\big[S_n(\theta,x;\beta_l)\big] - S(\theta,x;\gamma;\alpha)\big\| \\
&\quad + \limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \sup_{0\leq j\leq k_n(\alpha)} r_j^{-1} \sup_{x\in I,\theta\in\Theta} \big\|\mathbb{E}_\gamma\big[S_n(\theta,x;\beta_j)\big] - S(\theta,x;\gamma;\alpha)\big\| \\
&\leq 2 \limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} r(\alpha)^{-1} \sup_{x\in I,\theta\in\Theta} \big\|\mathbb{E}_\gamma\big[S_n(\theta,x;\alpha)\big] - S(\theta,x;\gamma;\alpha)\big\| \\
&\leq 2C^{**}
\end{aligned}$$

because $\Gamma(\alpha) \subset \Gamma(\beta_j)$. Hence, there is an $n_0 \in \mathbb{N}$ so that for all $n \geq n_0$, $0 \leq l < j \leq k_n(\alpha)$, we have

$$\sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} r_l^{-1} \sup_{x\in I,\theta\in\Theta} \big\|\mathbb{E}_\gamma\big[S_n(\theta,x;\beta_l)\big] - \mathbb{E}_\gamma\big[S_n(\theta,x;\beta_j)\big]\big\| \leq 3C^{**} .$$

Subsequently, deduce that for any $n \geq n_0$, $\alpha \in [a,b]$, $\gamma \in \Gamma(\alpha)$, $0 \leq l < j \leq k_n(\alpha)$, we have

$$\begin{aligned}
&\mathbb{P}_\gamma\Big(\sup_{x\in I,\theta\in\Theta} \big\|S_n(\theta,x;\beta_j) - S_n(\theta,x;\beta_l)\big\| > C_{\mathrm{Lep}} r_l \Big) \\
&\leq \mathbb{P}_\gamma\Big(\sup_{x\in I,\theta\in\Theta} \big\|S_n(\theta,x;\beta_j) - \mathbb{E}_\gamma\big[S_n(\theta,x;\beta_j)\big]\big)\big\| \\
&\quad + \sup_{x\in I,\theta\in\Theta} \big\|\mathbb{E}_\gamma\big[S_n(\theta,x;\beta_l)\big] - \mathbb{E}_\gamma\big[S_n(\theta,x;\beta_j)\big]\big\| \\
&\quad + \sup_{x\in I,\theta\in\Theta} \big\|S_n(\theta,x;\beta_l) - \mathbb{E}_\gamma\big[S_n(\theta,x;\beta_l)\big]\big\| > C_{\mathrm{Lep}} r_l \Big) \\
&\leq \mathbb{P}_\gamma\Big(\sup_{x\in I,\theta\in\Theta} \big\|S_n(\theta,x;\beta_j) - \mathbb{E}_\gamma\big[S_n(\theta,x;\beta_j)\big]\big\|
\end{aligned}$$

$$
\begin{aligned}
&\quad + \sup_{x\in I,\theta\in\Theta} \big\|S_n(\theta,x;\beta_j) - \mathbb{E}_\gamma\big[S_n(\theta,x;\beta_l)\big]\big\| > (C_{\mathrm{Lep}} - 3C^{**})r_l\Big) \\
&\le \mathbb{P}_\gamma\Big(\sup_{x\in I,\theta\in\Theta} \big\|S_n(\theta,x;\beta_j) - \mathbb{E}_\gamma\big[S_n(\theta,x;\beta_j)\big]\big\| > \tfrac{C_{\mathrm{Lep}}-3C^{**}}{2} r_l\Big) \\
&\quad + \mathbb{P}_\gamma\Big(\sup_{x\in I,\theta\in\Theta} \big\|S_n(\theta,x;\beta_l) - \mathbb{E}_\gamma\big[S_n(\theta,x;\beta_l)\big]\big\| > \tfrac{C_{\mathrm{Lep}}-3C^{**}}{2} r_l\Big) \\
&\le 2 \max_{i\in\{j,l\}} \mathbb{P}_\gamma\Big(\sup_{x\in I,\theta\in\Theta} \big\|S_n(\theta,x;\beta_i) - \mathbb{E}_\gamma\big[S_n(\theta,x;\beta_i)\big]\big\| > \tfrac{C_{\mathrm{Lep}}-3C^{**}}{2} r_l\Big) .
\end{aligned}
$$

□

Proof of Lemma 3.1.10. As all assumptions of Lemma 3.1.9 are fulfilled, we only need to prove that the second summand in (3.1.11) is zero, i.e. that

$$
0 = \limsup_{n\to\infty} \Bigg\{ \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \sup_{a\le\beta\le\alpha} r(\alpha)^{-1} \Big(\mathbb{E}_\gamma\Big[\Big(\sup_{x\in I,\theta\in\Theta} \big\|S_n(\theta,x;\beta) - S(\theta,x;\gamma;\alpha)\big\|\Big)^2\Big]\Big)^{\frac12} \tag{3.1.17}
$$

$$
\cdot \log(n)^{\frac32} \sup_{0\le l\le j\le k_n(\alpha)} p_{lj}^{\frac12} \Bigg\} . \tag{3.1.18}
$$

The factor (3.1.17) is asymptotically dominated by the rate $r(a)r(b)^{-1}$ as can be seen by inserting $r(\beta)r(\beta)^{-1}$ so that

$$
(3.1.17) \le r(a)r(b)^{-1} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} r(\alpha)^{-1} \Big(\mathbb{E}_\gamma\Big[\Big(\sup_{x\in I,\theta\in\Theta} \big\|S_n(\theta,x;\alpha) - S(\theta,x;\gamma;\alpha)\big\|\Big)^2\Big]\Big)^{\frac12} ,
$$

where the supremum is asymptotically bounded by C^{**} according to (3.1.13). The second factor (3.1.18) can be dealt with by

$$
\begin{aligned}
&\sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \log(n)^{\frac32} \sup_{0\le l\le j\le k_n(\alpha)} p_{lj}^{\frac12} \\
&\le \log(n)^{\frac32} n^{-u(C_{\mathrm{Lep}})/2} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \sup_{0\le l\le j\le k_n(\alpha)} n^{u(C_{\mathrm{Lep}})/2} p_{lj}^{\frac12}
\end{aligned}
$$

and as $u(C_{\mathrm{Lep}})/2 > \frac{b}{2b+d} - \frac{a}{2a+d}$, we get

$$
r(a)r(b)^{-1} \log(n)^{\frac32} n^{-u(C_{\mathrm{Lep}})/2} = \left(\frac{n}{\log n}\right)^{\frac{b}{2b+d}-\frac{a}{2a+d}} \log(n)^{\frac32} n^{-u(C_{\mathrm{Lep}})/2} = o(1) ,
$$

concluding the proof by (3.1.12). □

Extending adaptivity from the gradient to the estimator

Let us give a specific set of conditions under which the gradient $S_n(\cdot,\cdot;\hat\alpha_n)$ defined in (3.1.9) is in fact adaptive and the adaptivity extends to the estimator $\widehat\theta_n^{\mathrm{adap}}(\cdot)$ defined in (3.1.10). This set of assumptions extends Assumption 3.1.4 by assuming that most of

the conditions hold uniformly over the nuisance parameter α as well. We further have to assume uniform L^2 convergence of the gradients and an exponential deviation inequality for the gradients as pointed out in the previous section. The extension of the convergence rates works by using Theorem 3.1.3 once again.

Assumption 3.1.11. Let $0 < a < b < \infty$, $\Theta \subset \Xi \subset \mathbb{R}^m$, $I \subset \mathbb{R}^d$, $(\Gamma(\alpha), \|\cdot\|_\alpha)$ be normed spaces, $\alpha \in [a,b]$, $M : \Xi \times I \times \bigcup_{\alpha\in[a,b]} \big(\Gamma(\alpha) \times \{\alpha\}\big) \to \mathbb{R}$ be a deterministic function, $M_n : \Xi \times I \times [a,b] \to \mathbb{R}$ be random functions; β_k, $r(\alpha)$, $\hat{k}$ be defined as in (3.1.6), (3.1.7) and (3.1.8), respectively; $\hat{\alpha}_n = \beta_{\hat{k}}$. Continuity of functions taking γ as arguments is to be understood with respect to the maximum of norms in the other arguments and the norm $\|\cdot\|_a$.

(B1) Assume that Θ is compact and convex with $\Theta = \overline{\mathrm{int}(\Theta)}$, Ξ is open and convex, I is compact, and $(\Gamma(\alpha), \|\cdot\|_\alpha)$ are compactly nested spaces, i.e. $\Gamma(\alpha) \subset \Gamma(\alpha')$ and $\Gamma(\alpha)$ is compact with respect to $\|\cdot\|_{\alpha'}$ whenever $\alpha' < \alpha$. Furthermore, $\Gamma(a)$ is compact with respect to $\|\cdot\|_a$. Additionally, for any α, $\alpha_n \nearrow \alpha$, it holds that

$$\bigcap_{n\in\mathbb{N}} \Gamma(\alpha_n) = \Gamma(\alpha) \ .$$

Moreover, there is a constant $\bar{C} < \infty$ so that for all $\theta, \theta' \in \Theta$, there is an $l \in \mathbb{N}_0$ and $\bar{\theta}_1, \ldots, \bar{\theta}_l \in \Theta$ so that with $\theta = \bar{\theta}_0$, $\theta' = \bar{\theta}_{l+1}$, we have

$$\bar{\theta}_{k+1} - \bar{\theta}_k = c_k e_{j_k} \ , \quad k = 0, \ldots, l$$

for some unit vectors e_{j_k} and some coefficients $c_k \in \mathbb{R}$ with $\sum_{k=0}^{l} |c_k| \leq \bar{C}\|\theta - \theta'\|$.

(B2) There is a set of permutation functions

$$\mathfrak{Z} \subset \{\zeta : \{1, \ldots, m\} \to \{1, \ldots, m\} : \zeta \text{ bijective}\} \ ,$$

so that M and M_n are invariant under permuting the first argument by $\zeta \in \mathfrak{Z}$.

(B3) The function M is continuous, i.e. the map

$$(\theta, x; \gamma; \alpha) \mapsto M(\theta, x; \gamma; \alpha)$$

is continuous. For every $x \in I$, $\alpha \in [a,b]$, $\gamma \in \Gamma(a)$, the contrast $M(\cdot, x; \gamma; \alpha)$ attains a minimum at $\theta_*(x; \gamma; \alpha)$ iff

$$\theta_*(x; \gamma; \alpha) \in \mathfrak{S}_{x;\gamma;\alpha} \ ,$$

where $\mathfrak{S}_{x;\gamma;\alpha} = \zeta(\mathfrak{S}_{x;\gamma;\alpha}) \subset \Theta$, $\zeta \in \mathfrak{Z}$ and $\#\mathfrak{S}_{x;\gamma;\alpha} = \#\mathfrak{Z}$. Furthermore, for any $x \in I$, $\alpha \in [a,b]$, $\gamma \in \Gamma(\alpha)$, $\theta_*(x; \gamma; \alpha) \in \mathfrak{S}_{x;\gamma;\alpha}$, there is a permutation $\zeta_{x;\alpha;\gamma} \in \mathfrak{Z}$ so that the maps $(x; \gamma; \alpha) \mapsto \zeta_{x;\alpha;\gamma}(\theta_*(x; \gamma; \alpha))$ are continuous.

(B4) For all $x \in I$, $\alpha \in [a,b]$, $\gamma \in \Gamma(a)$, the function $M(\cdot, x; \gamma; \alpha)$ is twice continuously differentiable in its first argument and the Hessian matrix

$$V_x\big(\theta_*(x; \gamma; \alpha); \gamma; \alpha\big) := \partial^2_{\theta^2} M\big(\theta_*(x; \gamma; \alpha), x; \gamma; \alpha\big)$$

is positive definite for all $\theta_*(x;\gamma;\alpha) \in \mathfrak{S}_{x;\gamma;\alpha}$. Particularly, the eigenvalues $\lambda^1_{x,\gamma;\alpha} \geq \cdots \geq \lambda^m_{x,\gamma;\alpha}$ of $V_x\big(\theta_*(x;\gamma;\alpha);\gamma;\alpha\big)$ are positive. Furthermore, the map

$$(x;\gamma;\alpha) \mapsto V_x\big(\theta_*(x;\gamma;\alpha);\gamma;\alpha\big)$$

is continuous.

(B5) The Hessian matrices $V_x(\cdot;\gamma;\alpha)$ are uniformly Lipschitz continuous in θ, i.e. for all $\theta, \theta' \in \Xi$, we have

$$\sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(a)} \sup_{x\in I} \big\|V_x(\theta;\gamma;\alpha) - V_x(\theta';\gamma;\alpha)\big\| \leq L_{\mathrm{Hess}}\|\theta-\theta'\| \,,$$

where $L_{\mathrm{Hess}} < \infty$ depends only on Ξ, I, a, b and $\Gamma(a)$.

(B6) There is an $\varepsilon^* > 0$ so that for all $0 < \varepsilon < \varepsilon^*$, $x \in I$, $\gamma \in \Gamma(a)$, $\alpha \in [a,b]$ the balls $\{\theta \in \Xi : \|\theta - \theta_*\| \leq \varepsilon\}$, $\theta_* \in \mathfrak{S}_{x;\gamma;\alpha}$ are disjoint.

(B7) The empirical contrast M_n is continuously differentiable in its first argument and for the gradients

$$S_n(\theta,x;\alpha) := \partial_\theta M_n(\theta,x;\alpha) \,, \quad S(\theta,x;\gamma;\alpha) := \partial_\theta M(\theta,x;\gamma;\alpha)$$

it holds that

$$\limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} r(\alpha)^{-2}\mathbb{E}_\gamma\bigg[\sup_{x\in I,\theta\in\Theta} \big\|S_n(\theta,x;\alpha) - S(\theta,x;\gamma;\alpha)\big\|^2\bigg] \leq C^{**} < \infty.$$

(B8) The empirical contrast M_n is uniformly consistent for M, i.e.

$$\lim_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(a)} \mathbb{P}_\gamma\bigg(\sup_{x\in I,\theta\in\Theta} \big|M_n(\theta,x;\alpha) - M(\theta,x;\gamma;a)\big| \geq \varepsilon\bigg) = 0 \,, \quad \varepsilon > 0 \,.$$

(B9) There is a constant $C_- > 0$ and a monotone function $u : [C_-,\infty) \to \big(1,\infty\big)$ with $u(t) \to \infty$, $t \to \infty$ so that for every $C_{\mathrm{Lep}} \geq C_-$,

$$\limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \sup_{0\leq l\leq j\leq k_n(\alpha)} n^{u(C_{\mathrm{Lep}})} p_{lj} < \infty \,,$$

where p_{lj} is defined in (3.1.12) and $0 \leq k_n(\alpha) \leq N-1$ with $N = \lceil \log n \rceil$ is chosen so that $\beta_{k_n(\alpha)} \leq \alpha \leq \beta_{k_n(\alpha)} + 1$.

Remark 3.1.12.

(i) Assumptions **(B1)** - **(B8)** imply Assumption 3.1.4 for any $\alpha \in [a,b]$ so that for any α, the estimators $\hat{\theta}_n(\cdot;\alpha) \in \operatorname{argmin}_{\theta\in\Theta} M_n(\theta,\cdot;\alpha)$ have convergence rate $r(\alpha)$.

(ii) In an identifiable model, the assumptions reduce in the same way as Assumption 3.1.4 does. A complete list of altered assumptions can be found in the Appendix, cf. Assumption A.2.2.

(iii) Note that in **(B8)**, we assume that every empirical contrast is consistent for all asymptotic contrasts independently of the actual nuisance parameter. In most estimation problems, this automatically holds when

$$\lim_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \mathbb{P}_\gamma\Big(\sup_{x\in I,\theta\in\Theta} \big|M_n(\theta,x;\alpha) - M(\theta,x;\gamma;\alpha)\big| \geq \varepsilon \Big) = 0 \,, \quad \varepsilon > 0 \,,$$

as adjusting the nuisance parameter is in most cases only used in order to tweak convergence rates, not to ensure consistency. Consider for example kernel density estimators that are typically uniformly consistent whenever the bandwidth h fulfils

$$h \to 0 \,, \qquad \frac{\log n}{nh^d} \to 0 \,,$$

so that varying h depending on nuisance parameters will not ruin consistency as long as the displayed constraints are not violated.

Theorem 3.1.13. *Under Assumption 3.1.11, for any $C_{\mathrm{Lep}} \geq C_-$ with $u(C_{\mathrm{Lep}}) > \frac{b}{2b+d} - \frac{a}{2a+d}$, the estimator $\widehat{\theta}_n^{\mathrm{adap}}(\cdot)$ defined in* (3.1.10) *is a uniformly adaptive estimator for $\theta_* \in \mathfrak{S}_{x;\gamma;\alpha}$, i.e.*

$$\lim_{\eta\to\infty} \limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \mathbb{P}_\gamma\Big(r(\alpha)^{-1} \sup_{x\in I} \min_{\theta_*\in\mathfrak{S}_{x;\gamma;\alpha}} \big\|\widehat{\theta}_n^{\mathrm{adap}}(x) - \theta_*\big\| \geq \eta \Big) = 0 \,.$$

Proof of Theorem 3.1.13. The proof works essentially analogous to the one of Theorem 3.1.6. Let us apply Theorem 3.1.3 for $\Gamma = \big\{(\alpha,\gamma) : \alpha \in [a,b], \gamma \in \Gamma(\alpha)\big\}$, which is compact with respect to $\max\{|\cdot|, \|\cdot\|_a\}$ as can be seen by a simple topological argument, cf. Lemma 3.1.14. Further use $r_{n,\gamma} = t_{n,\gamma} = r(\alpha)^{-1}$, $\phi_n = \mathrm{id}$, $\eta = \varepsilon^*$.

The first point of (i) is proved in the same way as for Theorem 3.1.6. The second part of (i) is also proved accordingly, one only has to exchange the uniform L^1-error of the gradients with the adaptive version attained by Lemma 3.1.10, i.e.

$$\begin{aligned} &\limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} r(\alpha)^{-1} \mathbb{E}_\gamma\Big[\sup_{x\in I,\theta\in\Theta} \big\|S_n(\theta,x;\hat{\alpha}_n) - S(\theta,x;\gamma;\alpha)\big\| \Big] \\ \leq &(C_{\mathrm{Lep}} + C^{**}) \exp(d(b-a)) \,. \end{aligned}$$

In order to prove uniform consistency of the estimator $\widehat{\theta}_n^{\mathrm{adap}}(\cdot)$, we first use **(B8)**, yielding

$$\lim_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \mathbb{P}_\gamma\Big(\sup_{x\in I,\theta\in\Theta} \big|M_n(\theta,x;\hat{\alpha}_n) - M(\theta,x;\gamma;\alpha)\big| \geq \varepsilon \Big)$$

$$\leq \lim_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(a)} \mathbb{P}_\gamma\Big(\sup_{x\in I,\theta\in\Theta} \big|M_n(\theta,x;\alpha) - M(\theta,x;\gamma;a)\big| \geq \varepsilon \Big) = 0 \;, \quad \varepsilon > 0$$

and then proceed analogously to the proof of Theorem 3.1.6

□

Lemma 3.1.14. *Under Assumption* **(B1)**, *the set* $\Gamma = \big\{(\alpha,\gamma) : \alpha \in [a,b], \gamma \in \Gamma(\alpha)\big\}$ *is compact with respect to* $\max\{|\cdot|, \|\cdot\|_a\}$.

Proof of Lemma 3.1.14. As $[a,b]\times\Gamma(a)$ is compact with respect to $\max\{|\cdot|,\|\cdot\|_a\}$, it is enough to show that Γ is a closed subset thereof. Let $\big((\alpha_n,\gamma_n)\big)_n \subset \Gamma$ converge to some $(\alpha_*,\gamma_*) \in [a,b]\times\Gamma(a)$. If there is a subsequence (n_k) so that $\alpha_{n_k} \geq \alpha_*$ for all $k\in\mathbb{N}$, then

$$\gamma_{n_k} \in \Gamma(\alpha_{n_k}) \subset \Gamma(\alpha_*) \;, \quad \text{for all } k \in \mathbb{N}$$

and since $\Gamma(\alpha_*)$ is closed with respect to $\|\cdot\|_a$, we have

$$\gamma_* = \lim_{k\to\infty} \gamma_{n_k} = \lim_{n\to\infty} \gamma_n \in \Gamma(\alpha_*) \;.$$

Hence, assume that there is an $n_* \in \mathbb{N}$ so that for all $n \geq n_*$, we have $\alpha_n < \alpha_*$. Without loss of generality assume $\alpha_n \nearrow \alpha_*$. Then, for any $\tilde{n}\in\mathbb{N}$ and any $n \geq \tilde{n}$, we have

$$\gamma_n \in \Gamma(\alpha_n) \subset \Gamma(\alpha_{\tilde{n}}) \;,$$

so that particularly $\gamma_* \in \Gamma(\alpha_{\tilde{n}})$ for all $\tilde{n}\in\mathbb{N}$. Since $\bigcap_{n\in\mathbb{N}} \Gamma(\alpha_n) = \Gamma(\alpha_*)$, the assertion follows. □

3.2 Techniques for examining uniform estimation errors

Let us describe some techniques for bounding uniform estimation errors often occurring in local M-estimation problems. These methods will be applied in Chapter 4 to mixture of regressions models.

3.2.1 Non-stochastic errors

Hölder smoothness constraints on the function to estimate are particularly useful when deriving bounds on the bias of an estimator. Consider for example a kernel density estimator $\hat{\ell}_n$, cf. Section 2.5, estimating any probability density $\ell : \mathbb{R}^d \to [0,\infty)$, i.e. $\int \ell = 1$. The bias of $\hat{\ell}_n$ can always be described by

$$\begin{aligned} \mathbb{E}\big[\hat{\ell}_n(x)\big] - \ell(x) = \Big(K_h * \ell\Big)(x) - \ell(x) &= \int \big(\ell(x+z) - \ell(x)\big) K_h(z)\,\mathrm{d}z \\ &= \int \big(\ell(x+hz) - \ell(x)\big) K(z)\,\mathrm{d}z \;, \quad x\in\mathbb{R}^d \end{aligned} \tag{3.2.1}$$

because $\int K_h = 1$. Integrals like the one in (3.2.1) are particularly easy to treat when ℓ comes from a Hölder class and one uses a kernel of higher order, cf. Definition 2.5.1.

Lemma 3.2.1. *Let $0 < a \leq b < \infty$, $K : \mathbb{R}^d \to \mathbb{R}$ be a kernel of order b with support $[-1,1]^d$; $I \subset \mathbb{R}^d$, $U \subset \mathbb{R}$ be compact with $I = \overline{\operatorname{int}(I)}$ as well as $L > 0$. Then, for any compact cuboid $J \subset \operatorname{int}(I)$ containing an open subset, there is some constant $0 < C_{\mathrm{Hol}} < \infty$ depending only on $[a,b]$, L, U and K so that*

$$\sup_{\alpha\in[a,b]} \sup_{h\in(0,\infty)} h^{-\alpha} \sup_{\ell\in H(\alpha,L,U)} \sup_{x\in J} \left| \int \big(\ell(x) - \ell(x+hz)\big) K(z)\,\mathrm{d}z \right| \leq C_{\mathrm{Hol}} \,.$$

Proof. Fix any $\ell \in H(\alpha,L,U)$, $x \in J$, $\alpha \in [a,b]$ and $h \in (0,\infty)$. Using the Taylor expansion of order $\lfloor\alpha\rfloor$ of ℓ around x and using that K is a kernel of order b, we get for some $\tau \in [0,1]$ and independently of ℓ, x, α, n and h that

$$\begin{aligned}
&\left| \int K(z)\big(\ell(hz+x) - \ell(x)\big)\,\mathrm{d}z \right| \\
&\leq \left| \sum_{|k|\in\{1,\ldots,\lfloor\alpha\rfloor-1\}} \frac{h^{|k|}}{k!} \partial^k \ell(x) \underbrace{\int K(z) z^k \,\mathrm{d}z}_{=0} \right| + \left| \sum_{|k|=\lfloor\alpha\rfloor} \frac{h^{\lfloor\alpha\rfloor}}{k!} \int z^k K(z) \partial^k \ell(x+\tau hz)\,\mathrm{d}z \right| \\
&= \left| \sum_{|k|=\lfloor\alpha\rfloor} \frac{h^{\lfloor\alpha\rfloor}}{k!} \int z^k K(z) \Big(\partial^k \ell(x+\tau hz) - \partial^k \ell(x)\Big)\,\mathrm{d}z \right| \\
&\leq \sum_{|k|=\lfloor\alpha\rfloor} \frac{L h^\alpha \tau^{\alpha-\lfloor\alpha\rfloor}}{k!} \int \|z\|^\alpha |K(z)|\,\mathrm{d}z \\
&= \frac{L d^{\lfloor\alpha\rfloor} \tau^{\alpha-\lfloor\alpha\rfloor} h^\alpha}{\lfloor\alpha\rfloor!} \int \|z\|^\alpha |K(z)|\,\mathrm{d}z \\
&\leq \frac{L d^b h^\alpha}{\lfloor a\rfloor!} \int \|z\|^a |K(z)|\,\mathrm{d}z \\
&\leq C_{\mathrm{Hol}} h^\alpha
\end{aligned}$$

because according to the multinomial theorem, we have

$$\sum_{\substack{0\leq k_1,\ldots,k_d\leq m \\ k_1+\ldots+k_d=m}} \frac{1}{k_1!\ldots k_d!} = \frac{1}{m!} \cdot \underbrace{(1+\ldots+1)}_{d \text{ times}}^m \,.$$

□

3.2.2 Uniform stochastic errors

Exponential deviation inequalities

In order to examine uniform stochastic errors, one often discretizes the supremum and uses exponential deviation inequalities of pointwise errors. The following three inequalities are used throughout this thesis.

The first exponential deviation inequality is Bernstein's well-known inequality. For a proof cf. Pollard (2012). This inequality gives an exponential bound on the tail probability of sums of uniformly bounded centred random variables.

Lemma 3.2.2 (Bernstein's inequality)**.** *Let $X_1, \dots, X_n$ be independent centred random variables with finite second moments so that $|X_j| \leq R$ for some $R > 0$. Let $S_n = \sum_{j=1}^n X_j$, then*

$$\mathbb{P}\big(|S_n| \geq t\big) \leq 2 \exp\left(- \frac{t^2}{2 \operatorname{Var}(S_n) + \frac{3}{2} t R} \right), \quad t > 0 .$$

While Bernstein's inequality is a powerful tool, it cannot be applied when the random variables are not uniformly bounded. This is often a problem when dealing with log-likelihood functions because the likelihood approaching zero translates to the log-likelihood approaching $-\infty$. The following inequality drops the boundedness constraint by imposing constraints on all moments of the random variables. It was introduced by Bennett (1962, p. 37-38).

Lemma 3.2.3 (Bennett's inequality)**.** *Let $X_1, \dots, X_n$ be independent centred random variables so that $\mathbb{E}\big[|X_j|^l\big] \leq l! M^{l-2} v_j / 2$, for every $l \geq 2$ and all j and some constants M, v_j. Let $S_n = \sum_{j=1}^n X_j$ and $v = v_1 + \ldots + v_n$, then*

$$\mathbb{P}\big(|S_n| \geq t\big) \leq 2 \exp\left(- \frac{1}{2} \frac{t^2}{v + Mt} \right), \quad t > 0 .$$

The third inequality is a Bernstein-type inequality for canonical U-statistics, also called "degenerate U-statistics", that can be found in Giné et al. (2000, p. 15). We use a slightly weaker version.

Lemma 3.2.4 (Giné et al. (2000))**.** *Let $(X_n)_n$ be a sequence of i.i.d. $\mathbb{R}^d$-valued random variables, defining a canonical U-statistic U_n with bounded canonical kernel $\chi : \mathbb{R}^{2d} \to \mathbb{R}$, i.e. for all $x, y \in \mathbb{R}^d$*

$$U_n = \sum_{\substack{j,k=1 \\ j \neq k}}^n \chi(X_j, X_k) , \qquad \chi(x,y) = \chi(y,x) , \qquad \mathbb{E}[\chi(X_1, x)] = \int_{\mathbb{R}^d} \chi(z, x) \, \mathrm{d}\mathbb{P}_{X_1}(z) = 0 .$$

Then there is a universal constant $T > 0$ so that for any $\omega > 0$, we have

$$\mathbb{P}\big(|U_n| > \omega\big) \leq T \exp\left(- T^{-1} \min\left\{ \frac{\omega}{C}, \left(\frac{\omega}{B}\right)^{\frac{2}{3}}, \left(\frac{\omega}{A}\right)^{\frac{1}{2}} \right\} \right),$$

where

$$A := \|\chi\|_\infty, \qquad B^2 := n \big\| \mathbb{E}\big[\chi^2(X_1, \cdot)\big] \big\|_\infty, \qquad C^2 := n(n-1) \mathbb{E}\big[\chi^2(X_1, X_2)\big] . \quad (3.2.2)$$

Uniform stochastic L^ρ-errors

The following two theorems are convergence rate results for the uniform L^ρ-error of random processes of the shape

$$M_n(\theta, x; h) := \frac{1}{n} \sum_{k=1}^{n} \tau(Y_k, \theta) K_h(X_k - x) \ , \quad \text{or} \tag{3.2.3}$$

$$M_n(\theta, x; h) := \frac{1}{n(n-1)} \sum_{\substack{j,k=1 \\ j \neq k}}^{n} \tau(Y_j, Y_k, \theta) K_h(X_j - x) K_h(X_k - x) \ , \tag{3.2.4}$$

where τ is a sufficiently smooth function, K is some kernel function and h some bandwidth parameter and the distributions of the Y_k and X_k depend on some parameter γ. The desired result has the form

$$\limsup_{n \to \infty} \left(\frac{\log n}{n h^d} \right)^{-\frac{\rho}{2}} \mathbb{E}_\gamma \left[\sup_{x \in J} \sup_{\theta \in \Theta} \left| M_n(\theta, x; h) - \mathbb{E}_\gamma[M_n(\theta, x; h)] \right|^\rho \right] \leq C < \infty \ ,$$

uniformly over the parameter γ and variations of bandwidth sequences h for some $C < \infty$, where $J \subset \mathrm{int}(I)$. For both types of random processes M_n, the general scheme of proof is identical, the details differ to some degree due to the differences in structure.

First, one discretizes the estimation error by laying a grid $\Theta_n \times J_n$ on the compact space $\Theta \times J$ getting finer at some rate $\delta_n \to 0$ that is used for balancing purposes. Write $T_n = M_n - \mathbb{E}_\gamma[M_n]$ for the centred process. The discretization error

$$\mathbb{E}_\gamma \left[\Big| \sup_{x \in J, \theta \in \Theta} |T_n(\theta, x; h)| - \sup_{y \in J_n, \vartheta \in \Theta_n} |T_n(\vartheta, y; h)| \Big|^\rho \right] \tag{3.2.5}$$

gets small fast enough by assuming that the function τ is Lipschitz continuous and bounded or more general versions thereof if necessary. As it turns out, being Lipschitz continuous with integrable Lipschitz constant can replace Lipschitz continuity. An exponential deviation inequality holding for the centred process T_n and uniform integrability of τ can replace boundedness.

The exponential inequality on the tail probability is used in order to get bounds on the discrete estimation error in the form

$$\begin{aligned}
&\mathbb{E}_\gamma \left[\sup_{x \in J_n, \theta \in \Theta_n} |T_n(\theta, x; h)|^\rho \right] \\
\leq &\Big(\frac{\log n}{n h^d} \Big)^{\frac{\rho}{2}} \left[a^\rho + \int_a^\infty \rho \omega^{\rho - 1} \mathbb{P}_\gamma \Big(\sup_{x \in J_n, \theta \in \Theta_n} |T_n(\theta, x; h)| > \Big(\frac{\log n}{n h^d} \Big)^{\frac{1}{2}} \omega \Big) \, \mathrm{d}\omega \right] \\
\leq &\Big(\frac{\log n}{n h^d} \Big)^{\frac{\rho}{2}} \left[a^\rho + C_{\Theta, J} \delta_n^{-d-m} \rho \int_a^\infty \omega^{\rho - 1} n^{-C_a \omega} \, \mathrm{d}\omega \right] ,
\end{aligned}$$

where $C_{\Theta,J}$ depends only on Θ and J. The terms δ_n, a and C_a need to be balanced in a way so that the second summand and the discretization error (3.2.5) are negligible.

Note that for the U-statistic process M_n defined in (3.2.4), one would typically decompose the centred process into a canonical U-statistic and a linear process before discretizing. By using the notation $Z_j = (Y_j, X_j^T)^T$, write for any fixed $x \in I$, $\theta \in \Theta$,

$$M_n(\theta, x; h) - \mathbb{E}_\gamma[M_n(\theta, x; h)] = \frac{1}{n(n-1)} \sum_{\substack{j,k=1 \\ j \neq k}}^{n} U_n\big(Z_j, Z_k, \theta, x; h\big) \tag{3.2.6}$$

$$+ \frac{2}{n} \sum_{j=1}^{n} u_n^*(Z_j, \theta, x; h) - \mathbb{E}_\gamma[u_n(Z_1, Z_2, \theta, x; h)] \,, \tag{3.2.7}$$

where for $z = (z_1, z_2^T)^T, y = (y_1, y_2^T)^T \in \mathbb{R} \times \mathbb{R}^d$

$$\begin{aligned} U_n(z, y, \theta, x; h) :=\, & u_n(z, y, \theta, x; h) - u_n^*(z, \theta, x; h) - u_n^*(y, \theta; x; h) \\ & + \mathbb{E}_\gamma[u_n(Z_1, Z_2, \theta, x; h)] \,, \end{aligned} \tag{3.2.8}$$

$$u_n(z, y, \theta, x; h) := \tau(z_1, y_1, \theta) K_h(z_2 - x) K_h(y_2 - x) \,, \tag{3.2.9}$$

$$u_n^*(z, \theta, x; h) := \mathbb{E}_\gamma[u_n(Z_1, z, \theta, x; h)] = \mathbb{E}_\gamma\big[\tau(z_1, Y_1, \theta) K_h(X_1 - x)\big] \cdot K_h(z_2 - x) \,. \tag{3.2.10}$$

Now, (3.2.6) is a canonical U-statistic as

$$\mathbb{E}_\gamma[U_n(Z, z, \theta, x; h)] = 0 \,, \quad z \in \mathbb{R} \times \mathbb{R}^d$$

to which Giné's inequality, Lemma 3.2.4, is usually applicable. On the other hand, (3.2.7) is a centred linear process that can usually be treated by one of the inequalities given in Lemmata 3.2.2 and 3.2.3 as described before.

The first theorem gives a result on functions M_n of the type (3.2.3).

Theorem 3.2.5. *Let Γ be a non-empty set, $\big(\Omega, \mathcal{A}, (\mathbb{P}_\gamma)_{\gamma \in \Gamma}\big)$ be a statistical model and probability densities be given by*

$$(y, x) \mapsto f_\gamma(y|x) \ell_\gamma(x) \,, \qquad (y, x) \in \mathbb{R} \times I \,, \gamma \in \Gamma \,, \tag{3.2.11}$$

where $I \subset \mathbb{R}^d$ is a compact cuboid containing an open subset and is the support of the ℓ_γ as well as $\sup_{\gamma \in \Gamma} \|\ell_\gamma\|_\infty < \infty$. Furthermore, let $\big((Y_n, X_n^T)^T\big)_n$ be sequences of i.i.d. random vectors with joint density (3.2.11) under $\mathbb{P}_\gamma$.

Let $K : \mathbb{R}^d \to \mathbb{R}$ be a Lipschitz continuous and bounded L^2-kernel; for some non-empty set A, $(h_n(\alpha))_{n \in \mathbb{N}}$, $\alpha \in A$ be sequences of bandwidth parameters so that

$$\sup_{\alpha \in A} h_n(\alpha) \to 0 \,, \quad \sup_{\alpha \in A} \frac{\log n}{n h_n(\alpha)^d} \to 0 \,.$$

Let $\tau : \mathbb{R} \times \Theta \to \mathbb{R}$ be a function, where $\Theta \subset \mathbb{R}^m$ is compact and convex with $\Theta = \overline{\mathrm{int}(\Theta)}$ and $\rho \in [1, \infty)$ so that

$$\sup_{\gamma \in \Gamma} \int (\sup_{\theta \in \Theta} |\tau(y, \theta)|)^\rho \sup_{x \in I} f_\gamma(y|x) \, \mathrm{d}y =: C_\tau < \infty \tag{3.2.12}$$

and

$$\big|\tau(y, \theta) - \tau(y, \tilde{\theta})\big| \leq \Psi_\tau(y, \theta, \tilde{\theta}) \|\theta - \tilde{\theta}\| \, , \quad y \in \mathbb{R}, \theta, \tilde{\theta} \in \Theta \, , \tag{3.2.13}$$

where Ψ_τ is a non-negative function so that

$$\sup_{\gamma \in \Gamma} \int \sup_{\theta, \tilde{\theta} \in \Theta} \Psi_\tau^\rho(y, \theta, \tilde{\theta}) \sup_{x \in I} f_\gamma(y|x) \, \mathrm{d}y < \infty \, . \tag{3.2.14}$$

If for the function $M_n : \Theta \times I \times (0, \infty) \to \mathbb{R}$ given by

$$M_n(\theta, x; h) := \frac{1}{n} \sum_{k=1}^{n} \tau(Y_k, \theta) K_h(X_k - x) \, ,$$

there are constants $C_1, C_2 < \infty$ independent of θ, x, h, γ, so that

$$\mathbb{P}_\gamma\Big(\big|M_n(\theta, x; h) - \mathbb{E}_\gamma\big[M_n\big(\theta, x; h\big)\big]\big| \geq t\Big) \leq 2 \exp\left(- \frac{t^2 n h^d}{C_1 + C_2 t} \right) , \quad t > 0 \, , \tag{3.2.15}$$

then for any compact $J \subset \mathrm{int}(I)$

$$\limsup_{n \to \infty} \sup_{\gamma \in \Gamma} \sup_{\alpha \in A} \left(\frac{\log n}{n h_n(\alpha)^d} \right)^{-\frac{\rho}{2}} \mathbb{E}_\gamma \left[\sup_{x \in J} \sup_{\theta \in \Theta} \big| M_n\big(\theta, x; h_n(\alpha)\big) - \mathbb{E}_\gamma\big[M_n\big(\theta, x; h_n(\alpha)\big)\big] \big|^\rho \right] \leq C \, ,$$

where $C < \infty$ depends on Γ, C_1, C_2, C_τ, τ, $\|K\|_\infty$, L_K, ρ, I, Θ but is free from n and the sequences of bandwidth parameters.

Remark 3.2.6.

(i) Note that assumption (3.2.15) is typically achieved by applying one of the exponential deviation inequalities given by Lemmata 3.2.2 and 3.2.3. In order to use Lemma 3.2.2, one might assume boundedness of τ as well as

$$\sup_{\alpha \in A} \sup_{\gamma \in \Gamma} \sup_{\theta \in \Theta, x \in I} n h_n(\alpha)^d \mathrm{Var}_\gamma\big(M_n(\theta, x; h_n(\alpha))\big) \leq C_{\mathrm{Var}} \, , \quad C_{\mathrm{Var}} < \infty \, .$$

(ii) Assumption (3.2.13) is fulfilled if for example the function τ is Lipschitz continuous in its second argument uniformly over its first argument, i.e. $\sup_y |\tau(y, \vartheta) - \tau(y, \theta)| \leq L_\tau \|\vartheta - \theta\|$ for some $L_\tau < \infty$.

The second theorem gives a result on functions M_n of the type (3.2.4).

Theorem 3.2.7. *Let Γ be a non-empty set, $\big(\Omega, \mathcal{A}, (\mathbb{P}_\gamma)_{\gamma\in\Gamma}\big)$ be a statistical model and probability densities be given by*

$$(y,x) \mapsto f_\gamma(y|x)\ell_\gamma(x)\ , \qquad (y,x) \in \mathbb{R}\times I\ , \gamma\in\Gamma\ , \tag{3.2.16}$$

where $I \subset \mathbb{R}^d$ is a compact cuboid containing an open subset and is the support of the ℓ_γ as well as $\sup_{\gamma\in\Gamma}\|\ell_\gamma\|_\infty < \infty$. Furthermore, let $(Z_n)_n = \big((Y_n, X_n^T)^T\big)_n$ be sequences of i.i.d. random vectors with joint density (3.2.16) *under $\mathbb{P}_\gamma$.*

Let $K : \mathbb{R}^d \to \mathbb{R}$ be a Lipschitz continuous and bounded L^2-kernel; for some non-empty set A, $(h_n(\alpha))_{n\in\mathbb{N}}$, $\alpha \in A$ be sequences of bandwidth parameters so that

$$\sup_{\alpha\in A} h_n(\alpha) \to 0\ , \quad \sup_{\alpha\in A}\frac{\log n}{nh_n(\alpha)^d} \to 0\ .$$

Let $\tau : \mathbb{R}\times\mathbb{R}\times\Theta \to [0,\infty)$ be a bounded function that is symmetric in its first two arguments as well as Lipschitz continuous in its third argument uniformly over all other arguments, i.e.

$$\sup_{z,y}|\tau(z,y,\vartheta) - \tau(z,y,\theta)| \le L_\tau\|\vartheta-\theta\|\ ,$$

where $\Theta \subset \mathbb{R}^m$ is compact and convex with $\Theta = \overline{\mathrm{int}(\Theta)}$ and $L_\tau > 0$ is a constant. Then, for the function $M_n : \Theta\times I\times(0,\infty) \to [0,\infty)$ given by

$$M_n(\theta,x;h) := \frac{1}{n(n-1)}\sum_{\substack{j,k=1\\ j\neq k}}^{n}\tau(Y_j,Y_k,\theta)K_h(X_j-x)K_h(X_k-x)\ ,$$

we have for any $\rho\in[1,\infty)$ and any compact $J\subset\mathrm{int}(I)$

$$\limsup_{n\to\infty}\sup_{\gamma\in\Gamma}\sup_{\alpha\in A}\left(\frac{\log n}{nh_n(\alpha)^d}\right)^{-\frac{\rho}{2}}\mathbb{E}_\gamma\left[\sup_{x\in J}\sup_{\theta\in\Theta}\big|M_n\big(\theta,x;h_n(\alpha)\big) - \mathbb{E}_\gamma\big[M_n\big(\theta,x;h_n(\alpha)\big)\big]\big|^\rho\right] \le C\ ,$$

where $C<\infty$ depends on $\|\tau\|_\infty$, L_τ, $\|K\|_\infty$, L_K, ρ, I, Θ, but is free from n and the sequences of bandwidth parameters.

Remark 3.2.8.

(i) Note that Theorem 3.2.7 can be generalized similarly to the result on linear processes, cf. Theorem 3.2.5, but it is not necessary for the remainder of this thesis.

(ii) Whenever we have a function M_n mapping to $\mathbb{R}^k$, $k>1$ and wish to apply one of the Theorems 3.2.5 or 3.2.7, it is enough to assume that every coordinate projection $\psi_i(M_n)$, $i=1,\dots,k$ of the function M_n fulfils the respective assumptions. Because then by Jensen's inequality for sums,

$$\begin{aligned}&\mathbb{E}_\gamma\Big[\sup_{x\in J}\sup_{\theta\in\Theta}\big\|M_n\big(\theta,x;h_n(\alpha)\big) - \mathbb{E}_\gamma[M_n\big(\theta,x;h_n(\alpha)\big)]\big\|_1^\rho\Big]\\ &\le k^{\rho-1}\sum_{i=1}^{k}\mathbb{E}_\gamma\Big[\sup_{x\in J}\sup_{\theta\in\Theta}\big|\psi_i\big(M_n\big(\theta,x;h_n(\alpha)\big)\big) - \mathbb{E}_\gamma\big[\psi_i\big(M_n\big(\theta,x;h_n(\alpha)\big)\big)\big]\big|^\rho\Big]\ ,\end{aligned}$$

which yields the result.

3.2.3 Methods specific to adaptive estimation

In this section, we give techniques for proving **(B9)** for differentiable contrast functions in the form of either a linear or a U-statistic process as defined in (3.2.3) and (3.2.4), respectively, i.e.

$$M_n(\theta, x; h) = \frac{1}{n}\sum_{k=1}^{n} \tau(Y_k, \theta) K_h(X_k - x) \ , \quad \text{or} \tag{3.2.17}$$

$$M_n(\theta, x; h) = \frac{1}{n(n-1)} \sum_{\substack{j,k=1 \\ j \neq k}}^{n} \tau(Y_j, Y_k, \theta) K_h(X_j - x) K_h(X_k - x) \ . \tag{3.2.18}$$

That is, for any compact $J \subset \mathrm{int}(I)$, we prove the existence of some $C_- \geq 0$ and a monotone function $u : [C_-, \infty) \to (0, \infty)$ with $u(t) \to \infty$, $t \to \infty$ so that for every $C_{\mathrm{Lep}} \geq C_-$,

$$\limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \sup_{0\leq l\leq j\leq k_n(\alpha)} n^{u(C_{\mathrm{Lep}})} p_{lj} < \infty \ ,$$

where

$$\begin{aligned} p_{lj} &= 2\,\mathbb{P}\Big(\sup_{x\in J, \theta\in\Theta} \big\| T_n(\theta, x; h_j) \big\| > \tfrac{C_{\mathrm{Lep}} - 3C^{**}}{2} r_l \Big) \ , \\ T_n(\theta, x; h) &= S_n(\theta, x; h) - \mathbb{E}_\gamma[S_n(\theta, x; h)] \ , \\ h(\alpha) &= \left(\frac{\log n}{n} \right)^{\frac{1}{2\alpha+d}} \ , \quad h_l = h(\beta_l) \ , \quad r(\alpha) = h(\alpha)^\alpha \ , \quad r_l = r(\beta_l) \ . \end{aligned}$$

The main tools for deriving u and C_{Lep} are also used when proving the results on uniform L^ρ convergence rates of stochastic errors in Section 3.2.2, namely discretization, treating the discrete error with exponential deviation inequalities and the discretization error by smoothness arguments.

First, assume we are dealing with a linear process as defined in (3.2.17).

Lemma 3.2.9. *Let M_n be a linear process as defined in* (3.2.17) *that is differentiable in θ and all assumptions of Theorem 3.2.5 hold for the coordinate projections of the gradient*

$$S_n(\theta, x; h) := \frac{1}{n}\sum_{k=1}^{n} \partial_\theta \tau(Y_k, \theta) K_h(X_k - x) \ .$$

Then, for any positive constants $c_1, c_2 > 0$,

$$\begin{aligned} C_- &= c_1^{-1} \left(c_2 + 1 + 32 \max\{C_1, C_2\}(d+m) \cdot \max\left\{ \frac{1}{2(d+m)-1}, \frac{b+d+1}{2a+d} + 1 \right\} \right) \ , \\ u(C_{\mathrm{Lep}}) &= \frac{c_1 C_{\mathrm{Lep}} - c_2}{32 \max\{C_1, C_2\}(d+m)} \ , \end{aligned}$$

we have

$$\limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \sup_{0\le l\le j\le k_n(\alpha)} n^{u(C_{\mathrm{Lep}})}\tilde{p}_{lj} < \infty\ , \quad C_{\mathrm{Lep}} \ge C_-\ , \quad \textit{where}$$
$$\tilde{p}_{lj} = \mathbb{P}\Big(\sup_{x\in J,\theta\in\Theta} \|T_n(\theta,x;h_j)\| > \big(c_1 C_{\mathrm{Lep}} - c_2\big) r_l \Big)\ .$$

Next, we are dealing with U-statistic processes as defined in (3.2.18).

Lemma 3.2.10. *Let M_n be a U-statistic process as defined in* (3.2.18) *that is differentiable in θ and all assumptions of Theorem 3.2.7 hold for the coordinate projections of the gradient*

$$S_n(\theta,x;h) = \frac{1}{n(n-1)} \sum_{\substack{j,k=1\\ j\neq k}}^{n} \partial_\theta \tau(Y_j,Y_k,\theta) K_h(X_j - x) K_h(X_k - x)\ .$$

For any positive constants $\tilde{c}_1, \tilde{c}_2 > 0$, let

$$\begin{aligned}
C_- &= \max\{\widetilde{C}_1, \widetilde{C}_2\}\ ,\\
\widetilde{C}_1 &= \tilde{c}_1^{-1}\left[\tilde{c}_2 + 4 + 64T^2(d+m)^2 \max\left\{ \left(\frac{1}{2(d+m)-1}\right)^2, \left(\frac{b+2d+1}{2a+d}+1\right)^2 \right\}\right]\ ,\\
\widetilde{C}_2 &= 2\tilde{c}_1^{-1}\left[\tilde{c}_2/2 + 1 + 32\max\{C_1,C_2\}(d+m)\max\left\{ \frac{1}{2(d+m)-1}, \frac{b+d+1}{2a+d}+1 \right\}\right]\ ,\\
u(C_{\mathrm{Lep}}) &= \min\left\{ \frac{T^{-1}\sqrt{\tilde{c}_1 C_{\mathrm{Lep}} - \tilde{c}_2}}{8(d+m)}, \frac{\tilde{c}_1 C_{\mathrm{Lep}} - \tilde{c}_2}{64\max\{C_1,C_2\}(d+m)} \right\}\ ,
\end{aligned}$$

where T is the universal constant in Lemma 3.2.4 and C_1, C_2 are defined in Theorem 3.2.5.
Then we have

$$\limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \sup_{0\le l\le j\le k_n(\alpha)} n^{u(C_{\mathrm{Lep}})}\tilde{p}_{lj} < \infty\ , \quad C_{\mathrm{Lep}} \ge C_-\ , \quad \textit{where}$$
$$\tilde{p}_{lj} = \mathbb{P}\Big(\sup_{x\in J,\theta\in\Theta} \|T_n(\theta,x;h_j)\| > \big(\tilde{c}_1 C_{\mathrm{Lep}} - \tilde{c}_2\big) r_l \Big)\ .$$

The constant $\widetilde{C}_1$ and the first term in the minimum in the definition of u are relevant to the U-statistic terms when decomposing T_n. The other terms are used to treat the linear remainder. The proofs can be found in Section 5.1.

4 Applications

4.1 Finite mixtures of normal regressions

Consider a mixture model of m normal regressions with multivariate covariates, expanding the model with univariate regressors considered by Huang et al. (2013). That is, we have observations (Y_i, X_i) taking values in $\mathbb{R} \times I$ for some compact cuboid $I \subset \mathbb{R}^d$ containing an open subset and a latent model variable Π_i taking values in $\{1, \dots, m\}$ for some $m \geq 2$ so that $\mathbb{P}(\Pi_i = c | X_i = x) = \pi_c(x)$, where $\pi_c : I \to (0,1)$, $\sum_{c=1}^m \pi_c = 1$, are mixing functions and X_i have the common density $\ell : I \to (0, \infty)$. The observations have the relation

$$Y_i = \sum_{c=1}^{m} \mathbb{1}_{\Pi_i = c}\big(\sigma_c(X_i)\varepsilon_{c,i} + \mu_c(X_i)\big) \ ,$$

where

$$\varepsilon_{c,i} \perp\!\!\!\perp \Pi_i \,|\, X_i \ , \qquad \varepsilon_{c,i} \,|\, X_i = x \sim \mathcal{N}(0,1) \ ,$$

the functions $\mu_c : I \to \mathbb{R}$, $\sigma_c : I \to (0, \infty)$ are location and scaling functions to be estimated along with the mixing functions π_c so that the parameter of interest is given by the function

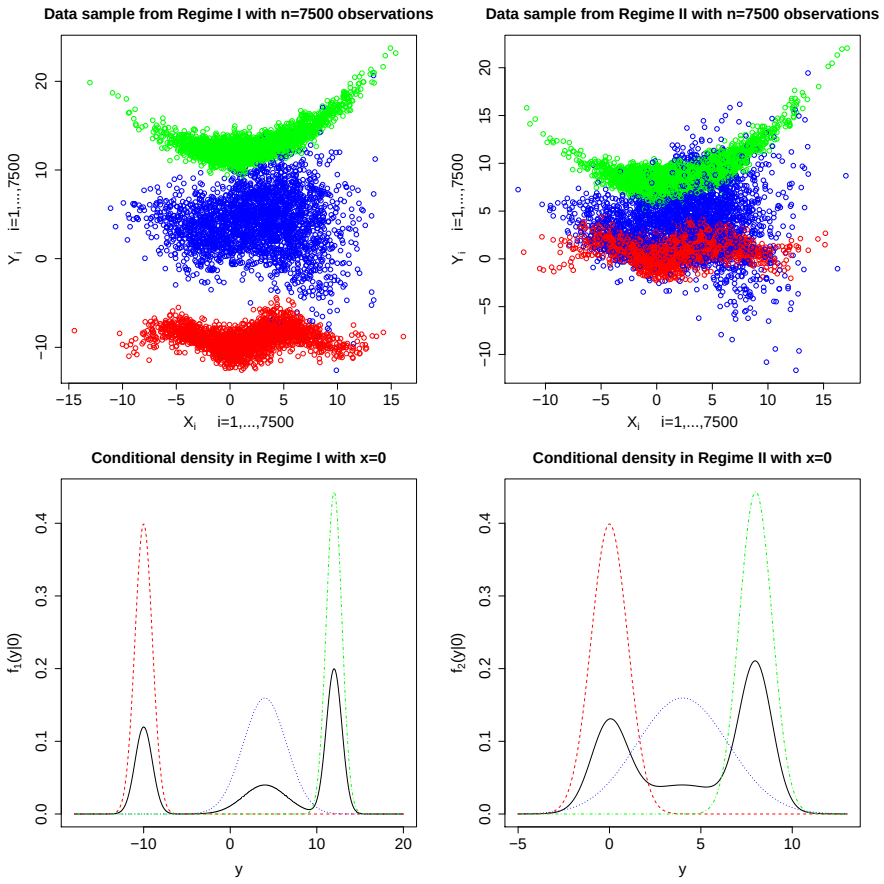
$$\theta(\cdot) = \big(\pi_1(\cdot), \dots, \pi_{m-1}(\cdot), \mu_1(\cdot), \dots, \mu_m(\cdot), \sigma_1(\cdot), \dots, \sigma_m(\cdot)\big)^T : I \to \mathbb{R}^{3m-1} \ .$$

Lemma 4.1.1. *For all $x \in I$, the conditional density of Y_i given $X_i = x$ exists and is given by*

$$f_{Y|X}^{\theta(\cdot)}(y|x) := \sum_{c=1}^{m} \pi_c(x)\phi\big(y\big|\mu_c(x), \sigma_c^2(x)\big) \ , \quad y \in \mathbb{R} \ .$$

Especially, the joint distribution of Y_i and X_i is given by

$$\begin{aligned} f_{Y,X}(y,x) &:= f_{Y|X}^{\theta(\cdot)}(y|x)\ell(x) \\ &= \left[\sum_{c=1}^{m} \pi_c(x)\phi\big(y\big|\mu_c(x), \sigma_c^2(x)\big)\right] \cdot \ell(x) \ , \quad (y,x) \in \mathbb{R} \times I \ . \end{aligned}$$



	Regime I	Regime II
Covariate distribution	$\mathcal{N}(2,4)$	$\mathcal{N}(2,4)$
Component 1	$\pi_1(x) = 0.3 - 0.05\sin(0.2x)$ $\mu_1(x) = -10 + 2\sin(0.3x)^2$ $\sigma_1(x) = 1 + 0.2\sin(0.5x)$	$\pi_1(x) = 0.3 - 0.05\sin(0.2x)$ $\mu_1(x) = 2\sin(0.3x)^2$ $\sigma_1(x) = 1 + 0.2\sin(0.5x)$
Component 2	$\pi_2(x) = 0.25 + 0.2\sin(0.2x)$ $\mu_2(x) = 4 + \sin(0.5x)$ $\sigma_2(x) = 1.5 + \exp(x/8)$	$\pi_2(x) = 0.25 + 0.2\sin(0.2x)$ $\mu_2(x) = 4 + \sin(0.5x)$ $\sigma_2(x) = 1.5 + \exp(x/8)$
Component 3	$\mu_3(x) = 12 + 0.05x^2$ $\sigma_3(x) = 0.8 + 0.1\cos(x)$	$\mu_3(x) = 8 + 0.05x^2$ $\sigma_3(x) = 0.8 + 0.1\cos(x)$

Figure 4.1: Samples and densities from mixture of normal regressions models. Data points are color coded by their respective subpopulation, where red, blue and green correspond to the components 1-3, respectively. The black curves are the conditional densities given $X = 0$. Dashed curves illustrate the distribution within the respective subpopulation. Both regimes are identical up to translation of the components. In the data sample of Regime I, the presence of three components is directly visible. In the other regime this is not the case because the means are relatively close.

4.1.1 Identifiability

A statistical model is called identifiable if there are no two parameters yielding the same distribution. To be precise, let $\Theta_1 \subset \Theta_2$ be non-empty sets and $(\mathbb{P}_\theta)_{\theta\in\Theta_2}$ a statistical model. The subset $(\mathbb{P}_\theta)_{\theta\in\Theta_1}$ is called identifiable within $(\mathbb{P}_\theta)_{\theta\in\Theta_2}$ if for any $\theta_1 \in \Theta_1, \theta_2 \in \Theta_2$, $\mathbb{P}_{\theta_1} = \mathbb{P}_{\theta_2}$ implies $\theta_1 = \theta_2$.

When dealing with mixture models, one often has to weaken the definition due to the relabeling problem. Consider any normal mixture density

$$\sum_{c=1}^{m} \pi_c \phi(y|\mu_c, \sigma_c^2),$$

then any permutation of the components yields the same mixture. Hence, we will comply with the common solution for this problem. That is, a mixture of normal distributions is identifiable if for any $\theta_1 \in \Theta_1$, $\theta_2 \in \Theta_2$, $\mathbb{P}_{\theta_1} = \mathbb{P}_{\theta_2}$ implies that θ_1 is a relabeled version of θ_2.

We give two identifiability results. The first one imposes differentiability conditions on the location and scaling functions and is stronger than needed for the estimation procedure. Particularly when having no differentiability, we may use a weaker result deduced directly from the classical identifiability result for normal mixture densities, cf. Teicher (1963).

In Huang et al. (2013), the identifiability proof is based on the concept of transversal curves, that is, differentiable curves mapping from $\mathbb{R}$ to $\mathbb{R}^2$ that may intersect, however not be tangential at intersections. To be precise, for any components $c \neq c'$, one assumes that

$$\Big(\big\| \big(\mu_c(x), \sigma_c(x)\big) - \big(\mu_{c'}(x), \sigma_{c'}(x)\big) \big\| + \big\| \big(\partial\mu_c(x), \partial\sigma_c(x)\big) - \big(\partial\mu_{c'}(x), \partial\sigma_{c'}(x)\big) \big\| \Big) \neq 0 \, .$$

As the set of intersection points S is discrete, one can show that the mixture is identifiable on all connected components of S^c. Subsequently, by approaching all intersection points from both sides with sequences, one deduces that the mixture is identifiable overall. We wish to extend this idea to differentiable curves mapping from $\mathbb{R}^d$ to $\mathbb{R}^2$. As we show later, we only need intersection points to be not tangential in one direction, i.e. for any $c \neq c'$ and any intersection point x of $(\mu_c, \sigma_c), (\mu_{c'}, \sigma_{c'})$, there needs to be a direction $z \in \mathbb{R}^d$, $\|z\| = 1$, so that $\partial_z(\mu_c(x), \sigma_c(x)) \neq \partial_z(\mu_{c'}(x), \sigma_{c'}(x))$. Hence, for those intersection points x, we need to have that

$$d > D^{c,c'}(x) \, ,$$

where

$$D^{c,c'}(x) := \max \Big\{ k \in \{1, \dots, d\} | \exists \, z_1, \dots, z_k \in \mathbb{R}^d, \|z_i\| = 1, \; z_i \text{ linearly independent} : \\ \forall i \in \{1, \dots, k\} \; \partial_{z_i}\mu_c(x) = \partial_{z_i}\mu_{c'}(x) \, , \partial_{z_i}\sigma_c(x) = \partial_{z_i}\sigma_{c'}(x) \Big\} \, .$$

Theorem 4.1.2. *Let $\mathcal{O} \subset \mathbb{R}^d$ be an open connected set. Assume that $\pi_c : \mathcal{O} \to (0,1)$ are continuous functions with $\sum_{c=1}^m \pi_c = 1$, and $\mu_c : \mathcal{O} \to \mathbb{R}$ and $\sigma_c : \mathcal{O} \to (0,\infty)$ are differentiable functions, $c = 1,\dots,m$; any two curves $(\mu_c(\cdot), \sigma_c(\cdot))$, $(\mu_{c'}(\cdot), \sigma_{c'}(\cdot))$ fulfil*

$$\big\|\big(\mu_c(x), \sigma_c(x)\big) - \big(\mu_{c'}(x), \sigma_{c'}(x)\big)\big\| + \big(d - D^{c,c'}(x)\big) \neq 0\ , \quad x \in \mathcal{O}.$$

Then the family of mixtures

$$f(\cdot|x) := \sum_{c=1}^m \pi_c(x)\phi\big(\,\cdot\,\big|\mu_c(x), \sigma_c^2(x)\big)\ , \quad x \in \mathcal{O}$$

is identifiable within all families of mixtures of normals indexed by $\mathcal{O}$ with at least two and at most m components, positive mixing functions and differentiable location and scaling functions. I.e. if there are $m \geq V \geq 2$, positive mixing functions $\lambda_1,\dots,\lambda_V : \mathcal{O} \to (0,1)$ with $\sum_{v=1}^V \lambda_v = 1$, and differentiable functions $\nu_v : \mathcal{O} \to \mathbb{R}$, $\delta_v : \mathcal{O} \to (0,\infty)$, $v = 1,\dots,V$ so that for all $x \in \mathcal{O}$

$$f(\cdot|x) = \sum_{v=1}^V \lambda_v(x)\phi\big(\,\cdot\,\big|\nu_v(x), \delta_v^2(x)\big)\ ,$$

then we have $V = m$ and there is a permutation τ on $\{1,\dots,m\}$ so that for all $c = 1\dots,m$, $x \in \mathcal{O}$,

$$\lambda_{\tau(c)}(x) = \pi_c(x)\ , \quad \nu_{\tau(c)}(x) = \mu_c(x)\ , \quad \delta_{\tau(c)}(x) = \sigma_c(x)\ . \tag{4.1.1}$$

The following result derived from the one on normal mixture densities by Teicher (1963) is weaker than Theorem 4.1.2 but suffices for the uniform estimation results we aim to prove.

Theorem 4.1.3. *Let $I \subset \mathbb{R}^d$ be a compact cuboid containing an open subset. Assume that $\pi_c : I \to (0,1)$, $\sum_{c=1}^m \pi_c = 1$, $\mu_c : I \to \mathbb{R}$ and $\sigma_c : I \to (0,\infty)$ are continuous functions; any two curves $(\mu_c(\cdot), \sigma_c(\cdot))$, $(\mu_{c'}(\cdot), \sigma_{c'}(\cdot))$ with $c \neq c'$ fulfil*

$$\big(\mu_c(x), \sigma_c(x)\big) \neq \big(\mu_{c'}(x), \sigma_{c'}(x)\big)\ , \quad x \in I\ .$$

Then, the family of mixtures

$$f(\cdot|x) := \sum_{c=1}^m \pi_c(x)\phi\big(\,\cdot\,\big|\mu_c(x), \sigma_c^2(x)\big)\ , \quad x \in I$$

is identifiable within all families of mixtures of normals indexed by I with at least two and at most m components, positive mixing functions and continuous location and scaling functions.

The proofs for both theorems can be found in Section 5.2.1.

4.1.2 Estimation

We will use the local log-likelihood approach introduced by Huang et al. (2013). The relevant estimation theory on local M-estimators is covered in Chapter 3.

Assume the true parameter function to be

$$\theta_*(\cdot) = \big(\pi_1^*(\cdot), \dots, \pi_{m-1}^*(\cdot), \mu_1^*(\cdot), \dots, \mu_m^*(\cdot), \sigma_1^*(\cdot), \dots, \sigma_m^*(\cdot)\big)^T \, .$$

The log-likelihood for the conditional distribution of $Y|X = x$ is given by the function

$$\big(y, \theta(\cdot)\big) \mapsto \log\Big(f_{Y|X=x}^{\theta(\cdot)}(y|x)\Big) := \log\Bigg(\sum_{c=1}^{m} \pi_c(x)\phi\big(y\big|\mu_c(x), {\sigma_c}^2(x)\big)\Bigg) \, ,$$

where $\theta(\cdot)$ comes from a non-parametric function class. A local version is given by

$$g(y;\theta) := \log\big(f_{\mathrm{mix}}(y;\theta)\big) := \log\Bigg(\sum_{c=1}^{m} \pi_c\phi\big(y\big|\mu_c, \sigma_c^2\big)\Bigg) \, ,$$

where

$$\theta = (\pi_1, \dots, \pi_{m-1}, \mu_1, \dots, \mu_m, \sigma_1, \dots, \sigma_m)^T \in \mathfrak{X} \, , \qquad \mathfrak{X} = \mathcal{S}_m \times \mathbb{R}^m \times (0,\infty)^m$$

and

$$\mathcal{S}_m = \Bigg\{(\pi_1, \dots, \pi_{m-1})^T \in (0,1)^{m-1} : \sum_{j=1}^{m-1} \pi_j < 1\Bigg\} \, .$$

Referring to the Kullback-Leibler divergence, we know that the function

$$M(\cdot, x; \gamma) := \mathbb{E}_\gamma\big[g(Y;\cdot)\big|X = x\big] \cdot \ell(x)$$

is uniquely maximized by the true parameter $\theta_*(x)$ up to relabeling when the model is identifiable up to relabeling, where $\mathbb{E}_\gamma$ denotes the expectation with respect to the distribution $\mathbb{P}_\gamma$, which is defined by

$$\mathbb{P}_\gamma\big((Y,X) \in A\big) = \int_A f_{Y|X}^{\theta_*(\cdot)}(y|x)\ell(x)\,\mathrm{d}(y,x) \, , \qquad \gamma = (\theta_*(\cdot), \ell) \, . \tag{4.1.2}$$

Hence, let us define the set $\mathfrak{S}_{x;\gamma}$ to contain the true parameter $\theta_*(x)$ as well as all relabeled versions, i.e.

$$\mathfrak{S}_{x;\gamma} = \Big\{\theta_* \in \mathfrak{X} \,\Big|\, \theta_* \in \operatorname*{argmax}_{\theta\in\mathfrak{X}} M(\theta, x; \gamma)\Big\} \, . \tag{4.1.3}$$

An empirical version of the asymptotic contrast M is given by

$$M_n(\theta, x; h) := \frac{1}{n}\sum_{k=1}^{n} g(Y_k;\theta)K_h(X_k - x) = \sum_{k=1}^{n} \log\Bigg(\sum_{c=1}^{m} \pi_c\phi\big(Y_k\big|\mu_c, \sigma_c^2\big)\Bigg)K_h(X_k - x) \, ,$$

so that its maximizer

$$\hat{\theta}_n(\cdot; h) \in \operatorname*{argmax}_{\theta} M_n(\theta, \cdot; h) \tag{4.1.4}$$

is the proposed estimator of $x \mapsto \mathfrak{S}_{x;\gamma}$, which exists when we restrict parameters to come from a compact subset of $\mathfrak{X}$. Note that both the empirical and asymptotic contrasts are symmetric in $(\pi_1, \mu_1, \sigma_1), \ldots, (\pi_m, \mu_m, \sigma_m)$ so that any relabeled maximizer also maximizes the respective contrast.

4.1.3 Uniform rates of convergence

In order to achieve a setting in which the estimator $\hat{\theta}_n(\cdot; h)$ defined in (4.1.4) estimates the parameter function $\theta_*(\cdot)$ and relabeled versions thereof with standard non-parametric uniform convergence rate, we need to make further assumptions on the model. We will consider models in which the parameter functions $\pi_c^*(\cdot)$, $\mu_c^*(\cdot)$, $\sigma_c^*(\cdot)$ and the covariate density ℓ are Hölder-α-smooth as defined in Section 2.4, Definition 2.4.1. Let us denote the parameter space by

$$\Theta = \Big\{(\pi_1, \ldots, \pi_{m-1}) \in U_\pi^{m-1} \Big| 1 - \textstyle\sum_{c=1}^{m-1} \pi_c \in U_\pi \Big\} \times U_\mu^m \times U_\sigma^m \subset \mathfrak{X}$$

for some sets $U_\pi \subset (0,1)$, $U_\mu \subset \mathbb{R}$, $U_\sigma \subset (0,\infty)$ and the set of admissible parameter functions by

$$\begin{aligned}
&\Delta(\alpha, L, U_\pi, U_\mu, U_\sigma, \varepsilon^\Delta) \\
:=&\Big\{\theta(\cdot) = \big(\pi_1(\cdot), \ldots, \pi_{m-1}(\cdot), \mu_1(\cdot), \ldots, \mu_m(\cdot), \sigma_1(\cdot), \ldots, \sigma_m(\cdot)\big)^T : I \to \Theta \Big| \\
&\quad \forall c = 1, \ldots, m-1 : \ \pi_c(\cdot) \in H(\alpha, L, U_\pi),\ 1 - \textstyle\sum_{c=1}^{m-1} \pi_c(x) \in U_\pi \ ,\ x \in I \ , \\
&\quad \forall c = 1, \ldots, m : \ \mu_c(\cdot) \in H(\alpha, L, U_\mu),\ \sigma_c(\cdot) \in H(\alpha, L, U_\sigma) \ , \\
&\quad \forall c \neq c', x : \ \big\| \big(\mu_c(x), \sigma_c(x)\big) - \big(\mu_{c'}(x), \sigma_{c'}(x)\big) \big\| \geq \varepsilon^\Delta \Big\}
\end{aligned} \tag{4.1.5}$$

for some $\varepsilon^\Delta > 0$ and any $\alpha, L > 0$. The last assumption in the definition of the parameter function set $\Delta(\alpha, L, U_\pi, U_\mu, U_\sigma, \varepsilon^\Delta)$ ensures that the parameter curves do not intersect and hence implies identifiability while simultaneously making $\Delta(\alpha, L, U_\pi, U_\mu, U_\sigma, \varepsilon^\Delta)$ compact with respect to the Hölder norm $\|\cdot\|_{\alpha/2}$ as stated in Remark 2.4.5 (ii) when assuming that the sets U_π, U_μ, U_σ are compact intervals.

Furthermore, the covariate density ℓ is supposed to take values in some set $U_\ell \subset (0,\infty)$, so that it is an element of the set

$$\mathcal{L}(\alpha, L, U_\ell) = \big\{\ell \in H(\alpha, L, U_\ell) : \textstyle\int \ell(x)\,\mathrm{d}x = 1\big\} \ .$$

In conclusion, the relevant model parameters must come from the class

$$\Gamma(\alpha) := \Delta(\alpha, L, U_\pi, U_\mu, U_\sigma, \varepsilon^\Delta) \times \mathcal{L}(\alpha, L, U_\ell)$$

that determines the underlying distributions for a given α. From now on include α in the definition of $\mathfrak{S}_{x;\gamma}$, cf. (4.1.3), i.e.

$$\mathfrak{S}_{x;\gamma;\alpha} = \left\{ \theta_* \in \mathfrak{X} \,\middle|\, \theta_* \in \operatorname*{argmax}_{\theta \in \mathfrak{X}} M(\theta, x; \gamma) \right\}, \quad x \in I,\ \alpha > 0,\ \gamma \in \Gamma(\alpha) .$$

The following assumptions will ensure uniform consistency of $\hat{\theta}_n(\cdot; h_n)$ with convergence rate $\left(\frac{n}{\log n}\right)^{\frac{\alpha}{2\alpha+d}}$ for the bandwidth choice $h_n = \left(\frac{\log n}{n}\right)^{\frac{1}{2\alpha+d}}$, which fulfils the assumptions

$$h_n \to 0 \quad \text{and} \quad \frac{\log n}{n h_n^d} \to 0 .$$

Assumption 4.1.4.

(N1) The sets I, U_π, U_μ, U_σ, U_ℓ are compact cuboids or intervals containing an open subset and $\varepsilon^\Delta > 0$. In particular, $1/m \in \operatorname{int}(U_\pi)$.

(N2) The kernel $K : \mathbb{R}^d \to \mathbb{R}$ is Lipschitz continuous with Lipschitz constant $L_K > 0$ and has support $[-1, 1]^d$.

(N3) The kernel K is of order α, cf. Definition 2.5.1.

Remark 4.1.5.

(i) According to Remark 2.4.5 (i), we can assume $\operatorname{diam} I \leq 1$ without loss of generality.

(ii) We assume $1/m \in \operatorname{int}(U_\pi)$ so that Θ and the set of model parameters are non-empty and Θ fulfils the latter part of Assumption 3.1.4 **(A1)**.

(iii) One assumes the parameter set Θ to be of this specific structure so that label switched versions of parameters in Θ also lie in Θ. Particularly, for any $x \in I$, $\gamma \in \Gamma(\alpha)$, we have $\mathfrak{S}_{x;\gamma;\alpha} \subset \Theta$.

(iv) As the kernel K is Lipschitz continuous and has bounded support, all moments exist.

(v) The bandwidth choice comes from balancing bias and variance. The logarithmic punitive term seems to be widely present in non-parametric estimation when examining uniform convergence rates, e.g. Giné and Guillou (2002) in kernel density estimation.

Theorem 4.1.6. *Under Assumption 4.1.4, given a compact cuboid $J \subset \operatorname{int}(I)$ containing an open subset, if we let $h_n = \left(\frac{\log n}{n}\right)^{\frac{1}{2\alpha+d}}$, the estimator $\hat{\theta}_n(\cdot; h_n)$ is uniformly consistent and has uniform convergence rate $\left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+d}}$, i.e. we have that*

$$\lim_{\eta \to \infty} \limsup_{n \to \infty} \sup_{\gamma \in \Gamma(\alpha)} \mathbb{P}_\gamma \left(\left(\frac{n}{\log n} \right)^{\frac{\alpha}{2\alpha+d}} \sup_{x \in J} \min_{\theta_* \in \mathfrak{S}_{x;\gamma;\alpha}} \left\| \hat{\theta}_n(x; h_n) - \theta_* \right\| \geq \eta \right) = 0 .$$

The proof of this result can be found in Section 5.2.2.

4.1.4 Curve estimation

In the last section, we showed that the estimator $\hat{\theta}_n(\cdot) = \hat{\theta}_n(\cdot; h)$ estimates the set function

$$x \mapsto \mathfrak{S}_{x;\gamma;\alpha}$$

uniformly consistently over any compact cuboid $J \subset \mathrm{int}(I)$ containing an open subset, where we fixed γ for simplicity. Assume

$$\sup_{x \in J} \min_{\theta_* \in \mathfrak{S}_{x;\gamma;\alpha}} \left\| \hat{\theta}_n(x) - \theta_* \right\| \leq \varepsilon^* \tag{4.1.6}$$

for some $\varepsilon^* > 0$. Then, for every $x \in J$, there is a permutation ζ_x switching the labels of $\hat{\theta}_n(\cdot)$ so that

$$\left| \zeta_x\big(\hat{\theta}_n(x)\big) - \theta_*(x) \right| \leq \varepsilon^* \ .$$

This permutation however depends on x as we do not have a natural way to correctly assign labels since the parameter curve $\theta_*(\cdot)$ is unknown.

In order to obtain estimates of the parameter curves on some finite grid $\mathcal{G} \subset J$, one has to overcome this problem of label switching. The fact that the estimator $\hat{\theta}_n(\cdot)$ is uniformly consistent allows us to formulate a procedure that ensures that the labels of the estimates at all grid points match with probability approaching one.

Consider estimation over the subset of parameter functions

$$\theta_*(\cdot) \in \Delta(\alpha, L, U_\pi, U_\mu, U_\sigma, \varepsilon^\Delta)$$

with $\alpha \geq 1$ for which

$$|\mu_c^*(x) - \mu_{c'}^*(x)| \geq \varepsilon^\Delta \ , \qquad c \neq c', \ x \in I \ , \tag{4.1.7}$$

which in fact defines a closed subset of $\Delta(\alpha, L, U_\pi, U_\mu, U_\sigma, \varepsilon^\Delta)$ so that Theorem 4.1.6 holds with the appropriately altered $\Gamma(\alpha)$.

Write $J = J_1 \times \ldots \times J_d$ and suppose that for some $0 < \delta < 1$, one wants to estimate the parameter curves on the finite equidistant grid

$$\mathcal{G}_\delta = \big(x_1 + \delta \cdot \mathbb{N}_0^d\big) \cap J \ ,$$

where $x_1 = (\min J_1, \ldots, \min J_d)^T$. Furthermore, suppose one orders $\mathcal{G}_\delta$ in a way so that subsequent indices translate to neighbouring grid points. To be precise, suppose that for $G_\delta = \#\mathcal{G}_\delta$, the points $x_1, \ldots, x_{G_\delta} \in \mathcal{G}_\delta$ are ordered so that $\|x_i - x_{i+1}\|_1 = \delta$ for all $i \in \{1, \ldots, G_\delta - 1\}$.

In addition, let ε^* in (4.1.6) and the grid width δ be so small that

$$\varepsilon^\Delta > 4\varepsilon^* + 2L\delta \tag{4.1.8}$$

and let for $i \in \{1, \dots, G_\delta\}$ the estimates

$$\hat{\theta}_n(x_i) = \big(\hat{\pi}_1(x_i), \dots, \hat{\pi}_m(x_i), \hat{\mu}_1(x_i), \dots, \hat{\mu}_m(x_i), \hat{\sigma}_1(x_i), \dots, \hat{\sigma}_m(x_i)\big)^T$$

be given, where we included $\hat{\pi}_m(x_i)$ for notational purposes. Also insert $\pi^*_m(\cdot)$ into $\theta_*(\cdot)$.

We will now give an iterative procedure to assign labels correctly. The idea is that, starting with some permuted estimate $\zeta_i\big(\hat{\theta}_n(x_i)\big)$ at some grid point x_i, permuting the labels of the estimate at the grid point x_{i+1} with a permutation ζ_{i+1} such that

$$\zeta_{i+1}(c) = \operatorname*{argmin}_{\tilde{c}=1,\dots,m} \big|\hat{\mu}_{\zeta_i(c)}(x_i) - \hat{\mu}_{\tilde{c}}(x_{i+1})\big| \,, \quad c = 1, \dots, m \tag{4.1.9}$$

yields a proper estimate of one of the curves in $\mathfrak{S}_{x;\gamma;\alpha}$ with probability approaching one. That is, one chooses a permutation such that the location estimates of neighbouring grid points are closest. Before deriving this rigorously, let us give some intuition why this should work. Whenever (4.1.6) holds, for every $\mu^*_c(x_i)$, there is an estimate $\hat{\mu}_{\tilde{c}}(x_i)$ with distance of at most ε^*. Since location functions with different labels have at least distance ε^Δ according to (4.1.7), this estimate $\hat{\mu}_{\tilde{c}}(x_i)$ needs to be unique in regard of (4.1.8). The location functions μ^*_c being Hölder-α-smooth with Hölder constant L lets us subsequently deduce that the choice (4.1.9) exists and is unique, both with probability approaching one.

Let us start with the rigorous derivation. According to (4.1.6), we know that for every $i \in \{1, \dots, G_\delta\}$, there is a permutation ζ_i on $\{1, \dots, m\}$ so that for all $c \in \{1, \dots, m\}$,

$$\big|\hat{\mu}_{\zeta_i(c)}(x_i) - \mu^*_c(x_i)\big| \le \varepsilon^* \,. \tag{4.1.10}$$

Without loss of generality assume that $\zeta_1 = \mathrm{id}$ and that the Hölder-α-smoothness of the parameter functions holds with respect to the norm $\|\cdot\|_1$ on I. As $\delta < 1$, we have for any $i \in \{1, \dots, G_\delta - 1\}$ that

$$\big|\mu^*_c(x_i) - \mu^*_c(x_{i+1})\big| \le L\delta \,, \tag{4.1.11}$$

because for $\alpha \ge 1$, μ^*_c is especially Lipschitz continuous with Lipschitz constant L. Combining (4.1.10) for $i+1$ and (4.1.11), we derive

$$\big|\hat{\mu}_{\zeta_{i+1}(c)}(x_{i+1}) - \mu^*_c(x_i)\big| \le \varepsilon^* + L\delta \,, \qquad i \in \{1, \dots, G_\delta - 1\} \tag{4.1.12}$$

by triangular inequality. Now, for any $c' \ne c$, $i \in \{1, \dots, G_\delta - 1\}$, combining (4.1.7), (4.1.8) and (4.1.12), we see that

$$\begin{aligned} \big|\mu^*_c(x_i) - \hat{\mu}_{\zeta_{i+1}(c')}(x_{i+1})\big| &\ge \big|\mu^*_c(x_i) - \mu^*_{c'}(x_i)\big| - \big|\mu^*_{c'}(x_i) - \hat{\mu}_{\zeta_{i+1}(c')}(x_{i+1})\big| \\ &> 4\varepsilon^* + 2L\delta - (\varepsilon^* + L\delta) \\ &= 3\varepsilon^* + L\delta \,. \end{aligned} \tag{4.1.13}$$

Because of (4.1.10), (4.1.12), (4.1.13) and then again (4.1.10), we deduce that

$$\big|\hat{\mu}_{\zeta_i(c)}(x_i) - \hat{\mu}_{\zeta_{i+1}(c)}(x_{i+1})\big| \le 2\varepsilon^* + L\delta \,,$$

$$\begin{aligned}\left|\hat{\mu}_{\zeta_i(c)}(x_i) - \hat{\mu}_{\zeta_{i+1}(c')}(x_{i+1})\right| &\geq \left|\hat{\mu}_{\zeta_{i+1}(c')}(x_{i+1}) - \mu^*_c(x_i)\right| - \left|\mu^*_c(x_i) - \hat{\mu}_{\zeta_i(c)}(x_i)\right| \\ &> 3\varepsilon^* + L\delta - \varepsilon^* \\ &= 2\varepsilon^* + L\delta\,.\end{aligned}$$

This especially means that on the set

$$A_{n\varepsilon^*} \;=\; \Big\{ \sup_{x\in J} \min_{\theta_*\in\mathfrak{S}_{x;\gamma;\alpha}} \big\|\hat{\theta}_n(x) - \theta_*\big\| \leq \varepsilon^* \Big\}\,,$$

for any $i \in \{1,\ldots,G_\delta - 1\}$, $c \in \{1,\ldots,m\}$, we have

$$\zeta_{i+1}(c) = \operatorname*{argmin}_{\tilde{c}=1,\ldots,m} \left|\hat{\mu}_{\zeta_i(c)}(x_i) - \hat{\mu}_{\tilde{c}}(x_{i+1})\right|\,.$$

Hence, starting with $\widetilde{\zeta}_1 = \mathrm{id}$, let us define

$$\widetilde{\zeta}_i(c) = \operatorname*{argmin}_{\tilde{c}=1,\ldots,m} \left|\hat{\mu}_{\widetilde{\zeta}_{i-1}(c)}(x_{i-1}) - \hat{\mu}_{\tilde{c}}(x_i)\right|\,, \qquad i \in \{2,\ldots,G_\delta\}\,,$$

which are well-defined permutations on the set $A_{n\varepsilon^*}$. In order to obtain permutations that are also well-defined on $A^c_{n\varepsilon^*}$, we propose for $i \in \{2,\ldots,G_\delta\}$, $c = 1,\ldots,m$,

$$\hat{\zeta}_1 \;=\; \mathrm{id}\,, \qquad \hat{\zeta}_i(c) \;=\; \operatorname*{argmin}_{\tilde{c}\in\{1,\ldots,m\}\setminus\{\hat{\zeta}_i(1),\ldots,\hat{\zeta}_i(c-1)\}} \left|\hat{\mu}_{\hat{\zeta}_{i-1}(c)}(x_{i-1}) - \hat{\mu}_{\tilde{c}}(x_i)\right|\,,$$

which in fact coincide with $\widetilde{\zeta}_i$ on $A_{n\varepsilon^*}$. Now, the curve estimators are given by

$$\hat{\theta}_{\hat{\zeta}}(x_i) := \big(\hat{\pi}_{\hat{\zeta}_i(1)}(x_i),\ldots,\hat{\pi}_{\hat{\zeta}_i(m)}(x_i),\hat{\mu}_{\hat{\zeta}_i(1)}(x_i),\ldots,\hat{\mu}_{\hat{\zeta}_i(m)}(x_i),\hat{\sigma}_{\hat{\zeta}_i(1)}(x_i),\ldots,\hat{\sigma}_{\hat{\zeta}_i(m)}(x_i)\big)^T\,.$$

Still under the assumption that $\zeta_1 = \mathrm{id}$, we see that

$$\Big\{ \sup_{x\in\mathcal{G}_\delta} \big\|\hat{\theta}_{\hat{\zeta}}(x) - \theta_*(x)\big\| \leq \varepsilon^* \Big\} \supset \Big\{ \sup_{x\in J} \min_{\theta_*\in\mathfrak{S}_{x;\gamma;\alpha}} \big\|\hat{\theta}_n(x) - \theta_*\big\| \leq \varepsilon^* \Big\}\,,$$

where the probability of the right-hand side converges to one due to uniform consistency of $\hat{\theta}_n(\cdot)$. When dropping this assumption, there must be a permutation ζ_* independent of x so that

$$\Big\{ \sup_{x\in\mathcal{G}_\delta} \big\|\zeta_*\big(\hat{\theta}_{\hat{\zeta}}(x)\big) - \theta_*(x)\big\| \leq \varepsilon^* \Big\} \supset \Big\{ \sup_{x\in J} \min_{\theta_*\in\mathfrak{S}_{x;\gamma;\alpha}} \big\|\hat{\theta}_n(x) - \theta_*\big\| \leq \varepsilon^* \Big\}\,.$$

4.1.5 Uniform adaptive estimation

Following the Lepski method for the gradients described in Section 3.1.1, we propose an adaptive estimator for parameters coming from

$$\Delta(a,L,U_\pi,U_\mu,U_\sigma,\varepsilon^\Delta) = \bigcup_{\alpha\in[a,b]} \Delta(\alpha,L,U_\pi,U_\mu,U_\sigma,\varepsilon^\Delta)$$

for some known $0 < a < b < \infty$, i.e. for the case in which the true smoothness parameter α is only known to come from an interval $[a, b]$. Note that L, U_π, U_σ, U_μ, Θ, I are assumed not to differ for varying α and are to have the same properties from the last section. Also note that the equality in the display above directly follows from inclusion of the Hölder classes as described in Remark 2.4.5 (i). Particularly, the set of all model parameters is now given by

$$\Gamma(a) = \bigcup_{\alpha \in [a,b]} \Gamma(\alpha) \, . \tag{4.1.14}$$

As in Section 3.1.1, define

$$\begin{aligned} S_n(\theta, x; h) &= \partial_\theta M_n(\theta, x; h) \, , \qquad S(\theta, x; \gamma) = \partial_\theta M(\theta, x; \gamma) \, , \\ \beta_k &= a + k\frac{b-a}{N}, \quad k = 0, \dots, N, \quad N = \lceil \log n \rceil \, , \\ h(\alpha) &= \left(\frac{\log n}{n}\right)^{\frac{1}{2\alpha+d}}, \quad r(\alpha) = h(\alpha)^\alpha = \left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+d}}, \quad r_k = r(\beta_k), \quad h_k = h(\beta_k). \end{aligned}$$

Subsequently, define the adaptive data driven grid point by

$$\hat{\alpha}_n = \beta_{\hat{k}},$$

where

$$\hat{k} = \max\left\{0 \le k \le N : \sup_{x \in I, \theta \in \Theta} \left\| S_n(\theta, x; h_k) - S_n(\theta, x; h_l) \right\| \le C_{\mathrm{Lep}} r_l \quad \forall \, 0 \le l \le k \right\} ,$$

where the Lepski constant $C_{\mathrm{Lep}} < \infty$ is to be chosen large enough. A specific lower bound on the Lepski constant can be derived from the discussion in the previous chapter, particularly Assumption **(B9)** and Lemma 3.2.9.

This construction yields the estimator

$$\widehat{\theta}_n^{\mathrm{adap}}(x) = \operatorname*{argmin}_{\theta \in \Theta} M_n(\theta, x; h_{\hat{k}}) \, .$$

In order to make use of the highest possible smoothness order b, we need to assume the kernel K to be of order b, cf. Definition 2.5.1. Therefore, we formulate the following alteration of Assumption **(N3)**.

Assumption 4.1.7.

($\widetilde{\mathbf{N3}}$) The kernel K is of order b.

Now, we can state the main result.

Theorem 4.1.8. *Let $0 < a < b < \infty$, K be a kernel fulfilling Assumptions* **(N2)** *and* **($\widetilde{\mathbf{N3}}$)**. *Then, under Assumption* **(N1)**, *for any compact cuboid $J \subset \mathrm{int}(I)$ containing an open subset, there is a $C_{\mathrm{Lep}} > 0$ so that*

$$\lim_{\eta \to \infty} \limsup_{n \to \infty} \sup_{\alpha \in [a,b]} \sup_{\gamma \in \Gamma(\alpha)} \mathbb{P}_\gamma \left(\left(\frac{n}{\log n}\right)^{\frac{\alpha}{2\alpha+d}} \sup_{x \in J} \min_{\theta_* \in \mathfrak{S}_{x;\gamma;\alpha}} \left\| \widehat{\theta}_n^{\mathrm{adap}}(x) - \theta_* \right\| \ge \eta \right) = 0 \, .$$

In order to prove both Theorems 4.1.6 and 4.1.8, it is enough to prove that Conditions **(B1)**-**(B9)** in Assumption 3.1.11 hold since this directly implies Theorem 4.1.8 by Theorem 3.1.13. As discussed in Section 3.1.1, Assumption 3.1.11 implies Assumption 3.1.4 so that Theorem 4.1.6 follows from Theorem 3.1.6. A complete proof can be found in Section 5.2.2. Here, we give some intuition and sketches.

Introduce some bounded, convex and open set $\Theta \subset \Xi \subset \mathfrak{X}$ with $\overline{\Xi} \subset \mathfrak{X}$. Condition **(B1)** follows directly from Assumption **(N1)** as well as (4.1.14) when using the Hölder norm $\|\cdot\|_{a/2}$ as defined in Definition 2.4.3, cf. Remark 2.4.5 (ii) and (iv). For the latter part of Condition **(B1)**, the polytope

$$D_m = \left\{(\pi_1, \dots, \pi_{m-1}) \in U_\pi^{m-1} \middle| 1 - \textstyle\sum_{c=1}^{m-1} \pi_c \in U_\pi \right\}$$

is the central issue. Starting with a vector on the boundary, one can make small changes in single arguments without leaving D_m, which results in a vector in $\mathrm{int}(D_m)$. Hence, one only needs to consider interior points. For two interior points λ_1, λ_2, a sufficiently small ε- neighbourhood around the connecting line segment $[\lambda_1, \lambda_2]$ is a subset of $\mathrm{int}(D_m)$ as it is an open convex set. Note that one can make small changes in single arguments without leaving the neighbourhood. Proceed by reducing the distance between λ_1 and λ_2. As it turns out, one can choose $\bar{C} = 4m - 1$.

Condition **(B2)** is given because normal mixture densities are invariant under relabeling, Condition **(B6)** holds because $\varepsilon^\Delta > 0$ is independent of x, γ, α. The continuity condition **(B3)** holds because the log-likelihood is continuous and one can interchange limit and integral by Lebesgue's theorem. Further, the continuity of the parameter functions is obvious as they all are assumed to be Hölder-α-smooth. The differentiability of the contrast functions is provided by the fact that the likelihood function of mixtures of normals are $\mathcal{C}^\infty$-functions and Lebesgue's theorem once again. In order to prove that the Hessian matrices $V_x(\theta_*(x); \gamma)$ of M are positive definite as demanded by Condition **(B4)**, one first deduces that the Hessian matrix is in fact given by the negative Fisher information and reduces to

$$\begin{aligned} V_x(\theta_*(x); \gamma) =& \partial_{\theta^2}^2 M(\theta_*(x), x; \gamma) \\ =& -\mathbb{E}_\gamma \left[\frac{(\partial_\theta f_{\mathrm{mix}}(Y; \theta_*(x)))(\partial_\theta f_{\mathrm{mix}}(Y; \theta_*(x)))^T}{f_{\mathrm{mix}}(Y; \theta_*(x))^2} \middle| X = x \right] \cdot \ell(x) \, . \end{aligned}$$

From this, one can derive linear equations in the form of

$$\partial_\theta f_{\mathrm{mix}}(y; \theta_*(x))^T v = 0 \, , \qquad \forall \, y \in \mathbb{R}$$

and shows that $\mathbb{R}^{3m-1} \ni v = 0$ by using that there is always a component in the mixture density f_{mix} that is dominated by all other components asymptotically according to the definition of $\Delta(\alpha, L, U_\pi, U_\mu, U_\sigma, \varepsilon^\Delta)$ and $\varepsilon^\Delta > 0$. This successively proves that the components of v need to be zero.

For the remaining conditions, one needs to bear in mind that the log-likelihood g and its derivatives are unbounded and the mixture density f_{mix} and its derivatives are not Lipschitz continuous in θ uniformly over all y. However, one can show that g and its derivatives are integrable with respect to $\sup_{\theta\in\Xi} f_{\text{mix}}(\cdot;\theta)$ and that f_{mix} and its derivatives are Lipschitz continuous with a Lipschitz constant that is integrable in y with respect to $\sup_{\theta\in\Xi} f_{\text{mix}}(\cdot;\theta)$ as well, which directly gives **(B5)**.

One can use Bennett's inequality, cf. Lemma 3.2.3 in order to attain an exponential deviation inequality for the gradients S_n, where bounding all moments is straightforward but lengthy. Accordingly, the conditions of Theorem 3.2.5 are fulfilled, which gives **(B7)** directly and **(B9)** by Lemma 3.2.9.

The application of Theorem 3.2.5 in order to prove uniform consistency of the empirical contrast, cf. **(B8)**, does not seem to be fruitful because the log-likelihood does not provide a Lipschitz constant that is integrable with respect to $\sup_{\theta\in\Xi} f_{\text{mix}}(\cdot;\theta)$. However, there is a workaround by using the fundamental theorem of calculus along edges of cuboids spanned by parameters θ and a single parameter $\tilde{\theta}$ so that the uniform error reduces to

$$\begin{aligned}
&\sup_{\theta\in\Theta, x\in J} |M_n(\theta,x;h) - M(\theta,x;\gamma)| \\
\leq\ &\sup_{x\in J} |M_n(\tilde{\theta},x;h) - M(\tilde{\theta},x;\gamma)| \\
&+ (4m-1)\operatorname{diam}(\Theta) \sup_{\theta\in\Theta, x\in J} \left\| S_n(\theta,x;h) - S(\theta,x;\gamma) \right\|_\infty . \qquad (4.1.15)
\end{aligned}$$

The first summand in (4.1.15) can now be treated by Theorem 3.2.5 and Lemma 3.2.1 because the supremum is only taken over the covariate values. The second summand is shown to be $o_{\mathbb{P}_\gamma}(1)$ by using Markov's inequality and arguments similar to the proof of **(B7)**.

4.2 A two-component mixture of location scale regressions

We propose a non-parametric regression model of the form

$$Y_i = W_i\big(\mu(X_i) + \varepsilon_{1,i}\big) + \big(1 - W_i\big)\sigma(X_i)\varepsilon_{2,i}\ , \quad i \in \mathbb{N}$$

for sequences of i.i.d. random vectors $(X_i)_{i\in\mathbb{N}}$ having support $I \subset \mathbb{R}^d$, where I is a compact cuboid containing an open subset, $d \geq 1$ and i.i.d. random variables $(Y_i)_{i\in\mathbb{N}}$, $(W_i)_{i\in\mathbb{N}}$, $(\varepsilon_{1,i})_{i\in\mathbb{N}}$ and $(\varepsilon_{2,i})_{i\in\mathbb{N}}$. The explanatory variables X_i and the response variables Y_i are observable, the latent variables W_i and the error variables $\varepsilon_{1,i}$ and $\varepsilon_{2,i}$ are not. The covariates X_i are assumed to have a Lebesgue density $\ell : I \to (0,\infty)$.

The unknown location and scaling functions $\mu : I \to \mathbb{R}$, $\sigma : I \to (0,\infty)$ are functions to be estimated as they partially determine the distributional relation between the explanatory

and response variables along with the unknown mixing function $p : I \to (0,1)$. That is because conditionally on $X_i = x$, the variables W_i are assumed to have a Bernoulli-distribution with parameter $p(x)$, i.e.

$$\mathbb{P}(W_i = 1|X_i = x) = p(x) \quad \text{and} \quad \mathbb{P}(W_i = 0|X_i = x) = 1 - p(x) \; .$$

Let us further assume that conditionally on $X_i = x$, the vectors $\varepsilon_{1,i}$ and $\varepsilon_{2,i}$ have zero-symmetric conditional densities denoted by f_x and $\bar{f}$, respectively, where we assume that $\bar{f}$ is known and f_x is not. If we furthermore have the conditional independence relations

$$\varepsilon_{1,i} \perp\!\!\!\perp W_i|X_i \quad \text{and} \quad \varepsilon_{2,i} \perp\!\!\!\perp W_i|X_i \; ,$$

we can deduce that the random vectors (Y_i, X_i) have a joint density. Therefore, we formulate the following simple lemma that was proved in Werner (2015).

Lemma 4.2.1. *For all $x \in I$, the conditional density of Y_i given $X_i = x$ exists and is given by*

$$f_{Y|X}^{\vartheta(\cdot)}(y|x) := \frac{1-p(x)}{\sigma(x)} \bar{f}\Big(\frac{y}{\sigma(x)}\Big) + p(x) f_x\big(y - \mu(x)\big) \; , \quad y \in \mathbb{R} \; ,$$

where $\vartheta(x) := (p(x), \sigma(x), \mu(x), f_x)$. Especially, the joint distribution of Y_i and X_i is given by

$$\begin{aligned} f_{Y,X}(y,x) &:= f_{Y|X}^{\vartheta(\cdot)}(y|x)\ell(x) \\ &= \left[\frac{1-p(x)}{\sigma(x)} \bar{f}\Big(\frac{y}{\sigma(x)}\Big) + p(x) f_x\big(y - \mu(x)\big)\right] \cdot \ell(x) \; , \quad (y,x) \in \mathbb{R} \times I \; . \end{aligned}$$

4.2.1 Identifiability

We will prove local identifiability, i.e. that for every $x \in I$ the conditional density $f_{Y|X}^{\vartheta(\cdot)}(\cdot|x)$ is identifiable within all mixture densities of the proposed type under some reasonable constraints. Hence, it is enough to consider the model without covariates. We need the known component density $\bar{f}$ and the unknown model density f to have finite third-order moments and be zero-symmetric. Therefore, we only examine the densities

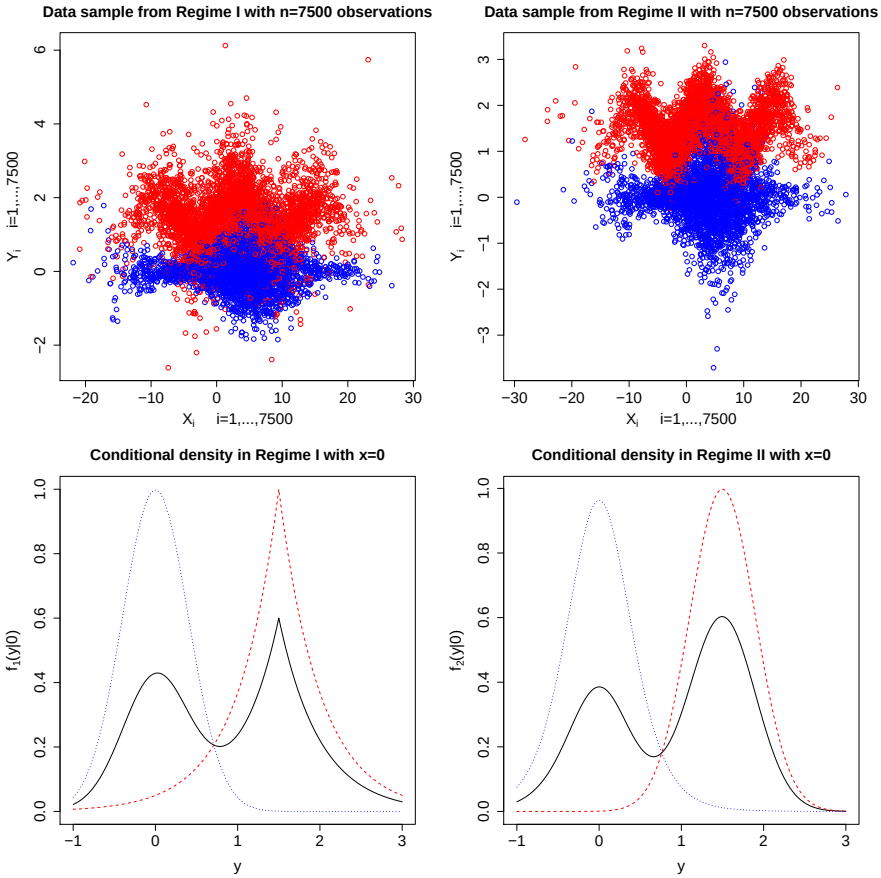
$$f_{\text{mix}}(y;\vartheta) = (1-p)\bar{f}(y/\sigma)/\sigma + pf(y-\mu) \; , \quad y \in \mathbb{R} \; ,$$

where

$$\vartheta = (p, \sigma, \mu, f)^T \in \mathfrak{X} := [0,1] \times (0,\infty) \times \mathbb{R} \times \mathcal{E}_3,$$

and $\bar{f} \in \mathcal{E}_3$ with

$$\mathcal{E}_3 = \{f : \mathbb{R} \to [0,\infty) |\ f \text{ even}, \textstyle\int f(x)\,\mathrm{d}x = 1, \int |x|^3 f(x)\,\mathrm{d}x < \infty\} \; .$$



	Regime I	Regime II
Covariate distribution	$\mathcal{N}(3,7)$	$\mathcal{N}(3,7)$
Error distributions	$\varepsilon_1 \sim \text{Laplace}(0,1/2)$ $\varepsilon_2 \sim \mathcal{N}(0,1)$	$\varepsilon_1 \sim \mathcal{N}(0,0.4)$ $\varepsilon_2 \sim \text{t}(7)$
Parameter functions	$p(x) = 0.6 - 0.15\sin(0.3x)$ $\sigma(x) = 0.4 + 0.25\sin(0.25x)$ $\mu(x) = 1.5 + 0.5\sin(0.5x)$	

Figure 4.2: Samples and densities from two-component mixtures of location scale regressions. Data points are color coded by their respective subpopulation, where red corresponds to the a priori unknown component and blue corresponds to the known component. The black curves are the conditional densities given $X = 0$. Dashed curves illustrate the distribution within the respective subpopulation. The parameter functions in both regimes are identical. However, different error distributions are chosen.

The following two examples illustrate why the model cannot be pointwise identifiable without imposing constraints.

Example 4.2.2.

(i) Let $\mu = 0$, $\sigma = 1$, $f = \bar{f}$. Then every p yields the same mixture.

(ii) For any $p, \bar{f}$ we have

$$\big(1 - \tfrac{p}{2}\big)\, \bar{f}(\cdot) + \tfrac{p}{2}\, \bar{f}(\cdot - 2) = (1 - p)\, \bar{f}(\cdot) + p\, f(\cdot - 1)$$

when $f(\cdot) = \big(\bar{f}(\cdot - 1) + \bar{f}(\cdot + 1)\big)/2$.

We will give two identifying assumptions. Both rely on the symmetry of the component densities. Note that whenever a density f is zero-symmetric, its characteristic function or Fourier transform

$$\varphi_f(t) = \int \exp(itz) f(z)\, \mathrm{d}z\ , \quad t \in \mathbb{R}$$

is real-valued.

The first assumption imposes a strong constraint on the true mixing parameter p_* but only mild conditions on the component densities $\bar{f}$ and f_*.

Assumption 4.2.3. The model parameter $\vartheta_* = (p_*, \sigma_*, \mu_*, f_*)^T \in \mathfrak{X}$ and the component density $\bar{f}$ fulfil

(I1) $\mu_* \in \mathbb{R}\backslash\{0\}$, $p_* \in (1/2, 1)$, $\sigma_* \in (0, \infty)$,

(I2) $\bar{f} \in \mathcal{E}_3$, $\varphi_{\bar{f}} > 0$,

(I3) $f_* \in \mathcal{E}_3$, $\varphi_{f_*} > 0$.

Theorem 4.2.4. *Let Assumption 4.2.3 hold. If for any $\vartheta \in \mathfrak{X}$, we have*

$$f_{\mathrm{mix}}(y; \vartheta_*) = f_{\mathrm{mix}}(y; \vartheta) \quad \textit{for almost all} \quad y \in \mathbb{R}, \tag{4.2.1}$$

then $\vartheta = \vartheta_*$.

Denote $\vartheta = (p, \sigma, \mu, f)^T$. By Fourier transforming both sides of (4.2.1) and using an addition formula for the trigonometric functions, we deduce

$$\big[(1 - p_*)\varphi_{\bar{f}}(\sigma_* t) - (1 - p)\varphi_{\bar{f}}(\sigma t)\big] \sin(\mu t) = p_* \varphi_{f_*}(t) \sin\big((\mu_* - \mu)t\big) \tag{4.2.2}$$

for all $t \in \mathbb{R}$. Examining the zeros of the terms on both sides yields $\mu = \mu_*$, which in fact implies the result, cf. Lemma 5.3.1. A complete proof can be found in Section 5.3.1.

The other identifiability result was introduced by Werner (2015) and is given for the sake of completion. It does not impose a restriction on the mixing parameter but strongly depends on the relationship of both component densities $\bar{f}$ and f_*. That is, the characteristic functions of those densities need to be distinguishable in the tails in one of the following manners.

Condition 4.2.5. For large $t \in \mathbb{R}$ it holds $\varphi_{f_*}(t) \neq 0$ and for all $\sigma > 0$, we have

$$\lim_{t\to\infty} \frac{\varphi_{\bar{f}}(\sigma t)}{\varphi_{f_*}(t)} = 0 \, .$$

Condition 4.2.6. For large $t \in \mathbb{R}$ it holds $\varphi_{f_*}(t)$, $\varphi_{\bar{f}}(t) \neq 0$ and for all $\sigma > 0$, we have

$$\lim_{t\to\infty} \frac{\varphi_{f_*}(t)}{\varphi_{\bar{f}}(\sigma t)} = 0 \, .$$

Further, we have that

$$\lim_{t\to\infty} \frac{\varphi_{\bar{f}}(\sigma t)}{\varphi_{\bar{f}}(\sigma' t)} = 0 \, , \qquad \forall\ 0 < \sigma' < \sigma \, . \tag{4.2.3}$$

Condition 4.2.7. For large $t \in \mathbb{R}$ it holds $\varphi_{f_*}(t), \varphi_{\bar{f}}(t) \neq 0$, we have (4.2.3) and further, there exists a $\sigma_0 \in (0, \infty)$ such that

$$\lim_{t\to\infty} \frac{\varphi_{f_*}(t)}{\varphi_{\bar{f}}(\sigma t)} = 0 \, , \qquad 0 < \sigma < \sigma_0 \quad \text{and} \quad \lim_{t\to\infty} \frac{\varphi_{\bar{f}}(\sigma t)}{\varphi_{f_*}(t)} = 0 \, , \qquad \sigma_0 < \sigma < \infty \, .$$

Moreover,

$$\lim_{t\to\infty} \frac{\varphi_{\bar{f}}(\sigma_0 t)}{\varphi_{f_*}(t)} = c \in [0, \infty] \setminus \{1\}.$$

Some typical examples of component densities fulfilling these conditions are given below.

Example 4.2.8.

(i) Condition 4.2.5 holds when $f_* \sim \text{Laplace}(\mu_1, \sigma_1)$ and $\bar{f} \sim \mathcal{N}(\mu_2, \sigma_2^2)$, cf. Werner (2015).

(ii) Condition 4.2.6 holds when $\bar{f} \sim t(\nu)$ and $f_* \sim \mathcal{N}(\mu_1, \sigma_1^2)$, cf. Werner (2015).

(iii) Condition 4.2.7 holds when $f_* \sim t(\nu)$, $\bar{f} \sim t(\nu_0)$ for $\nu \neq \nu_0$, cf. Werner (2015).

(iv) When both component densities are normals, none of the three conditions are fulfilled. However, any centred normal density fulfils **(I2)** and **(I3)**.

Admissible component densities f_* are aggregated in the function class

$$\mathcal{E}_3^{\bar{f}} = \{f \in \mathcal{E}_3 : (\bar{f}, f) \text{ meets one of the Conditions } 4.2.5 - 4.2.7\} \, .$$

The second identifying assumption is as follows.

Assumption 4.2.9. The model parameter $\vartheta_* = (p_*, \sigma_*, \mu_*, f_*)^T \in \mathfrak{X}$ and the known component density $\bar{f}$ fulfil

(I1) $\mu_* \in \mathbb{R}\backslash\{0\}$, $p_* \in (0, 1)$, $\sigma_* \in (0, \infty)$,

(I2) $\bar{f} \in \mathcal{E}_3$,

(I3) $f_* \in \mathcal{E}_3^{\bar{f}}$.

Theorem 4.2.10 (Werner (2015))**.** *Let Assumption 4.2.9 hold. If for any $\vartheta \in \mathfrak{X}$, we have*

$$f_{\text{mix}}(y;\vartheta_*) = f_{\text{mix}}(y;\vartheta) \quad \text{for almost all} \quad y \in \mathbb{R},$$

then $\vartheta = \vartheta_*$.

In order to prove this result, one examines (4.2.2) by dividing both sides by dominating characteristic functions and letting $t \to \infty$. By using that sin is periodic, this allows for assertions on the parameter. A full proof is given by Werner (2015) and can be found in the appendix with some modifications, cf. Section A.4.

Remark 4.2.11. Note that in both identifiability results, the additional conditions need only be imposed on the true parameter $\vartheta_* = (p_*, \sigma_*, \mu_*, f_*)^T$, it is then identifiable within the whole class of parameters $\mathfrak{X}$.

Both Theorems 4.2.4 and 4.2.10 directly give identifiability in the model with covariates.

Corollary 4.2.12. *Let for all $x \in I$ the parameter $\vartheta_*(x)$ and the component density $\bar{f}$ fulfil either Assumption 4.2.3 or Assumption 4.2.9. If for any $x \in I$, $\vartheta \in \mathfrak{X}$, we have*

$$f_{Y|X}^{\vartheta_*(\cdot)}(y|x) = f_{\text{mix}}(y;\vartheta) \quad \text{for almost all } y \in \mathbb{R}\,,$$

then $\vartheta_*(x) = \vartheta$.

Note that under the assumptions postulated in Corollary 4.2.12, we especially have identifiability over any subset of $\mathfrak{X}$. That will be useful once we restrict the parameters to a compact set.

4.2.2 Estimation

Let us construct a contrast function based on the component densities' symmetry that allows for minimization yielding an M-estimator for the parameter

$$\theta_*(x) := \big(p_*(x), \sigma_*(x), \mu_*(x)\big)^T \in (1/2, 1) \times (0, \infty) \times \mathbb{R}\backslash\{0\}$$

with $f_x^* \in \mathcal{E}_3$ under the identifying Assumption 4.2.3.

For now, consider the model without covariates. To be specific, consider a random variable Y with density

$$f_{\text{mix}}(y;\vartheta_*) = \tfrac{1-p_*}{\sigma_*}\bar{f}\big(\tfrac{y}{\sigma_*}\big) + p_* f_*(y - \mu_*)\,, \quad y \in \mathbb{R}\,,$$

where $\bar{f}, f_* \in \mathcal{E}_3$ and

$$\theta_* = (p_*, \sigma_*, \mu_*)^T \in (1/2, 1) \times (0, \infty) \times \mathbb{R}\backslash\{0\}\,, \quad \vartheta_* = (\theta_*^T, f_*)^T\,.$$

As $\bar{f}$ and f_* are symmetric densities, their characteristic functions $\varphi_{\bar{f}}$ and φ_{f_*} are real-valued. The characteristic function of the mixture $f_{\mathrm{mix}}(\cdot;\vartheta_*)$ is given by

$$\varphi_{f_{\mathrm{mix}}(\cdot;\vartheta_*)}(t) = (1-p_*)\,\varphi_{\bar{f}}(\sigma_* t) + p_*\, e^{it\mu_*}\,\varphi_{f_*}(t)\;.$$

Now, as $p_*\varphi_{f_*}(t)$ is real-valued for all $t \in \mathbb{R}$, so is

$$\Big(\varphi_{f_{\mathrm{mix}}(\cdot;\vartheta_*)}(t) - (1-p_*)\,\varphi_{\bar{f}}(\sigma_*\, t)\Big)\, e^{-it\mu_*}\;. \tag{4.2.4}$$

Since $\varphi_{f_{\mathrm{mix}}(\cdot;\vartheta_*)}(t)\, e^{-it\mu_*}$ is the characteristic function of $Y-\mu_*$, we deduce for the imaginary part of (4.2.4) that

$$\begin{aligned} 0 &= \Im\bigg(\Big(\varphi_{f_{\mathrm{mix}}(\cdot;\vartheta_*)}(t) - (1-p_*)\,\varphi_{\bar{f}}(\sigma_*\, t)\Big)\, e^{-it\mu_*}\bigg) \\ &= \mathbb{E}_{\vartheta_*}\Big[\sin\big((Y-\mu_*)\,t\big)\Big] + (1-p_*)\,\varphi_{\bar{f}}(\sigma_*\, t)\,\sin(t\mu_*) \end{aligned} \tag{4.2.5}$$

for all $t \in \mathbb{R}$, where $\mathbb{E}_{\vartheta_*}$ denotes the expectation with respect to the distribution $\mathbb{P}_{\vartheta_*}$, which fulfils

$$\mathbb{P}_{\vartheta_*}(Y \in A) = \int_A f_{\mathrm{mix}}(y;\vartheta_*)\,\mathrm{d}y\;.$$

Define the function $H : \mathbb{R}\times\mathbb{R}\times[0,1]\times(0,\infty)\times\mathbb{R} \to [-2,2]$,

$$H(y,t,\theta) = \sin\big((y-\mu)\,t\big) + (1-p)\,\varphi_{\bar{f}}(\sigma\, t)\,\sin(\mu t) \tag{4.2.6}$$

that has a contrast property as follows.

Proposition 4.2.13. *Let $\bar{f}$ fulfil* **(I2)** *and a parameter $\vartheta_* = \big(p_*, \sigma_*, \mu_*, f_*\big)^T \in \mathfrak{X}$ fulfil* **(I1)** *and* **(I3)**. *Then for $\theta \in [0,1]\times(0,\infty)\times\mathbb{R}$, we have*

$$\mathbb{E}_{\vartheta_*}\big[H(Y,t,\theta)\big] = 0 \quad \forall\; t\in\mathbb{R} \quad \Longleftrightarrow \quad \theta = \theta_* = \big(p_*,\sigma_*,\mu_*\big)^T\;.$$

In order to lose dependence of H on t, we integrate H with respect to some strictly positive density q that is chosen a priori, giving the non-negative contrast

$$M(\theta;\vartheta_*) := \int_{\mathbb{R}} \mathbb{E}_{\vartheta_*}\big[H^2(Y,t,\theta)\big]\,q(t)\,\mathrm{d}t\;, \tag{4.2.7}$$

which adopts the contrast property from $\theta \mapsto \mathbb{E}_{\vartheta_*}\big[H(Y,t,\theta)\big]$ as $q>0$.

Corollary 4.2.14. *Let $\bar{f}$ fulfil* **(I2)** *and a parameter $\vartheta_* = \big(p_*, \sigma_*, \mu_*, f_*\big)^T$ fulfil* **(I1)** *and* **(I3)**. *Then, the function $M(\cdot;\vartheta_*) : [0,1]\times(0,\infty)\times\mathbb{R} \to [0,4]$ defined in* (4.2.7) *is a discrepancy function, i.e. for $\theta \in [0,1]\times(0,\infty)\times\mathbb{R}$, we have*

$$M(\theta;\vartheta_*) = 0 \qquad \Longleftrightarrow \qquad \theta = \theta_* = \big(p_*,\sigma_*,\mu_*\big)^T\;.$$

Let us reintroduce the covariates. The asymptotic contrast is now given by the function

$$M(\theta, x; \gamma) := \int_{\mathbb{R}} \mathbb{E}_\gamma\big[H^2(Y,t,\theta)\big|X = x\big] q(t)\,\mathrm{d}t \cdot \ell^2(x)\ , \tag{4.2.8}$$

where once again $\mathbb{E}_\gamma$ denotes the expectation with respect to the distribution $\mathbb{P}_\gamma$, which is defined in (4.1.2), i.e.

$$\mathbb{P}_\gamma\big((Y,X)\in A\big) = \int_A f_{Y|X}^{\vartheta_*(\cdot)}(y|x)\ell(x)\,\mathrm{d}(y,x)\ , \qquad \gamma = \big(\theta_*(\cdot), \ell, f_\cdot^*(\cdot)\big)\ .$$

In order to estimate the contrast M, we use an empirical U-statistic estimator localized around x, i.e.

$$M_n(\theta, x; h) = \frac{1}{n\,(n-1)} \sum_{\substack{j,k=1\\ j\neq k}}^{n} \int H(Y_j,t,\theta)H(Y_k,t,\theta)q(t)\,\mathrm{d}t\, K_h(X_j - x)K_h(X_k - x)\ , \tag{4.2.9}$$

where $K : \mathbb{R}^d \to \mathbb{R}$ is a kernel, $h \in (0,\infty)$ is a bandwidth parameter. The contrast functions M and M_n adopt smoothness from the function H whenever the kernel K and the density q are smooth.

Finally, our estimator $\hat\theta_n : I \to \mathbb{R}^3$ for the parameter function $\theta_*(\cdot)$ is proposed by the minimizer of M_n, i.e.

$$\hat\theta_n(x;h) \in \operatorname*{argmin}_{\theta\in\Theta} M_n(\theta, x; h)\ , \tag{4.2.10}$$

which exists for example when Θ is compact since $M_n(\cdot, x; h)$ is continuous.

4.2.3 Uniform rates of convergence

In order to achieve a setting in which the estimator $\hat\theta_n$ defined in (4.2.10) estimates the parameter functions $p_*(\cdot)$, $\mu_*(\cdot)$, $\sigma_*(\cdot)$ with non-parametric uniform convergence rates, we need to make further assumptions on the model. We will only consider those models in which the functions $p_*(\cdot)$, $\mu_*(\cdot)$, $\sigma_*(\cdot)$ and the covariate density ℓ are Hölder-α-smooth as defined in Section 2.4.

Let us denote the parameter space by $\Theta = U_p \times U_\sigma \times U_\mu \subset (1/2,1)\times(0,\infty)\times\mathbb{R}\backslash\{0\}$ for some sets $U_p \subset (1/2,1)$, $U_\sigma \subset (0,\infty)$, $U_\mu \subset \mathbb{R}\backslash\{0\}$ and the set of admissible parameter functions by

$$\begin{aligned}\Delta(\alpha, L, U_p, U_\sigma, U_\mu) := \big\{\theta(\cdot) = \big(p(\cdot),\sigma(\cdot),\mu(\cdot)\big)^T : I \to \Theta\big| \\ p(\cdot)\in H(\alpha,L,U_p),\ \mu(\cdot)\in H(\alpha,L,U_\mu),\ \sigma(\cdot)\in H(\alpha,L,U_\sigma)\big\}\ ,\end{aligned}$$

for any $\alpha, L > 0$. The set of admissible covariate densities is given by

$$\mathcal{L}(\alpha, L, U_\ell) = \big\{\ell \in H(\alpha,L,U_\ell) : \textstyle\int \ell(x)\,\mathrm{d}x = 1\big\}$$

for some $U_\ell \subset (0,\infty)$. Note that under these assumptions, for every $\theta(\cdot) \in \Delta(\alpha, L, U_p, U_\sigma, U_\mu)$ and every $x \in I$, the parameter $\theta(x)$ fulfils Assumption **(I1)**.

Furthermore, we fix some family of component densities $(f_x^*)_{x\in I}$ fulfilling certain smoothness and shape conditions. The model parameters must come from the class

$$\Gamma(\alpha) := \Delta(\alpha, L, U_p, U_\sigma, U_\mu) \times \mathcal{L}(\alpha, L, U_\ell) \times \left\{(f_x^*)_{x\in I}\right\} .$$

The following assumptions will ensure uniform consistency of $\hat{\theta}_n(\cdot; h_n)$ with convergence rate $\left(\frac{n}{\log n}\right)^{\frac{\alpha}{2\alpha+d}}$ for the bandwidth choice $h_n = \left(\frac{\log n}{n}\right)^{\frac{1}{2\alpha+d}}$.

Assumption 4.2.15.

(R1) The sets I, U_p, U_σ, U_μ, U_ℓ are compact cuboids or intervals containing open subsets. For every $x \in I$, the characteristic function $\varphi_{f_x^*}$ of the density f_x^* is positive and for any $y \in \mathbb{R}$, we have $f_\cdot^*(y) \in H(\alpha, L(y), U)$ for some integrable and bounded function $L(\cdot)$ and some compact set $U \subset [0,\infty)$.

(R2) The known component density $\bar{f}$ fulfils Assumption **(I2)** and the maps $y \mapsto y\bar{f}(y)$, $\bar{f}$ and $\partial^2\varphi_{\bar{f}}$ are bounded as well as $\lim_{|t|\to\infty} t\partial\varphi_{\bar{f}}(t) = 0$.

(R3) The kernel $K : \mathbb{R}^d \to \mathbb{R}$ is Lipschitz continuous with Lipschitz constant $L_K > 0$ and has support $[-1,1]^d$.

(R4) The kernel K is of order α, cf. Definition 2.5.1.

(R5) The probability density q has a finite third absolute moment and is bounded.

Theorem 4.2.16. *Under Assumption 4.2.15, given a compact cuboid $J \subset \operatorname{int}(I)$ containing an open subset, if we let $h_n = \left(\frac{\log n}{n}\right)^{\frac{1}{2\alpha+d}}$, the estimator $\hat{\theta}_n(\cdot; h_n)$ has the uniform convergence rate $\left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+d}}$, i.e. we have that*

$$\lim_{\eta\to\infty} \limsup_{n\to\infty} \sup_{\gamma\in\Gamma(\alpha)} \mathbb{P}_\gamma\left(\left(\frac{n}{\log n}\right)^{\frac{\alpha}{2\alpha+d}} \sup_{x\in J} \left\|\hat{\theta}_n(x; h_n) - \theta_*(x)\right\| \geq \eta\right) = 0 .$$

The proof of this result can be found in Section 5.3.2.

Remark 4.2.17.

(i) We only stated results tailored to the identifiability Assumption 4.2.3. If one wishes analogous results under Assumption 4.2.9, one chooses $U_p \subset (0,1)$ and modifies the assumptions on the family of functions $(f_x^*)_{x\in I}$.

(ii) One can also formulate a non-parametric class containing a variety of families of component densities $(f_x^*)_{x\in I}$ and inserting this class into $\Gamma(\alpha)$. Assumptions and notation become quite lengthy. Particularly, compactness of those classes can become an intricate problem. As the unknown component density is not estimated here, we refrain from pursuing this in detail. Considerations for the specific case $\alpha \leq 1$ can be found in the appendix, cf. Section A.5.

4.2.4 Uniform adaptive estimation

Once again, following the Lepski method for the gradients described in Section 3.1.1, we propose an adaptive estimator for parameters coming from

$$\Delta(a, L, U_p, U_\sigma, U_\mu) = \bigcup_{\alpha\in[a,b]} \Delta(\alpha, L, U_p, U_\sigma, U_\mu)$$

for some known $0 < a < b < \infty$, i.e. for the case in which the true smoothness parameter α is only known to come from an interval $[a, b]$. Note that L, U_π, U_σ, U_μ, Θ, I are assumed not to differ for varying α and to have the same properties from the last section. The equality in the display above follows directly from inclusion of the Hölder classes as described in Remark 2.4.5 (i). Particularly, the set of all model parameters is now given by

$$\Gamma(a) = \bigcup_{\alpha\in[a,b]} \Gamma(\alpha) \, .$$

As in Section 3.1.1, define

$$\begin{aligned} S_n(\theta, x; h) &= \partial_\theta M_n(\theta, x; h) \, , \qquad S(\theta, x; \gamma) = \partial_\theta M(\theta, x; \gamma) \, , \\ \beta_k &= a + k\frac{b-a}{N}, \quad k = 0, \dots, N, \quad N = \lceil \log n \rceil \, , \\ h(\alpha) &= \left(\frac{\log n}{n}\right)^{\frac{1}{2\alpha+d}}, \quad r(\alpha) = h(\alpha)^\alpha = \left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+d}}, \quad r_k = r(\beta_k), \quad h_k = h(\beta_k). \end{aligned}$$

Subsequently, define the adaptive data driven grid point by

$$\hat{\alpha}_n = \beta_{\hat{k}},$$

where

$$\hat{k} = \max\left\{0 \le k \le N : \sup_{x\in I, \theta\in\Theta} \left\|S_n(\theta, x; h_k) - S_n(\theta, x; h_l)\right\| \le C_{\mathrm{Lep}} r_l \quad \forall\, 0 \le l \le k\right\} \, ,$$

where the Lepski constant $C_{\mathrm{Lep}} < \infty$ is to be chosen large enough. A specific lower bound on the Lepski constant can be derived from the discussion in the previous chapter, particularly Assumption **(B̃7)** and Lemma 3.2.9.

This construction yields the adaptive estimator

$$\widehat{\theta}_n^{\mathrm{adap}}(x) = \operatorname*{argmin}_{\theta\in\Theta} M_n(\theta, x; h_{\hat{k}}) \, .$$

In order to make use of the highest possible smoothness order b, we need to assume the kernel K to be of order b, cf. Definition 2.5.1. Therefore, we formulate the following alteration of Assumption **(R4)**.

Assumption 4.2.18.

(R̃4) The kernel K is of order b.

Now, we can state the main result.

Theorem 4.2.19. *Let $0 < a < b < \infty$, K be a kernel fulfilling Assumptions* **(R3)** *and* $\widetilde{\mathbf{R4}}$. *Then, under Assumptions* **(R1)**, **(R2)**, **(R5)**, *for any compact cuboid $J \subset \operatorname{int}(I)$ containing an open subset, there is a $C_{\mathrm{Lep}} > 0$ so that*

$$\lim_{\eta\to\infty} \limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \mathbb{P}_\gamma\left(\left(\frac{n}{\log n}\right)^{\frac{\alpha}{2\alpha+d}} \sup_{x\in J} \left\|\widehat{\theta}_n^{\mathrm{adap}}(x) - \theta_*(x)\right\| \geq \eta\right) = 0 \,.$$

In order to prove Theorems 4.2.16 and 4.2.19, it is enough to prove that conditions $\widetilde{\mathbf{B1}}$-$\widetilde{\mathbf{B7}}$ in the appendix hold, cf. Assumption A.2.2. This set of assumptions is equivalent to Assumption 3.1.11 for identified models. A complete proof can be found in Section 5.3. An outline is given here.

Just like in the mixture of normal regressions model in Section 4.1, introduce some bounded, convex and open set

$$\Theta \subset \Xi \subset (1/2,1) \times (0,\infty) \times \mathbb{R}\backslash\{0\} \quad \text{with} \quad \overline{\Xi} \subset (1/2,1) \times (0,\infty) \times \mathbb{R}\backslash\{0\}.$$

$\widetilde{\mathbf{B1}}$ is given by Assumption **(R1)** and Remark 2.4.5 (ii) and (iv). Note that we use the Hölder norm $\|\cdot\|_{a/2}$ once again. The latter part of $\widetilde{\mathbf{B1}}$ is given by Remark 3.1.5 (i).

The continuity of M in $\widetilde{\mathbf{B2}}$ is given by the fact that the function H defined in (4.2.6) is continuous and bounded, which allows for application of Lebesgue's theorem. The contrast property is given by Proposition 4.2.13 and the continuity of the parameter function $\theta_*(\cdot)$ by the coordinates of $\theta_*(\cdot)$ being Hölder-α-smooth.

In order to prove that the Hessian matrix of M evaluated at the true parameter $\theta_*(x)$ is positive definite, cf. $\widetilde{\mathbf{B3}}$, one uses that the function H is zero at the true parameter so that the Hessian matrix V_x is of the form

$$V_x\big(\theta_*(x);\gamma\big) = 2\int \mathbb{E}_\gamma\Big[\partial_\theta H\big(Y,t,\theta_*(x)\big)\Big|X=x\Big]^T \mathbb{E}_\gamma\Big[\partial_\theta H\big(Y,t,\theta_*(x)\big)\Big|X=x\Big] q(t)\,\mathrm{d}t \cdot \ell^2(x) \,.$$

When examining the equation system

$$0 = v^T V_x\big(\theta_*(x);\gamma\big)v \,, \quad v = (v_1,v_2,v_3) \in \mathbb{R}^3$$

one directly derives that for all $t \in \mathbb{R}$,

$$\begin{aligned} 0 &= v_1 \mathbb{E}_\gamma\Big[\partial_p H\big(Y,t,\theta_*(x)\big)\Big|X=x\Big] + v_2 \mathbb{E}_\gamma\Big[\partial_\sigma H\big(Y,t,\theta_*(x)\big)\Big|X=x\Big] \\ &\quad + v_3 \mathbb{E}_\gamma\Big[\partial_\mu H\big(Y,t,\theta_*(x)\big)\Big|X=x\Big] \\ &= -v_1 \varphi_{\bar{f}}\big(\sigma_*(x)t\big)\sin\big(\mu_*(x)t\big) + v_2 t\big(1-p_*(x)\big)\partial\varphi_{\bar{f}}\big(\sigma_*(x)t\big)\sin\big(\mu_*(x)t\big) - v_3 t p_*(x)\varphi_{f_x^*}(t) \,. \end{aligned} \tag{4.2.11}$$

The first and second summand in (4.2.11) are zero for $t \in \frac{\pi}{\mu_*(x)}\mathbb{Z}$ so that $v_3 = 0$. Differentiating the remaining summands in t and examining the behaviour in a neighbourhood

around 0 yields $v_1 = v_2 = 0$ as well.

In this model, the Lipschitz continuity of the Hessian matrix in **($\widetilde{\mathbf{B4}}$)** is a matter of direct calculation.

In order to derive the uniform L^2 convergence rates in **($\widetilde{\mathbf{B5}}$)**, the common technique of decomposing the estimation error in squared bias and variance works. The bias term is mainly handled by the convolution technique described in Lemma 3.2.1. The variance term is treated directly by applying Theorem 3.2.7 for $\rho = 2$ to the gradients. The conditions in the theorem are easily verified as H is smooth and bounded. Subsequently, **($\widetilde{\mathbf{B7}}$)** is directly given by Lemma 3.2.10.

As H is bounded, Theorem 3.2.7 applied to the contrast functions M_n yields that the contrast's uniform stochastic estimation error fulfils

$$\sup_{\alpha\in[a,b]}\sup_{\gamma\in\Gamma(a)}\mathbb{P}_\gamma\Big(\sup_{\theta,x}\big|M_n\big(\theta,x;h_n(\alpha)\big)-\mathbb{E}_\gamma\big[M_n\big(\theta,x;h_n(\alpha)\big)\big]\big|\geq\eta\Big)=o(1)\;,\quad\eta>0$$

by Markov's inequality. A direct calculation gives that the bias converges to zero as well so that the contrast M_n is uniformly consistent. Hence, **($\widetilde{\mathbf{B6}}$)** is fulfilled as well and Theorems 3.1.6 and 3.1.13 yield the desired results.

5 Proofs and auxiliary results

5.1 Proofs for Chapter 3

Proof of Theorem 3.1.2. Because of $\hat{\theta}_n(x;\gamma) = \operatorname{argmin}_\theta M_n(\theta, x;\gamma)$ and M being constant on $\mathfrak{S}_{x;\gamma}$, we have for all, $\gamma \in \Gamma$, $\theta_*(x;\gamma) \in \mathfrak{S}_{x;\gamma}$,

$$\begin{aligned}
0 &\le \sup_{x\in I}\Big[M\big(\hat{\theta}_n(x;\gamma),x;\gamma\big) - M\big(\theta_*(x;\gamma),x;\gamma\big)\Big] \\
&\le \sup_{x\in I}\Big[M\big(\hat{\theta}_n(x;\gamma),x;\gamma\big) - M_n\big(\hat{\theta}_n(x;\gamma),x;\gamma\big)\Big] \\
&\quad + \sup_{x\in I}\Big[M_n\big(\hat{\theta}_n(x;\gamma),x;\gamma\big) - M\big(\theta_*(x;\gamma),x;\gamma\big)\Big] \\
&\le \sup_{\theta\in\Theta}\sup_{x\in I}|M(\theta,x;\gamma) - M_n(\theta,x;\gamma)| + \sup_{x\in I}\Big[M_n\big(\theta_*(x;\gamma),x;\gamma\big) - M\big(\theta_*(x;\gamma),x;\gamma\big)\Big] \\
&\le 2\sup_{\theta\in\Theta}\sup_{x\in I}|M(\theta,x;\gamma) - M_n(\theta,x;\gamma)| \,.
\end{aligned}$$

Fix $\varepsilon > 0$. Because of $(*)$ there is an $\eta > 0$ so that for any $\gamma \in \Gamma$, $\theta_*(x;\gamma) \in \mathfrak{S}_{x;\gamma}$, the inequality

$$\sup_{x\in I}\Big[M(\hat{\theta}_n(x;\gamma),x;\gamma) - M(\theta_*(x;\gamma),x;\gamma)\Big] < \eta$$

implies

$$\sup_{x\in I}\min_{\theta_*\in\mathfrak{S}_{x;\gamma}}\big\|\hat{\theta}_n(x;\gamma) - \theta_*\big\| < \varepsilon \,,$$

giving

$$\begin{aligned}
\Big\{\sup_{x\in I}\min_{\theta_*\in\mathfrak{S}_{x;\gamma}}\big\|\hat{\theta}_n(x;\gamma) - \theta_*\big\| \ge \varepsilon\Big\} &\subset \Big\{\sup_{x\in I}\Big[M(\hat{\theta}_n(x;\gamma),x;\gamma) - M(\theta_*(x;\gamma),x;\gamma)\Big] \ge \eta\Big\} \\
&\subset \Big\{\sup_{\theta\in\Theta}\sup_{x\in I}|M(\theta,x;\gamma) - M_n(\theta,x;\gamma)| \ge \eta/2\Big\} \,.
\end{aligned}$$

Thus, by uniform consistency of the random functions M_n,

$$\lim_{n\to\infty}\sup_{\gamma\in\Gamma}\mathbb{P}_\gamma\Big(\sup_{x\in I}\min_{\theta_*\in\mathfrak{S}_{x;\gamma}}\big\|\hat{\theta}_n(x;\gamma) - \theta_*\big\| \ge \varepsilon\Big) = 0 \,.$$

□

The general idea for the proof of Theorem 3.1.3 is similar to the one of van der Vaart and Wellner (1996, Theorem 3.2.5) but the details are a bit more involved.

Proof of Theorem 3.1.3. For every $n \in \mathbb{N}$, $x \in I$, $\gamma \in \Gamma$, we can define a disjoint partition of $\Theta \backslash \mathfrak{S}_{x;\gamma}$ by $\bigcup_{j \in \mathbb{Z}} S_{jnx\gamma}$, where

$$S_{jnx\gamma} = \left\{ \theta \in \Theta : 2^{j-1} < r_{n,\gamma} \min_{\theta_* \in \mathfrak{S}_{x;\gamma}} \|\theta - \theta_*\| \leq 2^j \right\}.$$

Let us define for any $N, n \in \mathbb{N}$, $\gamma \in \Gamma$ the sets

$$A_{Nn\gamma} := \left\{ r_{n,\gamma} \sup_{x \in I} \min_{\theta_* \in \mathfrak{S}_{x;\gamma}} \left\| \hat{\theta}_n(x;\gamma) - \theta_* \right\| > 2^N \right\}$$

and show that $\lim_{N\to\infty} \limsup_{n\to\infty} \sup_{\gamma\in\Gamma} \mathbb{P}_\gamma(A_{Nn\gamma}) = 0$. In order to do that, we show for any $\eta > 0$ the inequality

$$\begin{aligned} \mathbb{P}_\gamma(A_{Nn\gamma}) \leq & \sum_{\substack{j \geq N \\ 2^j \leq \eta r_{n,\gamma}}} \mathbb{P}_\gamma\left(\sup_{x \in I} \sup_{\theta \in S_{jnx\gamma}} \min_{\theta_* \in \mathfrak{S}_{x;\gamma}} \left[M_n(\theta_*, x;\gamma) - M_n(\theta, x;\gamma) \right] \geq 0 \right) \\ & + \mathbb{P}_\gamma\left(2 \sup_{x \in I} \min_{\theta_* \in \mathfrak{S}_{x;\gamma}} \left\| \hat{\theta}_n(x;\gamma) - \theta_* \right\| \geq \eta \right). \end{aligned} \tag{5.1.1}$$

Therefore, let

$$\begin{aligned} \omega \in & \bigcap_{\substack{j \geq N \\ 2^j \leq \eta r_{n,\gamma}}} \left\{ \sup_{x \in I} \sup_{\theta \in S_{jnx\gamma}} \min_{\theta_* \in \mathfrak{S}_{x;\gamma}} \left[M_n(\theta_*, x;\gamma) - M_n(\theta, x;\gamma) \right] < 0 \right\} \\ & \cap \left\{ 2 \sup_{x \in I} \min_{\theta_* \in \mathfrak{S}_{x;\gamma}} \left\| \hat{\theta}_n(x;\gamma) - \theta_* \right\| < \eta \right\}. \end{aligned} \tag{5.1.2}$$

Then, for all $j \geq N$ with $2^j \leq \eta r_{n,\gamma}$ and all $x \in I$ we have $\hat{\theta}_n(x;\gamma)(\omega) \notin S_{jnx\gamma}$ because $\hat{\theta}_n(x;\gamma)$ minimizes $M_n(\cdot, x;\gamma)$. Hence, for all $x \in I$, either

$$r_{n,\gamma} \min_{\theta_* \in \mathfrak{S}_{x;\gamma}} \left\| \hat{\theta}_n(x;\gamma)(\omega) - \theta_* \right\| \leq 2^{N-1} \quad \text{or} \quad r_{n,\gamma} \min_{\theta_* \in \mathfrak{S}_{x;\gamma}} \left\| \hat{\theta}_n(x;\gamma)(\omega) - \theta_* \right\| > 2^{l_\gamma}, \tag{5.1.3}$$

where $l_\gamma = \max\{j \geq N : 2^j \leq \eta r_{n,\gamma}\}$ if such an l_γ exists. The latter case needs to be disproved. Therefore, assume that for some $x \in I$, $J \geq N$ with $2^J > \eta r_{n,\gamma}$, we have

$$r_{n,\gamma} \min_{\theta_* \in \mathfrak{S}_{x;\gamma}} \left\| \hat{\theta}_n(x;\gamma)(\omega) - \theta_* \right\| > 2^{J-1}.$$

Then,

$$2^{J-1} < r_{n,\gamma} \min_{\theta_* \in \mathfrak{S}_{x;\gamma}} \left\| \hat{\theta}_n(x;\gamma)(\omega) - \theta_* \right\| < r_{n,\gamma}\eta/2 < 2^{J-1},$$

according to the right-hand side of (5.1.2), a contradiction. Hence,

$$r_{n,\gamma} \sup_{x \in I} \min_{\theta_* \in \mathfrak{S}_{x;\gamma}} \left\| \hat{\theta}_n(x;\gamma)(\omega) - \theta_* \right\| \leq 2^{N-1} \leq 2^N$$

according to (5.1.3), giving $\omega \in A^c_{Nn\gamma}$ and via subadditivity we deduce (5.1.1). The second summand on the right-hand side of (5.1.1) converges uniformly over all $\gamma \in \Gamma$ to zero for all $\eta > 0$ according to assumption (ii), i.e.

$$\lim_{n\to\infty} \sup_{\gamma\in\Gamma} \mathbb{P}_\gamma \left(2 \sup_{x\in I} \min_{\theta_*\in\mathfrak{S}_{x;\gamma}} \left\|\hat{\theta}_n(x;\gamma) - \theta_*\right\| \geq \eta \right) = 0 \,. \tag{5.1.4}$$

Hence, we only need to handle the first summand. Choose $\eta > 0$ so that (i) is fulfilled. Then, because every $\theta_*(x;\gamma) \in \mathfrak{S}_{x;\gamma}$ minimizes $M(\cdot, x;\gamma)$, for any $j \geq N$ so that $2^j \leq \eta r_{n,\gamma}$, which is equivalent to $2^j/r_{n,\gamma} \leq \eta$ and any $\gamma \in \Gamma$, we have

$$\begin{aligned}
&\sup_{x\in I} \sup_{\theta\in S_{jnx\gamma}} \min_{\theta_*\in\mathfrak{S}_{x;\gamma}} \Big[M(\theta_*, x;\gamma) - M(\theta, x;\gamma) \Big] \\
&= \sup_{x\in I} \sup_{\varepsilon\in\left(2^{j-1}/r_{n,\gamma}, 2^j/r_{n,\gamma}\right]} \sup_{*} \Big[M\big(\theta_*(x;\gamma), x;\gamma\big) - M(\theta, x;\gamma) \Big] \\
&= \sup_{\varepsilon\in\left(2^{j-1}/r_{n,\gamma}, 2^j/r_{n,\gamma}\right]} \sup_{x\in I} \sup_{*} \Big[M\big(\theta_*(x;\gamma), x;\gamma\big) - M(\theta, x;\gamma) \Big] \\
&\leq C_1 \sup_{\varepsilon\in\left(2^{j-1}/r_{n,\gamma}, 2^j/r_{n,\gamma}\right]} -\varepsilon^2 = -C_1 \left(\frac{2^{j-1}}{r_{n,\gamma}} \right)^2
\end{aligned}$$

according to the first part of (i), where the suprema indexed with $*$ are taken over $\big\{\theta \in \Theta : \min_{\theta_*\in\mathfrak{S}_{x;\gamma}} \|\theta - \theta_*\| = \varepsilon\big\}$. Now, the fact that $M(\cdot, x;\gamma)$ is constant on $\mathfrak{S}_{x;\gamma}$, (5.1.1), (5.1.4), the display above and Markov's inequality give

$$\begin{aligned}
&\lim_{N\to\infty} \limsup_{n\to\infty} \sup_{\gamma\in\Gamma} \mathbb{P}_\gamma(A_{Nn\gamma}) \\
&= \lim_{N\to\infty} \limsup_{n\to\infty} \sup_{\gamma\in\Gamma} \mathbb{P}_\gamma\Big(r_{n,\gamma} \sup_{x\in I} \min_{\theta_*\in\mathfrak{S}_{x;\gamma}} \left\|\hat{\theta}_n(x;\gamma) - \theta_*\right\| \geq 2^N \Big) \\
&\leq \lim_{N\to\infty} \limsup_{n\to\infty} \sup_{\gamma\in\Gamma} \sum_{\substack{j\geq N \\ 2^j\leq \eta r_{n,\gamma}}} \mathbb{P}_\gamma\Big(\sup_{x\in I} \sup_{\theta\in S_{jnx\gamma}} \min_{\theta_*\in\mathfrak{S}_{x;\gamma}} \Big[M_n(\theta_*, x;\gamma) - M_n(\theta, x;\gamma) \Big] \geq 0 \Big) \\
&= \lim_{N\to\infty} \limsup_{n\to\infty} \sup_{\gamma\in\Gamma} \sum_{\substack{j\geq N \\ 2^j\leq \eta r_{n,\gamma}}} \mathbb{P}_\gamma\Big(\sup_{x\in I} \sup_{\theta\in S_{jnx\gamma}} \min_{\theta_*\in\mathfrak{S}_{x;\gamma}} \Big[W_n(\theta_*, x;\gamma) - W_n(\theta, x;\gamma) \\
&\qquad\qquad + M(\theta_*, x;\gamma) - M(\theta, x;\gamma) \Big] \geq 0 \Big) \\
&\leq \lim_{N\to\infty} \limsup_{n\to\infty} \sup_{\gamma\in\Gamma} \sum_{\substack{j\geq N \\ 2^j\leq \eta r_{n,\gamma}}} \mathbb{P}_\gamma\Big(\sup_{x\in I} \sup_{\theta\in S_{jnx\gamma}} \min_{\theta_*\in\mathfrak{S}_{x;\gamma}} |W_n(\theta_*, x;\gamma) - W_n(\theta, x;\gamma)| \\
&\qquad\qquad + \sup_{x\in I} \sup_{\theta\in S_{jnx\gamma}} \min_{\theta_*\in\mathfrak{S}_{x;\gamma}} \Big[M(\theta_*, x;\gamma) - M(\theta, x;\gamma) \Big] \geq 0 \Big) \\
&\leq \lim_{N\to\infty} \limsup_{n\to\infty} \sup_{\gamma\in\Gamma} \sum_{\substack{j\geq N \\ 2^j\leq \eta r_{n,\gamma}}} \mathbb{P}_\gamma\Big(\sup_{x\in I} \sup_{\theta\in S_{jnx\gamma}} \min_{\theta_*\in\mathfrak{S}_{x;\gamma}} |W_n(\theta_*, x;\gamma) - W_n(\theta, x;\gamma)| \geq C_1 \frac{2^{2j-2}}{r^2_{n,\gamma}} \Big)
\end{aligned}$$

$$\leq \lim_{N\to\infty} \limsup_{n\to\infty} \sum_{j\geq N} \sup_{\gamma\in\Gamma} \mathbb{1}_{\frac{2^j}{r_{n,\gamma}}\leq\eta} \frac{r_{n,\gamma}^2}{C_1 2^{2j-2}} \mathbb{E}_\gamma \left[\sup_{x\in I} \sup_{\theta\in S_{jnx\gamma}} \min_{\theta_*\in\mathfrak{S}_{x;\gamma}} |W_n(\theta_*,x;\gamma) - W_n(\theta,x;\gamma)| \right]. \tag{5.1.5}$$

We will use the second point in (i) in order to treat the $\limsup_{n\to\infty}$ term in (5.1.5) for fixed $N \in \mathbb{N}$ by Fatou's lemma for the counting measure on $\{N, N+1, \ldots\}$. To be precise, we need to show that the summands in (5.1.5) are uniformly bounded in $n \geq n_0$ by a function in $j \geq N$ that is summable for some $n_0 \in \mathbb{N}$.

The second point in (i) gives

$$\limsup_{n\to\infty} \sup_{\gamma\in\Gamma} \sup_{\substack{j\geq N \\ \frac{2^j}{r_{n,\gamma}}\leq\eta}} \frac{t_{n,\gamma}}{\phi_n(2^j/r_{n,\gamma})} \mathbb{E}_\gamma \left[\sup_{x\in I} \sup_{\theta\in S_{jnx\gamma}} \min_{\theta_*\in\mathfrak{S}_{x;\gamma}} |W_n(\theta_*,x;\gamma) - W_n(\theta,x;\gamma)| \right] \leq C_2.$$

In particular, for any $\kappa > 0$, there is an $n_0 \in \mathbb{N}$ so that for every $\gamma \in \Gamma$, $n \geq n_0$, $j \geq N$ with $2^j \leq \eta r_{n,\gamma}$, we have

$$\mathbb{E}_\gamma \left[\sup_{x\in I} \sup_{\theta\in S_{jnx\gamma}} \min_{\theta_*\in\mathfrak{S}_{x;\gamma}} |W_n(\theta_*,x;\gamma) - W_n(\theta,x;\gamma)| \right] \leq \frac{\big(C_2+\kappa\big)\phi_n(2^j/r_{n,\gamma})}{t_{n,\gamma}} .$$

Hence, for every $\gamma \in \Gamma$, $n \geq n_0$, $j \geq N$ with $2^j \leq \eta r_{n,\gamma}$, the summands in (5.1.5) can be treated by

$$\begin{aligned} &\frac{r_{n,\gamma}^2}{C_1 2^{2j-2}} \mathbb{E}_\gamma \left[\sup_{x\in I} \sup_{\theta\in S_{jnx\gamma}} \min_{\theta_*\in\mathfrak{S}_{x;\gamma}} |W_n(\theta_*,x;\gamma) - W_n(\theta,x;\gamma)| \right] \\ &\leq \frac{r_{n,\gamma}^2}{C_1 2^{2j-2}} \frac{\big(C_2+\kappa\big)\phi_n(2^j/r_{n,\gamma})}{t_{n,\gamma}} . \end{aligned} \tag{5.1.6}$$

Since the function $\phi_n(\cdot)/\cdot^\alpha$ is decreasing, for any $z \geq 1$, $y > 0$, we have

$$\frac{\phi_n(zy)}{z^\alpha y^\alpha} \leq \frac{\phi_n(y)}{y^\alpha} \quad \text{so that} \quad \phi_n(zy) \leq z^\alpha \phi_n(y) .$$

As $2^j \geq 1$, this implies

$$(5.1.6) \leq \frac{r_{n,\gamma}^2 \big(C_2+\kappa\big) 2^{j\alpha} \phi_n(1/r_{n,\gamma})}{C_1 2^{2j-2} t_{n,\gamma}} \leq \frac{4\big(C_2+\kappa\big)}{C_1} \cdot \left(\frac{1}{2^{2-\alpha}}\right)^j ,$$

which clearly is summable in $j \geq N$. Hence, we can apply Fatou's lemma, so that for some κ independent of N, we have

$$\begin{aligned} &(5.1.5) \\ &\leq \lim_{N\to\infty} \sum_{j\geq N} \limsup_{n\to\infty} \sup_{\gamma\in\Gamma} \mathbb{1}_{\frac{2^j}{r_{n,\gamma}}\leq\eta} \frac{r_{n,\gamma}^2}{C_1 2^{2j-2}} \mathbb{E}_\gamma \left[\sup_{x\in I} \sup_{\theta\in S_{jnx\gamma}} \min_{\theta_*\in\mathfrak{S}_{x;\gamma}} |W_n(\theta_*,x;\gamma) - W_n(\theta,x;\gamma)| \right] \\ &\leq \lim_{N\to\infty} \sum_{j\geq N} \frac{4\big(C_2+\kappa\big)}{C_1} \cdot \left(\frac{1}{2^{2-\alpha}}\right)^j = 0 . \end{aligned}$$

□

Proofs for the uniform stochastic L^ρ-errors

The first technical lemma gives a discretization method that will be used for deriving uniform L^ρ-errors.

Lemma 5.1.1. *Let $J \subset \mathbb{R}^d$ be a compact cuboid, $\Theta \subset \mathbb{R}^m$ be compact and convex with $\Theta = \overline{\mathrm{int}(\Theta)}$, $\delta_n \to 0$ be a zero-sequence, $T_n : \Theta \times I \to \mathbb{R}$ be functions. Then there are nets $\Theta_n \times J_n \subset \Theta \times J$ so that*

$$\sup_{x\in J}\inf_{y\in J_n}\|x-y\| < \delta_n\,, \qquad \sup_{\vartheta\in\Theta}\inf_{\theta\in\Theta_n}\|\vartheta-\theta\| < \delta_n\,, \qquad \#(\Theta_n\times J_n) \le C_{\Theta,J}\delta_n^{-d-m}\,,$$

where $C_{\Theta,J}$ is independent of n. Furthermore,

$$\begin{aligned}&\sup_{x\in J,\theta\in\Theta}|T_n(\theta,x)|^\rho\\ &\le 2^{\rho-1}\Big(\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\le\delta_n}}|T_n(\theta,x)-T_n(\vartheta,y)|\Big)^\rho + 2^{\rho-1}\Big(\sup_{x\in J_n,\vartheta\in\Theta_n}|T_n(\vartheta,x)|\Big)^\rho\,.\end{aligned}$$

Proof of Lemma 5.1.1. The set $\Theta\times J$ is compact, hence totally bounded, so that for any $n\in\mathbb{N}$ there are finite sets $J_n \subset J, \Theta_n \subset \Theta$ so that

$$\sup_{x\in J}\inf_{y\in J_n}\|x-y\| < \delta_n\,, \qquad \sup_{\vartheta\in\Theta}\inf_{\theta\in\Theta_n}\|\vartheta-\theta\| < \delta_n$$

and we can choose $\Theta_n\times J_n$ so that

$$\#(\Theta_n\times J_n) \le C_{\Theta,J}\delta_n^{-d-m}\,,$$

and $C_{\Theta,J}$ is independent of n. That is because J is a cuboid and for Θ we can examine a superset that is a compact cuboid. Since for any monotone and continuous function $\varrho : A\to\mathbb{R}$, where $A\subset\mathbb{R}$ is compact, we have

$$\varrho(\sup A) = \sup_{x\in A}\varrho(x)\,, \qquad \varrho(\inf A) = \inf_{x\in A}\varrho(x)$$

and because $s\mapsto s^\rho$ is monotone, convex and especially continuous on $[0,\infty)$, we achieve the representation

$$\begin{aligned}&\sup_{x\in J,\theta\in\Theta}|T_n(\theta,x)|^\rho\\ &= \Big(\sup_{x\in J,\theta\in\Theta}|T_n(\theta,x)|\Big)^\rho\\ &= \Big(\underbrace{\sup_{x\in J,\theta\in\Theta}|T_n(\theta,x)| - \sup_{y\in J_n,\vartheta\in\Theta_n}|T_n(\vartheta,y)| + \sup_{y\in J_n,\vartheta\in\Theta_n}|T_n(\vartheta,y)|}_{\ge 0}\Big)^\rho\\ &\le 2^{\rho-1}\Big|\sup_{x\in J,\theta\in\Theta}|T_n(\theta,x)| - \sup_{y\in J_n,\vartheta\in\Theta_n}|T_n(\vartheta,y)|\Big|^\rho + 2^{\rho-1}\Big(\sup_{x\in J_n,\vartheta\in\Theta_n}|T_n(\vartheta,x)|\Big)^\rho\,.\end{aligned} \tag{5.1.7}$$

The first summand of (5.1.7) is handled by the fact that the term's base is non-negative, that $s \mapsto s^\rho$ is monotone and that

$$\begin{aligned}
\sup_{x\in J,\theta\in\Theta}|T_n(\theta,x)| - \sup_{y\in J_n,\vartheta\in\Theta_n}|T_n(\vartheta,y)| &= \sup_{x\in J,\theta\in\Theta}\Big[|T_n(\theta,x)| - \sup_{y\in J_n,\vartheta\in\Theta_n}|T_n(\vartheta,y)|\Big] \\
&= \sup_{x\in J,\theta\in\Theta}\inf_{y\in J_n,\vartheta\in\Theta_n}\Big[|T_n(\theta,x)| - |T_n(\vartheta,y)|\Big] \\
&\le \sup_{x\in J,\theta\in\Theta}\inf_{y\in J_n,\vartheta\in\Theta_n}\Big||T_n(\theta,x)| - |T_n(\vartheta,y)|\Big| \\
&\le \sup_{x\in J,\theta\in\Theta}\inf_{y\in J_n,\vartheta\in\Theta_n}\Big|T_n(\theta,x) - T_n(\vartheta,y)\Big| \\
&\le \sup_{x\in J,\theta\in\Theta}\inf_{\substack{y\in J_n,\vartheta\in\Theta_n\\ \|x-y\|,\|\vartheta-\theta\|\le\delta_n}}\Big|T_n(\theta,x) - T_n(\vartheta,y)\Big| \\
&\le \sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\le\delta_n}}|T_n(\theta,x) - T_n(\vartheta,y)|
\end{aligned}$$

by means of the reverse triangular inequality. □

Proof of Theorem 3.2.5. We will drop dependence of the bandwidth parameters on α and n for convenience but only give a hint where it comes into play. Note that throughout the proof, we will only use the notation $a_n \lesssim b_n$ if there is a constant $C > 0$ and an $n_0 \in \mathbb{N}$ so that for all $n \ge n_0$ we have $a_n \le Cb_n$ and the constant depends only on Γ, C_1, C_2, C_τ, τ, $\|K\|_\infty$, L_K, ρ, I, Θ, A or is universal. Also note that all calculations below hold independently of $\gamma \in \Gamma$. Let us define

$$\begin{aligned}
T_n(\theta,x;h) :=& M_n(\theta,x;h) - \mathbb{E}_\gamma\big[M_n(\theta,x;h)\big] \\
=& \frac{1}{n}\sum_{k=1}^n \tau(Y_k,\theta)K_h(X_k-x) - \mathbb{E}_\gamma[\tau(Y_k,\theta)K_h(X_k-x)]\,.
\end{aligned}$$

By using Lemma 5.1.1, for a sequence $\delta_n \to 0$ specified later, there are nets $\Theta_n \times J_n \subset \Theta \times J$ so that

$$\sup_{x\in J}\inf_{y\in J_n}\|x-y\| < \delta_n\,, \qquad \sup_{\vartheta\in\Theta}\inf_{\theta\in\Theta_n}\|\vartheta-\theta\| < \delta_n\,, \qquad \#(\Theta_n\times J_n) \le C_{\Theta,J}\delta_n^{-d-m}\,,$$

where $C_{\Theta,J}$ is independent of n. We further achieve the representation

$$\begin{aligned}
&\sup_{x\in J,\theta\in\Theta}|T_n(\theta,x;h)|^\rho \\
\le& 2^{\rho-1}\bigg(\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\le\delta_n}}|T_n(\theta,x;h) - T_n(\vartheta,y;h)|\bigg)^\rho + 2^{\rho-1}\Big(\sup_{x\in J_n,\vartheta\in\Theta_n}|T_n(\vartheta,x;h)|\Big)^\rho\,.
\end{aligned} \tag{5.1.8}$$

Let us treat the first summand of (5.1.8). By Jensen's inequality for sums and monotonicity and continuity of $s \mapsto s^\rho$ on $(0,\infty)$, we deduce

$$\mathbb{E}_\gamma\bigg[\bigg(\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\le\delta_n}}|T_n(\theta,x;h) - T_n(\vartheta,y;h)|\bigg)^\rho\bigg]$$

$$
\begin{aligned}
\leq\ & 4^{\rho-1}\Bigg\{\mathbb{E}_\gamma\Bigg[\Bigg(\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\leq\delta_n}}\Bigg|\frac{1}{n}\sum_{k=1}^{n}\big(\tau(Y_k,\vartheta)-\tau(Y_k,\theta)\big)K_h(X_k-x)\Bigg|\Bigg)^{\rho}\Bigg]\\
&+\mathbb{E}_\gamma\Bigg[\Bigg(\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\leq\delta_n}}\Bigg|\frac{1}{n}\sum_{k=1}^{n}\tau(Y_k,\theta)\big(K_h(X_k-x)-K_h(X_k-y)\big)\Bigg|\Bigg)^{\rho}\Bigg]\\
&+\mathbb{E}_\gamma\Bigg[\Bigg(\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\leq\delta_n}}\Bigg|\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}_\gamma\Big[\big(\tau(Y_k,\vartheta)-\tau(Y_k,\theta)\big)K_h(X_k-x)\Big]\Bigg|\Bigg)^{\rho}\Bigg]\\
&+\mathbb{E}_\gamma\Bigg[\Bigg(\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\leq\delta_n}}\Bigg|\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}_\gamma\Big[\tau(Y_k,\theta)\big(K_h(X_k-x)-K_h(X_k-y)\big)\Big]\Bigg|\Bigg)^{\rho}\Bigg]\Bigg\}\\
\leq\ & 4^{\rho-1}\Bigg\{\mathbb{E}_\gamma\Bigg[\Bigg(\frac{1}{n}\sum_{k=1}^{n}\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\leq\delta_n}}\big|\big(\tau(Y_k,\vartheta)-\tau(Y_k,\theta)\big)K_h(X_k-x)\big|\Bigg)^{\rho}\Bigg] && (5.1.9)\\
&+\mathbb{E}_\gamma\Bigg[\Bigg(\frac{1}{n}\sum_{k=1}^{n}\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\leq\delta_n}}\big|\tau(Y_k,\theta)\big(K_h(X_k-x)-K_h(X_k-y)\big)\big|\Bigg)^{\rho}\Bigg] && (5.1.10)\\
&+\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\leq\delta_n}}\Big(\mathbb{E}_\gamma\Big[\big|\big(\tau(Y_1,\vartheta)-\tau(Y_1,\theta)\big)K_h(X_1-x)\big|\Big]\Big)^{\rho} && (5.1.11)\\
&+\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\leq\delta_n}}\Big(\mathbb{E}_\gamma\Big[\big|\tau(Y_1,\theta)\big(K_h(X_1-x)-K_h(X_1-y)\big)\big|\Big]\Big)^{\rho}\Bigg\}. && (5.1.12)
\end{aligned}
$$

Again by Jensen's inequality for sums,

$$
\begin{aligned}
&(5.1.9)+(5.1.10)\\
\leq\ &\mathbb{E}_\gamma\Bigg[\frac{1}{n}\sum_{k=1}^{n}\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\leq\delta_n}}\big|\big(\tau(Y_k,\vartheta)-\tau(Y_k,\theta)\big)K_h(X_k-x)\big|^{\rho}\Bigg]\\
&+\mathbb{E}_\gamma\Bigg[\frac{1}{n}\sum_{k=1}^{n}\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\leq\delta_n}}\big|\tau(Y_k,\theta)\big(K_h(X_k-x)-K_h(X_k-y)\big)\big|^{\rho}\Bigg]\\
=\ &\mathbb{E}_\gamma\Bigg[\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\leq\delta_n}}\big|\big(\tau(Y_1,\vartheta)-\tau(Y_1,\theta)\big)K_h(X_1-x)\big|^{\rho}\Bigg] && (5.1.13)\\
&+\mathbb{E}_\gamma\Bigg[\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\leq\delta_n}}\big|\tau(Y_1,\theta)\big(K_h(X_1-x)-K_h(X_1-y)\big)\big|^{\rho}\Bigg]. && (5.1.14)
\end{aligned}
$$

By Jensen's inequality, monotonicity and continuity of $s\mapsto s^\rho$ on $(0,\infty)$ and (3.2.13), we get

$$
(5.1.11)\leq(5.1.13)\leq\frac{\delta_n^\rho}{h^{\rho d}}4^{\rho-1}\|K\|_\infty^\rho\int\sup_{\theta,\tilde\theta\in\Theta}\Psi_\tau^\rho(y,\theta,\tilde\theta)\sup_{\gamma\in\Gamma,x\in I}f_\gamma(y|x)\mathrm{d}y=O\Big(\frac{\delta_n^\rho}{h^{\rho d}}\Big)
$$

and additionally applying Lipschitz continuity of the kernel K and (3.2.12), we deduce

$$(5.1.12) \leq (5.1.14) \leq \frac{\delta_n^\rho}{h^{\rho(d+1)}} 4^{\rho-1} L_K^\rho C_\tau = O\left(\frac{\delta_n^\rho}{h^{\rho(d+1)}}\right).$$

This gives a bound for the discretizing error in (5.1.8), i.e.

$$\mathbb{E}_\gamma\left[\Big| \sup_{x\in J,\vartheta\in\Theta} |T_n(\vartheta,x;h)| - \sup_{y\in J_n,\theta\in\Theta_n} |T_n(\theta,y;h)|\Big|^\rho\right] = O\left(\frac{\delta_n^\rho}{h^{\rho(d+1)}}\right). \tag{5.1.15}$$

Now, let us focus on the discretized term $\sup_{x\in J_n,\theta\in\Theta_n} |T_n(\theta,x;h)|$ in (5.1.8). According to (3.2.15), for any fixed θ, x, h, n, we have that

$$\begin{aligned}\mathbb{P}_\gamma\left(|T_n(\theta,x;h)| > \omega\left(\frac{\log n}{nh^d}\right)^{\frac{1}{2}}\right) &\leq 2\exp\left(-\frac{\omega^2\frac{\log n}{nh^d}\cdot nh^d}{C_1+\omega C_2\left(\frac{\log n}{nh^d}\right)^{\frac{1}{2}}}\right)\\ &= 2\exp\left(-\frac{\omega^2\log n}{C_1+\omega C_2\left(\frac{\log n}{nh^d}\right)^{\frac{1}{2}}}\right), \quad \omega>0\,.\end{aligned}$$

We will apply the estimate

$$\mathbb{E}[|W|^\rho] \leq \int_0^\infty \rho\omega^{\rho-1}\mathbb{P}(|W|>\omega)\,\mathrm{d}\omega \leq a^\rho + \int_a^\infty \rho\omega^{\rho-1}\mathbb{P}(|W|>\omega)\,\mathrm{d}\omega\,, \quad a>0$$

to $W = \sup_{x\in J_n,\theta\in\Theta_n} |T_n(\theta,x;h)|\big(\frac{nh^d}{\log n}\big)^{\frac{1}{2}}$ in order to obtain a bound on the remainder in (5.1.8) as follows.

$$\begin{aligned}&\mathbb{E}_\gamma\left[\sup_{x\in J_n,\theta\in\Theta_n}|T_n(\theta,x;h)|^\rho\right]\\ \leq&\Big(\frac{\log n}{nh^d}\Big)^{\frac{\rho}{2}}\left[a^\rho+\int_a^\infty \rho\omega^{\rho-1}\mathbb{P}_\gamma\Big(\sup_{x\in J_n,\theta\in\Theta_n}|T_n(\theta,x;h)|>\Big(\frac{\log n}{nh^d}\Big)^{\frac{1}{2}}\omega\Big)\,\mathrm{d}\omega\right]\\ \leq&\Big(\frac{\log n}{nh^d}\Big)^{\frac{\rho}{2}}\left[a^\rho+\int_a^\infty \rho\omega^{\rho-1}\sum_{x\in J_n,\theta\in\Theta_n}\mathbb{P}_\gamma\Big(|T_n(\theta,x;h)|>\Big(\frac{\log n}{nh^d}\Big)^{\frac{1}{2}}\omega\Big)\,\mathrm{d}\omega\right]\\ \leq&\Big(\frac{\log n}{nh^d}\Big)^{\frac{\rho}{2}}\left[a^\rho+2C_{\Theta,J}\delta_n^{-d-m}\rho\int_a^\infty \omega^{\rho-1}\exp\left(-\frac{\omega^2\log n}{C_1+\omega C_2\left(\frac{\log n}{nh^d}\right)^{\frac{1}{2}}}\right)\mathrm{d}\omega\right]\\ \leq&\Big(\frac{\log n}{nh^d}\Big)^{\frac{\rho}{2}}\left[a^\rho+2C_{\Theta,J}\delta_n^{-d-m}\rho\int_a^\infty \omega^{\rho-1}\exp\Big(-\frac{\omega\log n}{C_a}\Big)\,\mathrm{d}\omega\right]\end{aligned} \tag{5.1.16}$$

for $C_a = \frac{C_1}{a}+1$ and n so large that $\left(\frac{\log n}{nh^d}\right)^{\frac{1}{2}} C_2 \leq 1$. Because then, for $\omega\in[a,\infty)$,

$$C_a\omega = \frac{C_1\omega}{a}+\omega \geq C_1+\omega C_2\left(\frac{\log n}{nh^d}\right)^{\frac{1}{2}}.$$

Note that there is some n_0 so that this holds uniformly over all α for $n \geq n_0$.

The integral on the right-hand side of (5.1.16) is handled by the representation for the incomplete rho integral, i.e.

$$\int_a^\infty \omega^j \exp(-\omega)\,\mathrm{d}\omega = j!\exp(-a)\sum_{k=0}^{j}\frac{a^k}{k!}\,, \quad j \in \mathbb{N},\ a > 0\,.$$

Deduce that for $j = \lceil \rho - 1 \rceil$, $a \geq 1$, n so large that $\log n \geq C_a$, we have

$$\begin{aligned}
\int_a^\infty \omega^{\rho-1}\exp\Big(-\frac{\omega\log n}{C_a}\Big)\,\mathrm{d}\omega &\leq \int_a^\infty \omega^j \exp\Big(-\frac{\omega\log n}{C_a}\Big)\,\mathrm{d}\omega \\
&\leq \Big(\frac{C_a}{\log n}\Big)^{j+1}\int_{\frac{a\log n}{C_a}}^\infty \omega^j\exp(-\omega)\,\mathrm{d}\omega \\
&= \Big(\frac{C_a}{\log n}\Big)^{j+1} j!\exp\Big(-\frac{a\log n}{C_a}\Big)\sum_{k=0}^{j}\frac{\Big(\frac{a\log n}{C_a}\Big)^k}{k!} \\
&\leq \Big(\frac{C_a}{\log n}\Big)^{j+1}(j+1)!a^j\frac{(\log n)^j}{C_a^j}n^{-\frac{a}{C_a}} \\
&\leq \frac{C_a a^j (j+1)!}{\log n}n^{-\frac{a}{2\max\{C_1,1\}}}\,,
\end{aligned}$$

where the last inequality holds because $\frac{a}{C_a} = \frac{a^2}{C_1+a}$ and $C_1 + a \leq 2\max\{C_1,1\}a$. By choosing $\delta_n = n^{-5/2}$, $a = 5(d+m)\max\{C_1,1\}$ and using

$$\frac{\delta_n^\rho}{h^{(d+1)\rho}} = \frac{n^{-5\rho/2}}{h^{(d+1)\rho}} \lesssim \Big(\frac{\log n}{nh^d}\Big)^{\frac{\rho}{2}}\cdot\frac{1}{n^{2\rho}h^{(d+2)\rho/2}} \lesssim \Big(\frac{\log n}{nh^d}\Big)^{\frac{\rho}{2}}\cdot\frac{1}{n^{2\rho}h^{2d\rho}} \lesssim \Big(\frac{\log n}{nh^d}\Big)^{\frac{\rho}{2}}\,, \tag{5.1.17}$$

we deduce

$$\begin{aligned}
&\mathbb{E}_\gamma\Big[\sup_{x\in J,\theta\in\Theta}|T_n(\theta,x;h)|^\rho\Big] \\
\leq& O\Big(\frac{\delta_n^\rho}{h^{(d+1)\rho}}\Big) + a^\rho\Big(\frac{\log n}{nh^d}\Big)^{\frac{\rho}{2}} + \frac{C_a a^j C_{\Theta,J}(j+1)!}{\log n}\Big(\frac{\log n}{nh^d}\Big)^{\frac{\rho}{2}}\delta_n^{-d-m}n^{-\frac{a}{2\max\{C_1,1\}}} \\
=& O\Big(\Big(\frac{\log n}{nh^d}\Big)^{\frac{\rho}{2}}\Big)\,,
\end{aligned}$$

concluding the proof.

□

Proof of Theorem 3.2.7. We will drop dependence of the bandwidth parameters on α and n for convenience but only give a hint where it comes into play. Note that throughout the proof, we will only use the notation $a_n \lesssim b_n$ if there is a constant $C > 0$ and an

$n_0 \in \mathbb{N}$ so that for all $n \geq n_0$ we have $a_n \leq C b_n$ and the constant depends only on $\|\tau\|_\infty$, L_τ, $\|K\|_\infty$, L_K, ρ, J, Θ, Γ, A or is universal. Also note that once again all calculations below hold independently of $\gamma \in \Gamma$.

Let us decompose the centred process $M_n - \mathbb{E}_\gamma[M_n]$ into a canonical U-statistic and a linear process as described in (3.2.6)-(3.2.10). For any fixed $x \in J, \theta \in \Theta$ write

$$\begin{aligned} M_n(\theta,x;h) - \mathbb{E}_\gamma[M_n(\theta,x;h)] =& \frac{1}{n(n-1)} \sum_{\substack{j,k=1 \\ j\neq k}}^{n} U_n\big(Z_j, Z_k, \theta, x; h\big) \\ &+ \frac{2}{n} \sum_{j=1}^{n} \Big[u_n^*(Z_j,\theta,x;h) - \mathbb{E}_\gamma\big[u_n(Z_1,Z_2,\theta,x;h)\big]\Big] \\ =:& \, T_n^1(\theta,x;h) + T_n^2(\theta,x;h) \, , \end{aligned}$$

where for $z = (z_1, z_2^T)^T, y = (y_1, y_2^T)^T \in \mathbb{R} \times J$,

$$\begin{aligned} U_n(z,y,\theta,x;h) :=& \, u_n(z,y,\theta,x;h) - u_n^*(z,\theta,x;h) - u_n^*(y,\theta,x;h) \\ &+ \mathbb{E}_\gamma[u_n(Z_1,Z_2,\theta,x;h)] \, , \\ u_n(z,y,\theta,x;h) :=& \, \tau(z_1,y_1,\theta) K_h(z_2 - x) K_h(y_2 - x) \, , \\ u_n^*(z,\theta,x;h) :=& \, \mathbb{E}_\gamma[u_n(Z_1,z,\theta,x;h)] = \mathbb{E}_\gamma\big[\tau(z_1,Y_1,\theta) K_h(X_1 - x)\big] \cdot K_h(z_2 - x) \, . \end{aligned}$$

Since $s \mapsto s^\rho$ is convex, we have

$$\begin{aligned} \big|M_n(\theta,x;h) - \mathbb{E}_\gamma[M_n(\theta,x;h)]\big|^\rho &\leq \big|T_n^1(\theta,x;h) + T_n^2(\theta,x;h)\big|^\rho \\ &\leq 2^\rho \left| \frac{1}{2} T_n^1(\theta,x;h) + \frac{1}{2} T_n^2(\theta,x;h) \right|^\rho \\ &= 2^\rho \left| \Big(\frac{1}{2} T_n^1(\theta,x;h) + \frac{1}{2} T_n^2(\theta,x;h) \Big)^\rho \right| \\ &\leq 2^\rho \left| \Big(\frac{1}{2} |T_n^1(\theta,x;h)| + \frac{1}{2} |T_n^2(\theta,x;h)| \Big)^\rho \right| \\ &\leq 2^{\rho-1} \Big(|T_n^1(\theta,x;h)|^\rho + |T_n^2(\theta,x;h)|^\rho \Big) \, . \end{aligned} \tag{5.1.18}$$

The linear error process T_n^2 in (5.1.18) can be directly treated by Theorem 3.2.5. Note that (3.2.13) and (3.2.14) are given by Lipschitz continuity of τ. Moreover, (3.2.12) is given by boundedness of τ, and (3.2.15) is given by Bernstein's inequality, cf. Lemma 3.2.2.

Let us deal with the term T_n^1 in (5.1.18). By using Lemma 5.1.1, for a sequence $\delta_n \to 0$ specified later, there are nets $\Theta_n \times J_n \subset \Theta \times J$ so that

$$\sup_{x \in J} \inf_{y \in J_n} \|x - y\| < \delta_n \, , \qquad \sup_{\vartheta \in \Theta} \inf_{\theta \in \Theta_n} \|\vartheta - \theta\| < \delta_n \, , \qquad \#(\Theta_n \times J_n) \leq C_{\Theta,J} \delta_n^{-d-m} \, ,$$

where $C_{\Theta,J}$ is independent of n. We further achieve the representation

$$\begin{aligned}
&\sup_{x\in J,\theta\in\Theta} |T_n^1(\theta,x;h)|^\rho \\
&\leq 2^{\rho-1}\Bigg(\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\leq\delta_n}} |T_n^1(\theta,x;h)-T_n^1(\vartheta,y;h)|\Bigg)^\rho + 2^{\rho-1}\Big(\sup_{x\in J_n,\vartheta\in\Theta_n} |T_n^1(\vartheta,x;h)|\Big)^\rho .
\end{aligned} \tag{5.1.19}$$

The first summand in (5.1.19) is treated by applying Jensen's inequality for sums, yielding

$$\begin{aligned}
&\mathbb{E}_\gamma\Bigg[\Bigg(\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\leq\delta_n}} |T_n^1(\theta,x;h)-T_n^1(\vartheta,y;h)|\Bigg)^\rho\Bigg] \\
&\leq 3^{\rho-1}\mathbb{E}_\gamma\Bigg[\Bigg(\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\leq\delta_n}} \Bigg|\frac{1}{n(n-1)}\sum_{\substack{j,k=1\\ j\neq k}}^n u_n(Z_j,Z_k,\theta,x;h)-u_n(Z_j,Z_k,\vartheta,y;h)\Bigg|\Bigg)^\rho\Bigg] \\
&\quad + 3^{\rho-1}\mathbb{E}_\gamma\Bigg[\Bigg(\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\leq\delta_n}} \Bigg|\frac{2}{n}\sum_{j=1}^n u_n^*(Z_j,\vartheta,x;h)-u_n^*(Z_j,\theta,y;h)\Bigg|\Bigg)^\rho\Bigg] \\
&\quad + 3^{\rho-1}\mathbb{E}_\gamma\Bigg[\Bigg(\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\leq\delta_n}} \Big|\mathbb{E}_\gamma\big[u_n(Z_1,Z_2,\theta,x;h)-u_n(Z_1,Z_2,\vartheta,y;h)\big]\Big|\Bigg)^\rho\Bigg] \\
&\leq 3^{\rho-1}\mathbb{E}_\gamma\Bigg[\Bigg(\frac{1}{n(n-1)}\sum_{\substack{j,k=1\\ j\neq k}}^n \sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\leq\delta_n}} \big|u_n(Z_j,Z_k,\theta,x;h)-u_n(Z_j,Z_k,\vartheta,y;h)\big|\Bigg)^\rho\Bigg]
\end{aligned} \tag{5.1.20}$$

$$+ 3^{\rho-1}\mathbb{E}_\gamma\Bigg[\Bigg(\frac{2}{n}\sum_{j=1}^n \sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\leq\delta_n}} \big|u_n^*(Z_j,\theta,x;h)-u_n^*(Z_j,\vartheta,y;h)\big|\Bigg)^\rho\Bigg] \tag{5.1.21}$$

$$+ 3^{\rho-1}\Bigg(\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\leq\delta_n}} \Big|\mathbb{E}_\gamma\big[u_n(Z_1,Z_2,\theta,x;h)-u_n(Z_1,Z_2,\vartheta,y;h)\big]\Big|\Bigg)^\rho . \tag{5.1.22}$$

Let us bound the occurring summands directly. The summands of the term (5.1.20) are bounded by

$$\begin{aligned}
&\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\leq\delta_n}} \Big|\tau(Y_j,Y_k,\vartheta)\cdot\Big(K_h(X_j-x)K_h(X_k-x)-K_h(X_j-y)K_h(X_k-y)\Big)\Big| \\
+ &\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\leq\delta_n}} \Big|\Big[\tau(Y_j,Y_k,\vartheta)-\tau(Y_j,Y_k,\theta)\Big]\cdot K_h(X_j-y)K_h(X_k-y)\Big| ,
\end{aligned}$$

of which the first factor in the first summand is bounded by $\|\tau\|_\infty$. The first kernel terms are handled by the equality $ab-cd=ab-ac+ac-cd$, $\|K_h\|_\infty=\|K\|_\infty\frac{1}{h^d}$ and the fact

that K is Lipschitz continuous, i.e.

$$\sup_{\substack{x,y\in J\\ \|x-y\|\le\delta_n}} |K_h(X_j-x)-K_h(X_j-y)| \le L_K\frac{1}{h^d}\cdot\frac{\delta_n}{h}\ ,$$

yielding

$$\begin{aligned}&\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\le\delta_n}} \big|\tau(Y_j,Y_k,\vartheta)\cdot\big(K_h(X_j-x)K_h(X_k-x)-K_h(X_j-y)K_h(X_k-y)\big)\big|\\ &\le 2\|\tau\|_\infty L_K\|K\|_\infty\frac{\delta_n}{h^{2d+1}}\ .\end{aligned}$$

By using the Lipschitz continuity of τ in its third argument, we derive for the second summand that

$$\begin{aligned}&\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\le\delta_n}} \Big|\Big[\tau(Y_j,Y_k,\vartheta)-\tau(Y_j,Y_k,\theta)\Big]\cdot K_h(X_j-y)K_h(X_k-y)\Big|\\ &\le L_\tau\delta_n\sup_{\substack{x,y\in J\\ \|x-y\|\le\delta_n}}\big|K_h(X_j-y)K_h(X_k-y)\big| \le L_\tau\|K\|_\infty^2\frac{\delta_n}{h^{2d}}\ .\end{aligned}$$

Hence,

$$(5.1.20)\le 3^{\rho-1}\bigg(2\|\tau\|_\infty L_K\|K\|_\infty\frac{\delta_n}{h^{2d+1}}+L_\tau\|K\|_\infty^2\frac{\delta_n}{h^{2d}}\bigg)^\rho\lesssim\frac{\delta_n^\rho}{h^{(2d+1)\rho}}\ .$$

Using similar arguments, we observe that the summands of (5.1.21) are bounded by

$$\begin{aligned}&\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\le\delta_n}}\big|u_n^*(Z_j,\vartheta,x;h)-u_n^*(Z_j,\theta,x;h)\big|\\ &+\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\le\delta_n}}\big|u_n^*(Z_j,\theta,x;h)-u_n^*(Z_j,\theta,y;h)\big|\\ \le&\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\le\delta_n}}\Big|\mathbb{E}_\gamma^Z\big[\big(\tau(Y,Y_1,\vartheta)-\tau(Y,Y_1,\theta)\big)K_h(X-x)\big]K_h(X_1-x)\Big|\\ &+\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\le\delta_n}}\Big|\mathbb{E}_\gamma^Z\Big[\tau(Y,Y_1,\theta)\Big(K_h(X-x)K_h(X_1-x)-K_h(X-y)K_h(X_1-y)\Big)\Big]\Big|\\ \le&\frac{\|K\|_\infty L_\tau\|\ell_\gamma\|_\infty\delta_n}{h^d}+\frac{\|\tau\|_\infty\|\ell_\gamma\|_\infty L_K\delta_n}{h^{2d}}+\frac{\|\tau\|_\infty\|K\|_\infty^2 L_K\delta_n}{h^{2d+1}}\ ,\end{aligned}$$

where we used the notation

$$\mathbb{E}_\gamma^Z\big[f(Y,Y_1,X,X_1)\big]=\iint f(y,Y_1,x,X_1)\,f_\gamma(y|x)\ell_\gamma(x)\,\mathrm{d}y\,\mathrm{d}x\ .$$

Thus, we conclude (5.1.21) $\lesssim \frac{\delta_n^\rho}{h^{(2d+1)\rho}}$.

The term (5.1.22) is bounded by (5.1.20) according to Jensen's inequality, i.e.

$$
\begin{aligned}
&(5.1.22)\\
&= 3^{\rho-1}\left(\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\le\delta_n}} \frac{1}{n(n-1)} \sum_{\substack{j,k=1\\ j\neq k}}^{n} \left|\mathbb{E}_\gamma\left[u_n(Z_j,Z_k,\vartheta,x;h)-u_n(Z_j,Z_k,\theta,y;h)\right]\right| \right)^\rho\\
&\le 3^{\rho-1}\left(\mathbb{E}_\gamma\left[\frac{1}{n(n-1)} \sum_{\substack{j,k=1\\ j\neq k}}^{n} \sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\le\delta_n}} \left|u_n(Z_j,Z_k,\vartheta,x;h)-u_n(Z_j,Z_k,\theta,y;h)\right| \right] \right)^\rho\\
&\le 3^{\rho-1}\mathbb{E}_\gamma\left[\left(\frac{1}{n(n-1)} \sum_{\substack{j,k=1\\ j\neq k}}^{n} \sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\le\delta_n}} \left|u_n(Z_j,Z_k,\vartheta,x;h)-u_n(Z_j,Z_k,\theta,y;h)\right| \right)^\rho \right]\\
&= (5.1.20)\ .
\end{aligned}
$$

This finally gives a bound on the discretization error in (5.1.19), i.e.

$$
\mathbb{E}_\gamma\left[2^{\rho-1}\left(\sup_{x\in J,\vartheta\in\Theta} |T_n^1(\vartheta,x;h)| - \sup_{y\in J_n,\theta\in\Theta_n} |T_n^1(\theta,y;h)| \right)^\rho \right] \lesssim \frac{\delta_n^\rho}{h^{(2d+1)\rho}}\ .
$$

Now, let us focus on the discrete error $\sup_{x\in J_n,\theta\in\Theta_n} |T_n^1(\theta,x;h)|^\rho$. First we notice that $T_n^1(\theta,x;h)$ is a canonical U-Statistic in $Z_1,\dots,Z_n$ because U_n is symmetric in its first two arguments. In order to bound the error

$$
\sup_{x\in J_n,\theta\in\Theta_n} |T_n^1(\theta,x;h)|\ ,
$$

we will need to examine the tail behaviour of $|T_n^1(\theta,x;h)|$, which can be done by means of the Bernstein-type inequality for canonical U-statistics introduced by Giné et al. (2000), given in Lemma 3.2.4. In order to derive the terms A, B, C described in (3.2.2), we first observe that when taking the expectation of a term involving a random K_h term, we lose one factor $\frac{1}{h^d}$ by integration, e.g.

$$
\begin{aligned}
\mathbb{E}_\gamma\left[\left|K_h(X_1-x)\right|\right] &\le \|\ell_\gamma\|_\infty\ ,\\
\mathbb{E}_\gamma\left[K_h^2(X_1-x)\right] &\le \frac{1}{h^d}\|\ell_\gamma\|_\infty \int K^2 \lesssim \frac{1}{h^d}\ .
\end{aligned}
$$

This yields

$$
\begin{aligned}
A &= \|U_n\|_\infty \lesssim \|u_n\|_\infty \le \|\tau\|_\infty \frac{\|K\|_\infty^2}{h^{2d}} \lesssim \frac{1}{h^{2d}}\ ,\\
B^2 &= n\left\|\mathbb{E}_\gamma\left[U_n^2(Z_1,\cdot,\theta,x;h)\right]\right\|_\infty \lesssim n\left\|\mathbb{E}_\gamma\left[u_n^2(Z_1,\cdot,\theta,x;h)\right]\right\|_\infty \lesssim \frac{n}{h^{3d}}\ .
\end{aligned}
$$

The same arguments apply to C^2, giving

$$
C^2 = n(n-1)\,\mathbb{E}_\gamma[U_n^2(Z_1,Z_2,\theta,x;h)]
$$

$$
\begin{aligned}
&= n(n-1)\Big(\mathbb{E}_\gamma[u_n^2(Z_1,Z_2,\theta,x;h)] + 4\,\mathbb{E}_\gamma[u_n(Z_1,Z_2,\theta,x;h)u_n^*(Z_1,\theta,x;h)] \\
&\quad + 4\big(\mathbb{E}_\gamma[u_n(Z_1,Z_2,\theta,x;h)]\big)^2 + 4\,\mathbb{E}_\gamma[{u_n^*}^2(Z_1,\theta,x;h)] \\
&\quad + 4\,\mathbb{E}_\gamma[u_n^*(Z_1,\theta,x;h)u_n^*(Z_2,\theta,x;h)]\Big) \\
&\lesssim n(n-1)\mathbb{E}_\gamma\big[\big|u_n(Z_1,Z_2,\theta,x;h)\big|\big]\,\|u_n\|_\infty \lesssim \frac{n^2}{h^{2d}}\,.
\end{aligned}
$$

Now, the Bernstein-type inequality given in Lemma 3.2.4 and the monotonicity of exp give for any fixed θ, x, h, γ and any $\omega > 0$,

$$
\begin{aligned}
&\mathbb{P}_\gamma\big(|T_n^1(\theta,x;h)| > \omega\big) \\
&= \mathbb{P}_\gamma\Big(\Big|\sum_{\substack{j,k=1\\ j\neq k}}^{n} U_n(Z_j,Z_k,\theta,x;h)\Big| > n(n-1)\omega\Big) \\
&\le T\exp\Big(-T^{-1}\min\Big\{\frac{n(n-1)\omega}{C},\ \Big(\frac{n(n-1)\omega}{B}\Big)^{\frac{2}{3}},\ \Big(\frac{n(n-1)\omega}{A}\Big)^{\frac{1}{2}}\Big\}\Big) \\
&\lesssim T\exp\Big(-\frac{1}{T}\min\Big\{\frac{n(n-1)h^d\omega}{n}, \Big(\frac{n(n-1)h^{\frac{3d}{2}}\omega}{\sqrt{n}}\Big)^{\frac{2}{3}}, \big(n(n-1)h^{2d}\omega\big)^{\frac{1}{2}}\Big\}\Big) \\
&\lesssim T\exp\Big(-\frac{1}{T}nh^d\min\big\{\omega,\omega^{\frac{2}{3}},\omega^{\frac{1}{2}}\big\}\Big) \\
&\le T\exp\Big(-\frac{1}{T}nh^d\omega\Big)\mathbb{1}_{\omega\in[0,1)} + T\exp\Big(-\frac{1}{T}nh^d\omega^{\frac{1}{2}}\Big)\mathbb{1}_{\omega\in[1,\infty)}\,,
\end{aligned}
$$

where $T > 0$ is a universal constant. We will apply the estimate

$$
\mathbb{E}[|W|^\rho] \le \int_0^\infty \rho\omega^{\rho-1}\mathbb{P}(|W| > \omega)\,\mathrm{d}\omega \le a^\rho + \int_a^\infty \rho\omega^{\rho-1}\mathbb{P}(|W| > \omega)\,\mathrm{d}\omega\,, \quad a > 0
$$

to $W = \sup_{x\in J_n,\theta\in\Theta_n}|T_n^1(\theta,x;h)|\big(\frac{nh^d}{\log n}\big)^{\frac{1}{2}}$ in order to obtain an estimate of the remainder by

$$
\begin{aligned}
&\mathbb{E}_\gamma\Big[\sup_{x\in J_n,\theta\in\Theta_n}|T_n^1(\theta,x;h)|^\rho\Big] \\
&\le \Big(\frac{\log n}{nh^d}\Big)^{\frac{\rho}{2}}\Big[a^\rho + \int_a^\infty \rho\omega^{\rho-1}\mathbb{P}_\gamma\Big(\sup_{x\in J_n,\theta\in\Theta_n}|T_n^1(\theta,x;h)| > \Big(\frac{\log n}{nh^d}\Big)^{\frac{1}{2}}\omega\Big)\,\mathrm{d}\omega\Big] \\
&\le \Big(\frac{\log n}{nh^d}\Big)^{\frac{\rho}{2}}\Big[a^\rho + \int_a^\infty \rho\omega^{\rho-1}\sum_{x\in J_n,\theta\in\Theta_n}\mathbb{P}_\gamma\Big(|T_n^1(\theta,x;h)| > \Big(\frac{\log n}{nh^d}\Big)^{\frac{1}{2}}\omega\Big)\,\mathrm{d}\omega\Big] \\
&\le \Big(\frac{\log n}{nh^d}\Big)^{\frac{\rho}{2}}\Big[a^\rho + C_{\Theta,J}\delta_n^{-d-m}\int_a^\infty \rho\omega^{\rho-1}T\exp\Big(-\frac{1}{T}nh^d\Big(\frac{\log n}{nh^d}\Big)^{\frac{1}{2}}\omega\Big)\mathbb{1}_{\omega\in\left[0,\left(\frac{nh^d}{\log n}\right)^{1/2}\right)}\,\mathrm{d}\omega \\
&\quad + C_{\Theta,J}\delta_n^{-d-m}\int_a^\infty \rho\omega^{\rho-1}T\exp\Big(-\frac{1}{T}nh^d\Big(\frac{\log n}{nh^d}\Big)^{\frac{1}{4}}\omega^{\frac{1}{2}}\Big)\mathbb{1}_{\omega\in\left[\left(\frac{nh^d}{\log n}\right)^{1/2},\infty\right)}\,\mathrm{d}\omega\Big]
\end{aligned}
$$

$$\lesssim \Big(\frac{\log n}{nh^d}\Big)^{\frac{\rho}{2}} \Bigg[a^\rho + C_{\Theta,J}\delta_n^{-d-m} \int_a^{\max\left\{\left(\frac{nh^d}{\log n}\right)^{1/2},a\right\}} \rho\omega^{\rho-1} T \exp\Big(-\frac{1}{T}\log(n)\omega\Big)\,\mathrm{d}\omega$$
$$+ C_{\Theta,J}\delta_n^{-d-m} \int_{\max\left\{\left(\frac{nh^d}{\log n}\right)^{1/2},a\right\}}^{\infty} \rho\omega^{\rho-1} T \exp\Big(-\frac{1}{T}\log(n)\omega^{\frac{1}{2}}\Big)\,\mathrm{d}\omega \Bigg], \tag{5.1.23}$$

where we used

$$nh^d\Big(\frac{\log n}{nh^d}\Big)^{\frac{1}{2}} = \Big(\frac{nh^d}{\log n}\Big)^{\frac{1}{2}}\log n\,, \quad nh^d\Big(\frac{\log n}{nh^d}\Big)^{\frac{1}{4}} = \Big(\frac{nh^d}{\log n}\Big)^{\frac{3}{4}}\log n$$

and the fact that $\sup_\alpha \frac{\log n}{nh_n(\alpha)^d} \to 0$ implies $\inf_\alpha nh_n(\alpha)^d \geq \log n$ for large enough n.

The integrals on the right-hand side of (5.1.23) are handled by the representation for the incomplete rho integral, i.e. for $l \in \mathbb{N}, a > 0$

$$\int_a^\infty \omega^l \exp(-\omega)\,\mathrm{d}\omega = l!\exp(-a)\sum_{k=0}^{l}\frac{a^k}{k!}\,.$$

For a choice of $a \geq 1$ that will be specified later on and $l := \lceil \rho - 1 \rceil$, this and a substitution yield

$$\begin{aligned}
&\int_a^{\max\left\{\left(\frac{nh^d}{\log n}\right)^{1/2},a\right\}} \rho\omega^{\rho-1} T \exp\big(-T^{-1}\log(n)\omega\big)\,\mathrm{d}\omega \\
&\leq \int_a^\infty \rho\omega^l T \exp\big(-T^{-1}\log(n)\omega\big)\,\mathrm{d}\omega \\
&= \rho\frac{T^{l+2}}{\log(n)^{l+1}}\int_{T^{-1}\log(n)a}^\infty \omega^l\exp\big(-\omega\big)\,\mathrm{d}\omega \\
&= \rho\frac{T^{l+2}l!}{\log(n)^{l+1}}\sum_{k=0}^{l}\left(\frac{\big(T^{-1}\log(n)a\big)^k}{k!}\right)\cdot\exp(-T^{-1}\log(n)a) \\
&\lesssim \frac{n^{-T^{-1}a}}{\log n}\,.
\end{aligned}$$

By using the transformation $\omega \mapsto \frac{\omega^2T^2}{\log(n)^2}$, we get

$$\begin{aligned}
&\int_{\max\left\{\left(\frac{nh^d}{\log n}\right)^{\frac{1}{2}},a\right\}}^\infty \rho\omega^{\rho-1} T \exp\big(-T^{-1}\log(n)\omega^{\frac{1}{2}}\big)\,\mathrm{d}\omega \\
&\leq \int_{\max\left\{\left(\frac{nh^d}{\log n}\right)^{\frac{1}{2}},a\right\}}^\infty \rho\omega^l T \exp\big(-T^{-1}\log(n)\omega^{\frac{1}{2}}\big)\,\mathrm{d}\omega \\
&= \int_{\max\left\{\left(\frac{nh^d}{\log n}\right)^{\frac{1}{4}},a^{\frac{1}{2}}\right\}_{T^{-1}\log n}}^\infty \frac{2\rho T^{2l+3}}{(\log n)^{2l+2}}\omega^{2l+1}\exp\big(-\omega\big)\,\mathrm{d}\omega
\end{aligned}$$

$$
\begin{aligned}
&= \frac{2\rho T^{2l+3}}{(\log n)^{2l+2}}(2l+1)! \sum_{k=0}^{2l+1} \frac{\Big(\max\big\{\big(\frac{nh^d}{\log n}\big)^{\frac{1}{4}}, a^{\frac{1}{2}}\big\}T^{-1}\log n\Big)^k}{k!} \\
&\quad \cdot \exp\bigg(-\max\bigg\{\bigg(\frac{nh^d}{\log n}\bigg)^{\frac{1}{4}}, a^{\frac{1}{2}}\bigg\}T^{-1}\log n\bigg) \\
&\lesssim \frac{1}{\log n} b_n^{2l+1} n^{-T^{-1}b_n} \\
&\lesssim n^{-c}\,,
\end{aligned}
$$

where $b_n = \max\big\{\big(\frac{nh^d}{\log n}\big)^{\frac{1}{4}}, a^{\frac{1}{2}}\big\}$ converges to ∞ so that the last bound holds for any $c > 0$.

By choosing $\delta_n = n^{\frac{-T^{-1}a}{d+m}}$, $a \geq 1$ so large that independently of α,

$$
\frac{\delta_n^\rho}{h_n(\alpha)^{(2d+1)\rho}} = n^{-\frac{T^{-1}a\rho}{d+m}} h_n(\alpha)^{-(2d+1)\rho} \lesssim \bigg(\frac{\log n}{nh_n(\alpha)}\bigg)^{\frac{\rho}{2}}\,,
$$

$c > T^{-1}a$ and using that $\sup_\alpha \frac{\log n}{nh_n(\alpha)^d} \to 0$ implies $\sup_\alpha \frac{1}{h_n(\alpha)^d} \lesssim \frac{n}{\log n}$, we get

$$
\begin{aligned}
\mathbb{E}_\gamma\Big[\sup_{x\in J,\theta\in\Theta} |T_n^1(\theta,x;h)|^\rho\Big] &\lesssim \frac{\delta_n^\rho}{h^{(2d+1)\rho}} + \Big(\frac{\log n}{nh^d}\Big)^{\frac{\rho}{2}} + \Big(\frac{\log n}{nh^d}\Big)^{\frac{1}{2}} \delta_n^{-d-m}\bigg(\frac{n^{-T^{-1}a}}{\log n} + n^{-c}\bigg) \\
&\lesssim \Big(\frac{\log n}{nh^d}\Big)^{\frac{\rho}{2}} + \Big(\frac{\log n}{nh^d}\Big)^{\frac{\rho}{2}} \cdot \frac{1}{\log n} \\
&\lesssim \Big(\frac{\log n}{nh^d}\Big)^{\frac{\rho}{2}}\,,
\end{aligned}
$$

concluding the considerations of $\sup_{x\in J,\theta\in\Theta} |T_n^1(\theta,x;h)|^\rho$.

□

Proofs for Section 3.2.3

Proof of Lemma 3.2.9. We have to show that for any $C_{\mathrm{Lep}} > C_-$, we have

$$
\limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \sup_{0\leq l\leq j\leq k_n(\alpha)} n^{u(C_{\mathrm{Lep}})} \tilde{p}_{lj} < \infty
$$

with

$$
\begin{aligned}
\tilde{p}_{lj} &= \mathbb{P}_\gamma\Big(\sup_{x\in J,\theta\in\Theta} \|T_n(\theta,x;h_j)\| > \psi(C_{\mathrm{Lep}})r_l\Big) \\
&= \mathbb{P}_\gamma\Big(\sup_{x\in J,\theta\in\Theta} \|T_n(\theta,x;h_j)\| > \psi(C_{\mathrm{Lep}})h_l^{\alpha_l}\Big)\,, \qquad (5.1.24)
\end{aligned}
$$

where for abbreviation, we define the function $\psi : \big[c_1^{-1}(c_2+1), \infty\big) \to [1,\infty)$, $\psi(C_{\mathrm{Lep}}) := c_1 C_{\mathrm{Lep}} - c_2$ and note that ψ grows linearly in C_{Lep}.

We will deal with the term (5.1.24) by using a discretization approach similar to the one in the proof of Theorem 3.2.5. According to Lemma 5.1.1 with $\rho = 1$, for any sequence $\delta_n \to 0$, there are nets $\Theta_n \times J_n \subset \Theta \times J$ with

$$\sup_{x\in J}\inf_{y\in J_n}\|x-y\| < \delta_n\ , \qquad \sup_{\vartheta\in\Theta}\inf_{\theta\in\Theta_n}\|\vartheta-\theta\| < \delta_n\ , \qquad \#(\Theta_n\times J_n) \le C_{\Theta,J}\delta_n^{-d-m}\ ,$$

where $C_{\Theta,J}$ is independent of n. Furthermore, for any $h > 0$,

$$\begin{aligned}&\sup_{x\in J,\theta\in\Theta}\|T_n(\theta,x;h)\|\\ \le&\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\le\delta_n}}\|T_n(\theta,x;h)-T_n(\vartheta,y;h)\| + \sup_{x\in J_n,\theta\in\Theta_n}\|T_n(\theta,x;h)\|\ .\end{aligned} \tag{5.1.25}$$

The sequence δ_n will be specified later on in order to balance convergence rates. Now we may handle (5.1.24) by

$$\begin{aligned}&\mathbb{P}_\gamma\Big(\sup_{x\in J,\theta\in\Theta}\|T_n(\theta,x;h_j)\| > \psi(C_{\mathrm{Lep}})h_l^{\alpha_l}\Big)\\ \le&\,\mathbb{P}_\gamma\Big(\sup_{x\in J_n,\theta\in\Theta_n}\|T_n(\theta,x;h_j)\| > \psi(C_{\mathrm{Lep}})h_l^{\alpha_l}/2\Big) &(5.1.26)\\ &+\mathbb{P}_\gamma\left(\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\le\delta_n}}\|T_n(\theta,x;h_j)-T_n(\vartheta,y;h_j)\| > \psi(C_{\mathrm{Lep}})h_l^{\alpha_l}/2\right) &(5.1.27)\end{aligned}$$

according to (5.1.25). The term (5.1.26) can by handled by (3.2.15), i.e.

$$\begin{aligned}&\mathbb{P}_\gamma\Big(\sup_{x\in J_n,\theta\in\Theta_n}\|T_n(\theta,x;h_j)\| > \psi(C_{\mathrm{Lep}})h_l^{\alpha_l}/2\Big)\\ \le&\sum_{x\in J_n,\theta\in\Theta_n}\mathbb{P}_\gamma\big(\|T_n(\theta,x;h_j)\| > \psi(C_{\mathrm{Lep}})h_l^{\alpha_l}/2\big)\\ \le&C_{\Theta,J}\delta_n^{-d-m}2\exp\left(-\frac{1}{4}\frac{\psi^2(C_{\mathrm{Lep}})h_l^{2\alpha_l}nh_j^d}{C_1+C_2\psi(C_{\mathrm{Lep}})h_l^{\alpha_l}}\right).\end{aligned} \tag{5.1.28}$$

Now, by using $C_1 + C_2\psi(C_{\mathrm{Lep}}) \le 2\max\{C_1,C_2\}\psi(C_{\mathrm{Lep}})$, $\sup_{\alpha\in[a,b],n\in\mathbb{N}} h(\alpha) \le 1$, $h_l = \big(\frac{\log n}{n}\big)^{\frac{1}{2\alpha_l+d}}$ and the fact that $h_j \ge h_l$, we deduce

$$\begin{aligned}(5.1.28)\le&C_{\Theta,J}\delta_n^{-d-m}2\exp\left(-\frac{\psi(C_{\mathrm{Lep}})h_l^{2\alpha_l}nh_j^d}{8\max\{C_1,C_2\}}\right)\\ \le&C_{\Theta,J}\delta_n^{-d-m}2\exp\left(-\frac{\psi(C_{\mathrm{Lep}})nh_l^{2\alpha_l+d}}{8\max\{C_1,C_2\}}\right)\\ =&2C_{\Theta,J}\delta_n^{-d-m}\ n^{-\frac{\psi(C_{\mathrm{Lep}})}{8\max\{C_1,C_2\}}}\ .\end{aligned} \tag{5.1.29}$$

The term (5.1.27) is handled by arguments similar to the ones found in the proof of Theorem 3.2.5. Using Markov's inequality, $h_j \ge h_l$ and the arguments used to treat

(5.1.9)-(5.1.12) for $\rho = 1$ that showed that the expectation in the following display is $O(\delta_n h_j^{-d-1})$, we deduce that there is a constant $\widetilde{C}$ so that

$$\limsup_{n\to\infty} \sup_{*} \delta_n^{-1} h_l^{\alpha_l+d+1} \mathbb{P}_\gamma \Big(\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta \\ \|x-y\|,\|\vartheta-\theta\|\le\delta_n}} \big\| T_n(\theta,x;h_j) - T_n(\vartheta,y;h_j) \big\| > \psi(C_{\mathrm{Lep}}) h_l^{\alpha_l}/2 \Big)$$

$$\le \limsup_{n\to\infty} \sup_{*} \frac{2\delta_n^{-1} h_j^{d+1}}{\psi(C_{\mathrm{Lep}})} \mathbb{E}_\gamma \Big[\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta \\ \|x-y\|,\|\vartheta-\theta\|\le\delta_n}} \big\| T_n(\theta,x;h_j) - T_n(\vartheta,y;h_j) \big\| \Big] < \widetilde{C} < \infty \,, \quad (5.1.30)$$

where the suprema are taken over $\alpha \in [a,b]$, $\gamma \in \Gamma(\alpha)$, $0 \le l \le j \le k_n(\alpha)$. Combining (5.1.29) and (5.1.30) yields

$$\limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \sup_{0\le l\le j\le k_n(\alpha)} n^{u(C_{\mathrm{Lep}})} \tilde{p}_{lj}$$

$$\le \limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} 2C_{\Theta,J}\, n^{u(C_{\mathrm{Lep}})} \delta_n^{-d-m} n^{-\frac{\psi(C_{\mathrm{Lep}})}{8\max\{C_1,C_2\}}} \quad (5.1.31)$$

$$+ \limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \sup_{0\le l\le j\le k_n(\alpha)} \widetilde{C}\, n^{u(C_{\mathrm{Lep}})} \delta_n h_l^{-\alpha_l-d-1} \,, \quad (5.1.32)$$

which will be finite by choosing

$$\delta_n = \delta_n(C_{\mathrm{Lep}}) = n^{-\frac{\psi(C_{\mathrm{Lep}})}{16\max\{C_1,C_2\}(d+m)}} = n^{-\frac{c_1 C_{\mathrm{Lep}} - c_2}{16\max\{C_1,C_2\}(d+m)}} \,.$$

In order to treat (5.1.31), we see that

$$\begin{aligned}
&\log_n \left(n^{u(C_{\mathrm{Lep}})} \delta_n^{-d-m} n^{-\frac{\psi(C_{\mathrm{Lep}})}{8\max\{C_1,C_2\}}} \right) \\
&= \frac{c_1 C_{\mathrm{Lep}} - c_2}{32\max\{C_1,C_2\}(d+m)} + \frac{c_1 C_{\mathrm{Lep}} - c_2}{16\max\{C_1,C_2\}} - \frac{c_1 C_{\mathrm{Lep}} - c_2}{8\max\{C_1,C_2\}} \\
&= -\frac{\big(2(d+m)-1\big)\cdot\big(c_1 C_{\mathrm{Lep}} - c_2\big)}{32\max\{C_1,C_2\}(d+m)} \\
&\le -\frac{\big(2(d+m)-1\big)\cdot\big(c_1 C_- - c_2\big)}{32\max\{C_1,C_2\}(d+m)} \\
&\le -1
\end{aligned}$$

because

$$C_- \ge c_1^{-1} \left[c_2 + \frac{32\max\{C_1,C_2\}(d+m)}{2(d+m)-1} \right].$$

Hence, for all $C_{\mathrm{Lep}} \ge C_-$, we have

$$\begin{aligned}
(5.1.31) &= \limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} 2C_{\Theta,J}\, n^{u(C_{\mathrm{Lep}})} \delta_n^{-d-m} n^{-\frac{\psi(C_{\mathrm{Lep}})}{8\max\{C_1,C_2\}}} \\
&\le \limsup_{n\to\infty} n^{-1} = 0 \,.
\end{aligned}$$

Ad (5.1.32). Because

$$h_l^{-\alpha_l-d-1} = \left(\frac{n}{\log n}\right)^{\frac{\alpha_l+d+1}{2\alpha_l+d}} \leq n^{\frac{b+d+1}{2a+d}} ,$$

we have

$$\begin{aligned}
&\log_n \left(n^{u(C_{\mathrm{Lep}})}\delta_n h_l^{-\alpha_l-d-1}\right) \\
\leq& \frac{c_1 C_{\mathrm{Lep}} - c_2}{32\max\{C_1,C_2\}(d+m)} - \frac{c_1 C_{\mathrm{Lep}} - c_2}{16\max\{C_1,C_2\}(d+m)} + \frac{b+d+1}{2a+d} \\
=& -\frac{c_1 C_{\mathrm{Lep}} - c_2}{32\max\{C_1,C_2\}(d+m)} + \frac{b+d+1}{2a+d} \\
\leq& -\frac{c_1 C_{-} - c_2}{32\max\{C_1,C_2\}(d+m)} + \frac{b+d+1}{2a+d} \\
\leq& -1
\end{aligned}$$

because

$$C_- \geq c_1^{-1}\left[c_2 + \left(\frac{b+d+1}{2a+d}+1\right)32\max\{C_1,C_2\}(d+m)\right] .$$

Hence, for all $C_{\mathrm{Lep}} \geq C_-$, we have

$$\begin{aligned}
(5.1.32) =& \limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \sup_{0\leq l\leq j\leq k_n(\alpha)} \widetilde{C}\, n^{u(C_{\mathrm{Lep}})}\delta_n h_l^{-\alpha_l-d-1} \\
\leq& \limsup_{n\to\infty} n^{-1} = 0 ,
\end{aligned}$$

concluding the proof.

□

Proof of Lemma 3.2.10. The general scheme of this proof coincides with the one of Lemma 3.2.9. We have to prove that for any $C_{\mathrm{Lep}} \geq C_-$, we have

$$\limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \sup_{0\leq l\leq j\leq k_n(\alpha)} n^{u(C_{\mathrm{Lep}})}\tilde{p}_{lj} < \infty ,$$

where

$$\tilde{p}_{lj} = \mathbb{P}_\gamma\Big(\sup_{x\in J,\theta\in\Theta} \|T_n(\theta,x;h_j)\| > \psi(C_{\mathrm{Lep}})r_l\Big) ,$$

where for abbreviation, we define the function $\psi : [\tilde{c}_1^{-1}(\tilde{c}_2+4),\infty) \to [4,\infty)$, $\psi(C_{\mathrm{Lep}}) := \tilde{c}_1 C_{\mathrm{Lep}} - \tilde{c}_2$ and note that ψ grows linearly in C_{Lep}. Let us use the decomposition of the process T_n into a canonical U-statistic T_n^1 and a linear process T_n^2 as described in (3.2.6) - (3.2.10), i.e.

$$T_n(\theta,x;h) = T_n^1(\theta,x;h) + T_n^2(\theta,x;h)$$

$$= \frac{1}{n(n-1)} \sum_{\substack{j,k=1 \\ j \neq k}}^{n} U_n\big(Z_j, Z_k, \theta, x; h\big) + \frac{2}{n} \sum_{j=1}^{n} u_n^*(Z_j, \theta, x; h) - \mathbb{E}_\gamma[u_n(Z_1, Z_2, \theta, x; h)] \,,$$

where for $z = (z_1, z_2^T)^T, y = (y_1, y_2^T)^T \in \mathbb{R} \times J$

$$\begin{aligned} U_n(z, y, \theta, x; h) &= u_n(z, y, \theta, x; h) - u_n^*(z, \theta, x; h) - u_n^*(y, \theta, x; h) \\ &\quad + \mathbb{E}_\gamma[u_n(Z_1, Z_2, \theta, x; h)] \,, \\ u_n(z, y, \theta, x; h) &= \partial_\theta \tau(z_1, y_1, \theta) K_h(z_2 - x) K_h(y_2 - x) \,, \\ u_n^*(z, \theta, x; h) &= \mathbb{E}_\gamma[u_n(Z_1, z, \theta, x; h)] = \mathbb{E}_\gamma\big[\partial_\theta \tau(z_1, Y_1, \theta) K_h(X_1 - x)\big] \cdot K_h(z_2 - x). \end{aligned}$$

This decomposition yields

$$\begin{aligned} \tilde{p}_{lj} &= \mathbb{P}_\gamma\Big(\sup_{x \in J, \theta \in \Theta} \|T_n(\theta, x; h_j)\| > \psi(C_{\mathrm{Lep}}) r_l \Big) \\ &\leq \mathbb{P}_\gamma\Big(\sup_{x \in J, \theta \in \Theta} \|T_n^1(\theta, x; h_j)\| > \psi(C_{\mathrm{Lep}}) r_l / 2 \Big) \\ &\quad + \mathbb{P}_\gamma\Big(\sup_{x \in J, \theta \in \Theta} \|T_n^2(\theta, x; h_j)\| > \psi(C_{\mathrm{Lep}}) r_l / 2 \Big) \\ &=: \tilde{p}_{lj}^1 + \tilde{p}_{lj}^2 \,. \end{aligned} \tag{5.1.33}$$

The assumptions of Theorem 3.2.7 imply all assumptions of Theorem 3.2.5 for the linear process T_n^2, particularly the exponential deviation inequality

$$\mathbb{P}_\gamma\big(\big\|T_n^2(\theta, x; h)\big\| \geq t\big) \leq 2 \exp\left(- \frac{t^2 n h^d}{C_1 + C_2 t} \right), \quad t > 0 \,,$$

so that Lemma 3.2.9 with $c_1 = \tilde{c}_1/2$, $c_2 = \tilde{c}_2/2$ gives

$$\limsup_{n \to \infty} \sup_{\alpha \in [a,b]} \sup_{\gamma \in \Gamma(\alpha)} \sup_{0 \leq l \leq j \leq k_n(\alpha)} n^{u(C_{\mathrm{Lep}})} \tilde{p}_{lj}^2 < \infty \,.$$

We will deal with the term $\tilde{p}_{lj}^1$ defined in (5.1.33) by using a discretization approach once again. According to Lemma 5.1.1 with $\rho = 1$, for any sequence $\delta_n \to 0$, there are nets $\Theta_n \times J_n \subset \Theta \times J$ with

$$\sup_{x \in J} \inf_{y \in J_n} \|x - y\| < \delta_n \,, \qquad \sup_{\vartheta \in \Theta} \inf_{\theta \in \Theta_n} \|\vartheta - \theta\| < \delta_n \,, \qquad \#(\Theta_n \times J_n) \leq C_{\Theta, J} \delta_n^{-d-m} \,,$$

where $C_{\Theta,J}$ is independent of n. Furthermore, for any $h > 0$,

$$\begin{aligned} &\sup_{x \in J, \theta \in \Theta} \|T_n^1(\theta, x; h)\| \\ &\leq \sup_{\substack{x, y \in J, \vartheta, \theta \in \Theta \\ \|x - y\|, \|\vartheta - \theta\| \leq \delta_n}} \|T_n^1(\theta, x; h) - T_n^1(\vartheta, y; h)\| + \sup_{x \in J_n, \theta \in \Theta_n} \|T_n^1(\theta, x; h)\| \,. \end{aligned} \tag{5.1.34}$$

The sequence δ_n will be specified later on in order to balance convergence rates. Now we may handle $\tilde{p}_{lj}^1$ by

$$\begin{aligned}
&\mathbb{P}_\gamma\Big(\sup_{x\in J,\theta\in\Theta}\|T_n^1(\theta,x;h_j)\|>\psi(C_{\mathrm{Lep}})h_l^{\alpha_l}/2\Big)\\
\leq&\,\mathbb{P}_\gamma\Big(\sup_{x\in J_n,\theta\in\Theta_n}\|T_n^1(\theta,x;h_j)\|>\psi(C_{\mathrm{Lep}})h_l^{\alpha_l}/4\Big) &(5.1.35)\\
&+\mathbb{P}_\gamma\Bigg(\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\leq\delta_n}}\|T_n^1(\theta,x;h_j)-T_n^1(\vartheta,y;h_j)\|>\psi(C_{\mathrm{Lep}})h_l^{\alpha_l}/4\Bigg) &(5.1.36)
\end{aligned}$$

according to (5.1.34). The term (5.1.35) can by handled by Bernstein's inequality for U-statistics, cf. Lemma 3.2.4, just like we did in the proof of Theorem 3.2.7 and the fact that $h_j\geq h_l$, i.e.

$$\begin{aligned}
&\mathbb{P}_\gamma\Big(\sup_{x\in J_n,\theta\in\Theta_n}\|T_n^1(\theta,x;h_j)\|>\psi(C_{\mathrm{Lep}})h_l^{\alpha_l}/4\Big)\\
\leq&\sum_{x\in J_n,\theta\in\Theta_n}\mathbb{P}_\gamma\big(\|T_n^1(\theta,x;h_j)\|>\psi(C_{\mathrm{Lep}})h_l^{\alpha_l}/4\big)\\
\leq&\sum_{x\in J_n,\theta\in\Theta_n}\mathbb{P}_\gamma\big(\|U_n(Z_1,Z_2,\theta,x;h_j)\|>n(n-1)\psi(C_{\mathrm{Lep}})h_l^{\alpha_l}/4\big)\\
\leq&C_{\Theta,J}\delta_n^{-d-m}\Big\{T\exp\Big(-T^{-1}nh_j^d\psi(C_{\mathrm{Lep}})h_l^{\alpha_l}/4\Big)\mathbb{1}_{\psi(C_{\mathrm{Lep}})h_l^{\alpha_l}/4\in[0,1)}\\
&+T\exp\Big(-T^{-1}nh_j^d\sqrt{\psi(C_{\mathrm{Lep}})h_l^{\alpha_l}/4}\Big)\mathbb{1}_{\psi(C_{\mathrm{Lep}})h_l^{\alpha_l}/4\in[1,\infty)}\Big\}\\
\leq&C_{\Theta,J}\delta_n^{-d-m}\Big\{T\exp\Big(-T^{-1}n\psi(C_{\mathrm{Lep}})h_l^{\alpha_l+d}/4\Big)\mathbb{1}_{\psi(C_{\mathrm{Lep}})h_l^{\alpha_l}/4\in[0,1)}\\
&+T\exp\Big(-T^{-1}n\sqrt{\psi(C_{\mathrm{Lep}})}h_l^{\alpha_l/2+d}/2\Big)\mathbb{1}_{\psi(C_{\mathrm{Lep}})h_l^{\alpha_l}/4\in[1,\infty)}\Big\}. &(5.1.37)
\end{aligned}$$

By using $h_l=\big(\frac{\log n}{n}\big)^{\frac{1}{2\alpha_l+d}}$, we get that

$$nh_l^{\alpha_l/2+d}\geq nh_l^{\alpha_l+d}=n\Big(\frac{\log n}{n}\Big)^{\frac{\alpha_l+d}{2\alpha_l+d}}\geq\frac{n\log n}{n}=\log n$$

yielding

$$\begin{aligned}
(5.1.37)\leq&C_{\Theta,J}T\delta_n^{-d-m}\Big(n^{-T^{-1}\psi(C_{\mathrm{Lep}})/4}+n^{-T^{-1}\sqrt{\psi(C_{\mathrm{Lep}})}/2}\Big)\\
\leq&2C_{\Theta,J}T\delta_n^{-d-m}n^{-T^{-1}\sqrt{\psi(C_{\mathrm{Lep}})}/2}\,, &(5.1.38)
\end{aligned}$$

because $\psi(C_{\mathrm{Lep}})\geq 4$. The term (5.1.36) is handled by arguments similar to the ones found in the proof of Theorem 3.2.7. Using Markov's inequality, $h_j\geq h_l$ and the arguments used to treat (5.1.20) - (5.1.22) for $\rho=1$ that showed that the expectation in the

following display is $O(\delta_n h_j^{-2d-1})$, we deduce that there is a constant $\widetilde{C}$ so that

$$\begin{aligned}
&\limsup_{n\to\infty} \sup_{*} \delta_n^{-1} h_l^{\alpha_l+2d+1} \mathbb{P}_\gamma \Bigg(\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta \\ \|x-y\|,\|\vartheta-\theta\|\le\delta_n}} \left\| T_n^1(\theta,x;h_j) - T_n^1(\vartheta,y;h_j) \right\| > \psi(C_{\mathrm{Lep}}) h_l^{\alpha_l}/4 \Bigg) \\
&\le \limsup_{n\to\infty} \sup_{*} \frac{4\delta_n^{-1} h_j^{2d+1}}{\psi(C_{\mathrm{Lep}})} \mathbb{E}_\gamma \Bigg[\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta \\ \|x-y\|,\|\vartheta-\theta\|\le\delta_n}} \left\| T_n^1(\theta,x;h_j) - T_n^1(\vartheta,y;h_j) \right\| \Bigg] < \widetilde{C} < \infty \,,
\end{aligned} \tag{5.1.39}$$

where the suprema are taken over $\alpha \in [a,b]$, $\gamma \in \Gamma(\alpha)$, $0 \le l \le j \le k_n(\alpha)$.

Combining (5.1.38) and (5.1.39) yields

$$\begin{aligned}
&\limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \sup_{0\le l\le j\le k_n(\alpha)} n^{u(C_{\mathrm{Lep}})} \tilde{p}_{lj}^1 \\
&\le \limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} 2C_{\Theta,J} T\, n^{u(C_{\mathrm{Lep}})} \delta_n^{-d-m} n^{-T^{-1}\sqrt{\psi(C_{\mathrm{Lep}})}/2} && (5.1.40) \\
&\quad + \limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \sup_{0\le l\le j\le k_n(\alpha)} \widetilde{C}\, n^{u(C_{\mathrm{Lep}})} \delta_n h_l^{-\alpha_l-2d-1} \,, && (5.1.41)
\end{aligned}$$

which will be finite by choosing

$$\delta_n = \delta_n(C_{\mathrm{Lep}}) = n^{-\frac{T^{-1}\sqrt{\psi(C_{\mathrm{Lep}})}}{4(d+m)}} = n^{-\frac{T^{-1}\sqrt{\tilde{c}_1 C_{\mathrm{Lep}} - \tilde{c}_2}}{4(d+m)}} \,.$$

In order to treat (5.1.40), we see that

$$\begin{aligned}
&\log_n \left(n^{u(C_{\mathrm{Lep}})} \delta_n^{-d-m} n^{-T^{-1}\sqrt{\psi(C_{\mathrm{Lep}})}/2} \right) \\
&\le \frac{T^{-1}\sqrt{\psi(C_{\mathrm{Lep}})}}{8(d+m)} + \frac{T^{-1}\sqrt{\psi(C_{\mathrm{Lep}})}}{4} - \frac{T^{-1}\sqrt{\psi(C_{\mathrm{Lep}})}}{2} \\
&= - \frac{\big(2(d+m)-1\big) T^{-1}\sqrt{\psi(C_{\mathrm{Lep}})}}{8(d+m)} \\
&\le - \frac{\big(2(d+m)-1\big) T^{-1}\sqrt{\tilde{c}_1 C_- - \tilde{c}_2}}{8(d+m)} \\
&\le -1
\end{aligned}$$

because

$$C_- \ge \tilde{c}_1^{-1} \left[\tilde{c}_2 + T^2 \left(\frac{8(d+m)}{2(d+m)-1} \right)^2 \right].$$

Hence, for all $C_{\mathrm{Lep}} \ge C_-$, we have

$$(5.1.40) = \limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} 2C_{\Theta,J} T\, n^{u(C_{\mathrm{Lep}})} \delta_n^{-d-m} n^{-T^{-1}\sqrt{\psi(C_{\mathrm{Lep}})}/2}$$

$$\leq \limsup_{n\to\infty} n^{-1} = 0$$

Ad (5.1.41). Because

$$h_l^{-\alpha_l-2d-1} = \left(\frac{n}{\log n}\right)^{\frac{\alpha_l+2d+1}{2\alpha_l+d}} \leq n^{\frac{b+2d+1}{2a+d}} ,$$

we have

$$\begin{aligned}
&\log_n \left(n^{u(C_{\mathrm{Lep}})}\delta_n h_l^{-\alpha_l-2d-1}\right) \\
&\leq \frac{T^{-1}\sqrt{\psi(C_{\mathrm{Lep}})}}{8(d+m)} - \frac{T^{-1}\sqrt{\psi(C_{\mathrm{Lep}})}}{4(d+m)} + \frac{b+2d+1}{2a+d} \\
&= -\frac{T^{-1}\sqrt{\psi(C_{\mathrm{Lep}})}}{8(d+m)} + \frac{b+2d+1}{2a+d} \\
&\leq -\frac{T^{-1}\sqrt{\tilde{c}_1 C_- - \tilde{c}_2}}{8(d+m)} + \frac{b+2d+1}{2a+d} \\
&\leq -1
\end{aligned}$$

because

$$C_- \geq \tilde{c}_1^{-1}\left[\tilde{c}_2 + 64T^2(d+m)^2\left(\frac{b+2d+1}{2a+d}+1\right)^2\right] .$$

Hence, for all $C_{\mathrm{Lep}} \geq C_-$, we have

$$\begin{aligned}
(5.1.41) &= \limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \sup_{0\leq l\leq j\leq k_n(\alpha)} \widetilde{C}\, n^{u(C_{\mathrm{Lep}})}\delta_n h_l^{-\alpha_l-2d-1} \\
&\leq \limsup_{n\to\infty} n^{-1} = 0 ,
\end{aligned}$$

concluding the proof.

□

5.2 Proofs for Section 4.1

5.2.1 Proofs for the identifiability results

Proof of Theorem 4.1.2. Assume there is another representation

$$f(\cdot|x) = \sum_{v=1}^{V} \lambda_v(x)\phi\big(\,\cdot\,\big|\nu_v(x), \delta_v^2(x)\big) , \quad x \in \mathcal{O}$$

with $m \geq V \geq 2$, continuous functions $\lambda_1, \ldots, \lambda_V > 0$ and differentiable functions $\nu_1, \ldots, \nu_V$, $\delta_1, \ldots, \delta_V$.

Let

$$S := \big\{x \in \mathcal{O} : \big(\mu_c(x), \sigma_c(x)\big) = \big(\mu_{c'}(x), \sigma_{c'}(x)\big) \text{ , for some } c \neq c'\big\} .$$

First of all, for any $u \in \mathcal{O}$ with $(\mu_c(u), \sigma_c(u)) \neq (\mu_{c'}(u), \sigma_{c'}(u))$ for all $c, c' \in \{1, \dots, m\}$, $c \neq c'$, Teicher (1963, Proposition 1) gives $m = V$ and a permutation τ_u so that

$$\lambda_{\tau_u(c)}(u) = \pi_c(u) \ , \quad \nu_{\tau_u(c)}(u) = \mu_c(u) \ , \quad \delta_{\tau_u(c)}(u) = \sigma_c(u) \ .$$

Fix some $x \in S^c$. By continuity of the parameter curves μ_c, σ_c and S^c being open, there is a neighbourhood $U \subset S^c$ of x so that for all $u, u' \in U$, $c, c' \in \{1, \dots, m\}$, $c \neq c'$

$$\big(\mu_c(u), \sigma_c(u)\big) \neq \big(\mu_{c'}(u'), \sigma_{c'}(u')\big) \ .$$

Hence, $\tau_u = \tau_x$ for all $u \in U$ by continuity of the ν_c and δ_c, $c = 1, \dots, m$. As S^c is open, we get that for every connected component U' of S^c there is a permutation τ fulfilling (4.1.1) for all $u \in U'$.

Now, fix any $x \in S$. There is some $r \in \{1, \dots, (m-1)!\}$ and $c_1 \neq c_1', \dots, c_r \neq c_r'$ so that for any $j \in \{1, \dots, r\}$

$$\big(\mu_{c_j}(x), \sigma_{c_j}(x)\big) = \big(\mu_{c_j'}(x), \sigma_{c_j'}(x)\big) \ , \quad D^j(x) := D^{c_j, c_j'}(x) \leq d - 1$$

and that for all other pairs $c \neq c'$, we have

$$\big(\mu_c(x), \sigma_c(x)\big) \neq \big(\mu_{c'}(x), \sigma_{c'}(x)\big) \ , \tag{5.2.1}$$

where

$$D^{c,c'}(x) := \max \Big\{k \in \{1, \dots, d\} | \exists\, z_1, \dots, z_k \in \mathbb{R}^d, \|z_i\| = 1,\ z_i \text{ linearly independent} : \\ \forall i \in \{1, \dots, k\}\ \partial_{z_i}\mu_c(x) = \partial_{z_i}\mu_{c'}(x)\ , \partial_{z_i}\sigma_c(x) = \partial_{z_i}\sigma_{c'}(x)\Big\} \ .$$

According to our assumptions, for any $j \in \{1, \dots, r\}$ there are linearly independent vectors $x_1^j, \dots, x_{D^j(x)}^j$ so that for all $s \in \{1, \dots, D^j(x)\}$

$$\partial_{x_s^j}\big(\mu_{c_j}(x), \sigma_{c_j}(x)\big) = \partial_{x_s^j}\big(\mu_{c_j'}(x), \sigma_{c_j'}(x)\big)$$

and for any directional vector $z \in A_j := \{z \in \mathbb{R}^d | \, \|z\| = 1\} \backslash \operatorname{span}\{x_1^j, \dots, x_{D^j(x)}^j\}$

$$\partial_z\big(\mu_{c_j}(x), \sigma_{c_j}(x)\big) \neq \partial_z\big(\mu_{c_j'}(x), \sigma_{c_j'}(x)\big) \ .$$

Now, as for any $j \in \{1, \dots, r\}$ we have $\dim\big(\operatorname{span}\{x_1^j, \dots, x_{D^j(x)}^j\}\big) \leq d - 1$, it follows that

$$A := \bigcap_{j=1}^{r} A_j \neq \emptyset \ .$$

Hence, for any $z \in A$ and any $j \in \{1, \dots, r\}$

$$\partial_z \big(\mu_{c_j}(x), \sigma_{c_j}(x)\big) \neq \partial_z \big(\mu_{c'_j}(x), \sigma_{c'_j}(x)\big) \,. \tag{5.2.2}$$

Fix any $z \in A$. By continuity of the curves and (5.2.2), there is an $\varepsilon(z) > 0$ so that for any $y \in (x - \varepsilon(z)z, x + \varepsilon(z)z)\backslash\{x\}$, $j \in \{1, \dots, r\}$, we have

$$\big(\mu_{c_j}(y), \sigma_{c_j}(y)\big) \neq \big(\mu_{c'_j}(y), \sigma_{c'_j}(y)\big) \,.$$

Again by continuity of the parameter functions and by (5.2.1), for any other pair $c \neq c'$, there is an $\varepsilon^{c,c'} > 0$ so that for any $y \in (x - \varepsilon^{c,c'}z, x + \varepsilon^{c,c'}z)\backslash\{x\}$, we have

$$\big(\mu_c(y), \sigma_c(y)\big) \neq \big(\mu_{c'}(y), \sigma_{c'}(y)\big) \,.$$

By choosing $\varepsilon > 0$ minimal, we get that

$$\Big((x - \varepsilon z, x + \varepsilon z)\backslash\{x\}\Big) \cap S = \emptyset \,.$$

According to our prior observations, there are permutations τ_+, τ_- so that for all $c = 1, \dots, m$

$$\begin{aligned}
\nu_{\tau_-(c)}(u) = \mu_c(u) \,, \quad \delta_{\tau_-(c)}(u) = \sigma_c(u) \,, \quad u \in (x - \varepsilon z, x) \,, \\
\nu_{\tau_+(c)}(u) = \mu_c(u) \,, \quad \delta_{\tau_+(c)}(u) = \sigma_c(u) \,, \quad u \in (x, x + \varepsilon z) \,.
\end{aligned}$$

Hence, for any $c \in \{1, \dots, m\}$, $u \in (x - \varepsilon z, x + \varepsilon z)\backslash\{x\}$, we have

$$\begin{aligned}
\nu_c(u) =& \mu_{\tau_-^{-1}(c)}(u) \mathbb{1}_{u \in (x - \varepsilon z, x)} + \mu_{\tau_+^{-1}(c)}(u) \mathbb{1}_{u \in (x, x + \varepsilon z)} \\
\delta_c(u) =& \sigma_{\tau_-^{-1}(c)}(u) \mathbb{1}_{u \in (x - \varepsilon z, x)} + \sigma_{\tau_+^{-1}(c)}(u) \mathbb{1}_{u \in (x, x + \varepsilon z)}
\end{aligned}$$

By continuity of the parameter functions ν_c and δ_c, we have

$$\begin{aligned}
\lim_{\eta \to 0} \big(\mu_{\tau_-^{-1}(c)}(x - \eta z), \sigma_{\tau_-^{-1}(c)}(x - \eta z)\big) &= \lim_{\eta \to 0} \big(\nu_c(x - \eta z), \delta_c(x - \eta z)\big) \\
&= \lim_{\eta \to 0} \big(\nu_c(x + \eta z), \delta_c(x + \eta z)\big) \\
&= \lim_{\eta \to 0} \big(\mu_{\tau_+^{-1}(c)}(x + \eta z), \sigma_{\tau_+^{-1}(c)}(x + \eta z)\big)
\end{aligned}$$

so that

$$\big(\mu_{\tau_-^{-1}(c)}(x), \sigma_{\tau_-^{-1}(c)}(x)\big) = \big(\mu_{\tau_+^{-1}(c)}(x), \sigma_{\tau_+^{-1}(c)}(x)\big) \,.$$

So either, $\tau_+^{-1}(c) = \tau_-^{-1}(c)$ or

$$\big(\partial_z \mu_{\tau_-^{-1}(c)}(x), \partial_z \sigma_{\tau_-^{-1}(c)}(x)\big) \neq \big(\partial_z \mu_{\tau_+^{-1}(c)}(x), \partial_z \sigma_{\tau_+^{-1}(c)}(x)\big) \,,$$

a contradiction to the differentiability of the parameter functions ν_c and δ_c, as it implies

$$\lim_{\eta \to 0} \frac{\mu_{\tau_-^{-1}(c)}(x - \eta z) - \mu_{\tau_-^{-1}(c)}(x)}{\eta} = \lim_{\eta \to 0} \frac{\nu_c(x - \eta z) - \nu_c(x)}{\eta}$$

$$
\begin{aligned}
&= \lim_{\eta\to 0} \frac{\nu_c(x+\eta z) - \nu_c(x)}{\eta} \\
&= \lim_{\eta\to 0} \frac{\mu_{\tau_+^{-1}(c)}(x+\eta z) - \mu_{\tau_+^{-1}(c)}(x)}{\eta} \,, \\
\lim_{\eta\to 0} \frac{\sigma_{\tau_-^{-1}(c)}(x-\eta z) - \sigma_{\tau_-^{-1}(c)}(x)}{\eta} &= \lim_{\eta\to 0} \frac{\delta_c(x-\eta z) - \delta_c(x)}{\eta} \\
&= \lim_{\eta\to 0} \frac{\delta_c(x+\eta z) - \delta_c(x)}{\eta} \\
&= \lim_{\eta\to 0} \frac{\sigma_{\tau_+^{-1}(c)}(x+\eta z) - \sigma_{\tau_+^{-1}(c)}(x)}{\eta} \,.
\end{aligned}
$$

Hence, $\tau_+ = \tau_-$, completing the proof. □

The proof of Theorem 4.1.3 is an easier version of the one for Theorem 4.1.2.

Proof of Theorem 4.1.3. Assume there is another representation

$$
f(\cdot|x) = \sum_{v=1}^{V} \lambda_v(x)\phi\big(\,\cdot\,\big|\nu_v(x), \delta_v^2(x)\big) \,, \quad x \in I
$$

with $m \geq V \geq 2$, positive functions $\lambda_1, \ldots, \lambda_V$ and continuous functions $\nu_1, \ldots, \nu_V$, $\delta_1, \ldots, \delta_V$.

Like before, we see that for any $x \in I$ there is a neighbourhood U of x within the compact space I so that there is a permutation τ fulfilling (4.1.1) for all $u \in U$. This concludes the proof as there are at most countably many disjoint open intervals on the path between any two points $x, x' \in I$.

□

5.2.2 Proofs for the estimation results

In order to prove both Theorems 4.1.6 and 4.1.8, it is enough to prove Assumptions **(B1)** - **(B9)** according to Theorems 3.1.6 and 3.1.13. The complete proof is given by the assembly of the lemmata and proofs in this section.

Fix some bounded and convex open $\Theta \subset \Xi \subset \mathbb{R}^{3m-1}$ so that

$$
\overline{\Xi} = \Lambda_m \times [\mu_-, \mu_+]^m \times [\sigma_-, \sigma_+]^m \,,
$$

where

$$
\Lambda_m = \Big\{(\pi_1, \ldots, \pi_{m-1}) \in [\pi_-, \pi_+]^{m-1} \Big| 1 - \textstyle\sum_{c=1}^{m-1} \pi_c \in [\pi_-, \pi_+]\Big\} \tag{5.2.3}
$$

for some

$$
0 < \pi_- < 1/m < \pi_+ < 1 \,, \quad \mu_- < \mu_+ \,, \quad 0 < \sigma_- < \sigma_+ < \infty \,.
$$

Note that for fixed x, γ, h the contrast functions $M(\cdot, x; \gamma)$ and $M_n(\cdot, x; h)$ are defined on $\mathfrak{X} = \mathcal{S}_m \times \mathbb{R}^m \times (0, \infty)^m$, which is a superset of $\overline{\Xi}$. The model is identifiable over $\overline{\Xi}$ up to relabeling and the contrast property of M also holds over $\overline{\Xi}$ because it was established over $\mathcal{S}_m \times \mathbb{R}^m \times (0, \infty)^m$ as briefly discussed in Section 4.1.2.

Throughout the proofs in this section, let us use the notation $a_n \lesssim b_n$ only if $a_n \leq C b_n$ for $n \geq n_0$ and C depends only on I, $\overline{\Xi}$, U, U_π, U_μ, U_σ, U_ℓ, L, $[a, b]$ or is universal. In particular, the constant C then is independent of specific θ, x, h, α or γ.

Auxiliary results

Remark 5.2.1. We will deal with second-order derivatives of the log-likelihood function. Therefore, let us differentiate g twice. By minding the dependence $\pi_m = 1 - \pi_1 - \cdots - \pi_{m-1}$, we see that for any $t_1, t_2 \in \{\pi_1, \ldots, \pi_{m-1}, \sigma_1, \ldots, \sigma_m, \mu_1, \ldots, \mu_m\}$, we have

$$\begin{aligned}
\partial_{t_1} g(y; \theta) &= \frac{\partial_{t_1} f_{\mathrm{mix}}(y; \theta)}{f_{\mathrm{mix}}(y; \theta)} \\
\partial_{t_1} \partial_{t_2} g(y; \theta) &= \frac{\partial_{t_1} \partial_{t_2} f_{\mathrm{mix}}(y; \theta)}{f_{\mathrm{mix}}(y; \theta)} - \frac{\big(\partial_{t_1} f_{\mathrm{mix}}(y; \theta)\big)\big(\partial_{t_2} f_{\mathrm{mix}}(y; \theta)\big)}{f^2_{\mathrm{mix}}(y; \theta)} \,,
\end{aligned}$$

where for any $c, c' \in \{1, \ldots, m\}$, $\tilde{c}, \tilde{c}' \in \{1, \ldots, m-1\}$, $y \in \mathbb{R}$, $\theta \in \overline{\Xi}$,

$$\begin{aligned}
f_{\mathrm{mix}}(y; \theta) &= \sum_{c=1}^{m} \pi_c \phi(y | \mu_c, \sigma_c^2) \,, \\
\partial_{\pi_{\tilde{c}}} f_{\mathrm{mix}}(y; \theta) &= \phi(y | \mu_{\tilde{c}}, \sigma_{\tilde{c}}^2) - \phi(y | \mu_m, \sigma_m^2) \,, \\
\partial_{\mu_c} f_{\mathrm{mix}}(y; \theta) &= \pi_c (\mu_c - y) \sigma_c^{-2} \phi(y | \mu_c, \sigma_c^2) \,, \\
\partial_{\sigma_c} f_{\mathrm{mix}}(y; \theta) &= \pi_c \big((y - \mu_c)^2 - \sigma_c^2\big) \sigma_c^{-3} \phi(y | \mu_c, \sigma_c^2) \,, \\
\partial_{\pi_{\tilde{c}}} \partial_{\pi_{\tilde{c}'}} f_{\mathrm{mix}}(y; \theta) &= 0 \,, \\
\partial_{\mu_c} \partial_{\pi_{\tilde{c}'}} f_{\mathrm{mix}}(y; \theta) &= \mathbb{1}_{c=\tilde{c}'} \partial_{\mu_c} \phi(y | \mu_c, \sigma_c^2) - \mathbb{1}_{c=m} \partial_{\mu_m} \phi(y | \mu_m, \sigma_m^2) \,, \\
&= \Big(\mathbb{1}_{c=\tilde{c}'} - \mathbb{1}_{c=m}\Big) \cdot (\mu_c - y) \sigma_c^{-2} \phi(y | \mu_c, \sigma_c^2) \,, \\
\partial_{\sigma_c} \partial_{\pi_{\tilde{c}'}} f_{\mathrm{mix}}(y; \theta) &= \mathbb{1}_{c=\tilde{c}'} \partial_{\sigma_c} \phi(y | \mu_c, \sigma_c^2) - \mathbb{1}_{c=m} \partial_{\sigma_m} \phi(y | \mu_m, \sigma_m^2) \\
&= \Big(\mathbb{1}_{c=\tilde{c}'} - \mathbb{1}_{c=m}\Big) \cdot \big((y - \mu_c)^2 - \sigma_c^2\big) \sigma^{-3} \phi(y | \mu_c, \sigma_c^2) \\
\partial_{\mu_c} \partial_{\mu_{c'}} f_{\mathrm{mix}}(y; \theta) &= \mathbb{1}_{c=c'} \bigg\{ \Big(\big\{ \pi_c (\mu_c - y) \sigma_c^{-2} \big\}^2 + \pi_c \sigma_c^{-2} \Big) \phi(y | \mu_c, \sigma_c^2) \bigg\} \,, \\
\partial_{\sigma_c} \partial_{\mu_{c'}} f_{\mathrm{mix}}(y; \theta) &= \mathbb{1}_{c=c'} \bigg\{ \Big(\Big[\big(\pi_c (\mu_c - y)\big) \sigma_c^{-2} \Big] \Big[\pi_c \big((y - \mu_c)^2 - \sigma_c^2\big) \sigma_c^{-3} \Big] \\
&\qquad - 2 \pi_c (\mu_c - y) \sigma_c^{-3} \Big) \phi(y | \mu_c, \sigma_c^2) \bigg\} \,, \\
\partial_{\sigma_c} \partial_{\sigma_{c'}} f_{\mathrm{mix}}(y; \theta) &= \mathbb{1}_{c=c'} \bigg\{ \Big(\pi_c \sigma_c^2 + \pi_c (y - \mu_c) \sigma_c - \sigma_c^3 - \pi_c (y - \mu_c) \Big) \sigma_c^{-4} \phi(y | \mu_c, \sigma_c^2) \bigg\} \,.
\end{aligned}$$

Let us start with some auxiliary results that are proved in Section 5.2.3. The following lemma gives integrability of the log-likelihood function and its derivatives with respect to the model mixture densities.

Lemma 5.2.2. *Under Assumption* **(N1)**, *for any* $\rho \in [1, \infty)$ *and any function* $\tau \in \{g, \partial_{t_1} g, \partial_{t_1}\partial_{t_2} g\}$, *where* $t_1, t_2 \in \{\pi_1, \dots, \pi_{m-1}, \sigma_1, \dots, \sigma_m, \mu_1, \dots, \mu_m\}$, *we have*

$$\sup_{\gamma\in\Gamma} \int \Big(\sup_{\theta\in\overline{\Xi}} \big|\tau(y;\theta)\big| \Big)^{\rho} \sup_{x\in I} f_{Y|X}^{\theta_*(\cdot)}(y|x)\,\mathrm{d}y < \infty\,, \tag{5.2.4}$$

$$\int \Big(\sup_{\theta\in\overline{\Xi}} \big|\tau(y;\theta)\big| \Big)^{\rho} \sup_{\vartheta\in\overline{\Xi}} f_{\mathrm{mix}}(y;\vartheta)\,\mathrm{d}y < \infty\,. \tag{5.2.5}$$

Additionally, for any normal density $\phi(\cdot|0, s^2)$, *we have*

$$\int \Big(\sup_{\theta\in\overline{\Xi}} \big|\tau(y;\theta)\big| \Big)^{\rho} \phi(y|0, s^2)\,\mathrm{d}y < \infty\,. \tag{5.2.6}$$

The following lemma gives Lipschitz continuity of the mixture density and derivatives of the log-likelihood in θ with integrable Lipschitz constant.

Lemma 5.2.3. *Let Assumption* **(N1)** *hold.*

(i) For any $\theta, \tilde{\theta} \in \overline{\Xi}$, *we have*

$$|f_{\mathrm{mix}}(y;\theta) - f_{\mathrm{mix}}(y;\tilde{\theta})| \leq f_{\overline{\Xi}}(y;\theta,\tilde{\theta})\|\theta - \tilde{\theta}\|_1\,,$$

where $f_{\overline{\Xi}}$ *is a non-negative function so that*

$$\sup_{\theta,\tilde{\theta}\in\overline{\Xi}} f_{\overline{\Xi}}(\cdot;\theta,\tilde{\theta}) \leq c_* \phi(\cdot|0, s_*^2)$$

for some constants $0 < c_*, s_* < \infty$.

(ii) Let $0 < a \leq b < \infty$. *There is a constant* L^* *depending only on* I, $\overline{\Xi}$, a, b *and the Hölder constant* L *so that for any* $\alpha \in [a, b]$, $1 > \delta > 0$

$$\sup_{\substack{x,x'\in I \\ \|x-x'\|\leq\delta}} |f_{Y|X}^{\theta_*(\cdot)}(y|x) - f_{Y|X}^{\theta_*(\cdot)}(y|x')| \leq \sup_{\theta,\tilde{\theta}\in\overline{\Xi}} f_{\overline{\Xi}}(\cdot;\theta,\tilde{\theta}) L^* \delta^{\min\{1,\alpha\}}\,,$$

where $f_{\overline{\Xi}}$ *is defined in (i).*

(iii) For any functions $\tau \in \{\partial_{t_1} g, \partial_{t_1}\partial_{t_2} g\}$, *where* $t_1, t_2 \in \{\pi_1, \dots, \pi_{m-1}, \mu_1, \dots, \mu_m, \sigma_1, \dots, \sigma_m\}$ *and any* $y \in \mathbb{R}$, $\theta, \tilde{\theta} \in \overline{\Xi}$, *we have*

$$\big|\tau(y;\theta) - \tau(y;\tilde{\theta})\big| \leq f_\tau(y;\theta,\tilde{\theta})\|\theta - \tilde{\theta}\|\,,$$

where f_τ *is a non-negative function so that* $F^\rho(\cdot;\overline{\Xi}) := \sup_{\theta,\tilde{\theta}\in\overline{\Xi},\tau} f_\tau^\rho(\cdot;\theta,\tilde{\theta})$ *is integrable with respect to* $\sup_{\theta\in\overline{\Xi}} f_{\mathrm{mix}}(y;\theta)$ *for every* $\rho \geq 1$.

Lemma 5.2.4. *Let $0 < a \leq b < \infty$ and $(h_n(\alpha))_{n\in\mathbb{N}}$ for $\alpha \in [a,b]$ be sequences of bandwidth parameters so that*

$$\sup_{\alpha\in[a,b]} h_n(\alpha) \,, \quad \sup_{\alpha\in[a,b]} \frac{\log n}{n h_n(\alpha)^d} \to 0 \,.$$

Under Assumption **(N1)**, *for any compact $J \subset \operatorname{int}(I)$ and any function*

$$\tau \in \{g, \partial_{\pi_c} g, \partial_{\mu_c} g, \partial_{\sigma_c} g\} \,,$$

we have that

$$\sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(a)} n h_n(\alpha)^d \sup_{x\in J, \theta\in\overline{\Xi}} \operatorname{Var}_\gamma\left(\frac{1}{n}\sum_{k=1}^n \tau(Y_k;\theta) K_h(X_k - x)\right) = O(1) \,.$$

The following lemma gives an exponential deviation inequality for the empirical contrast's gradient.

Lemma 5.2.5. *Under Assumption 4.1.4, for any compact $J \subset \operatorname{int}(I)$, there are constants $C_1, C_2 > 0$ independent of n, h, θ, x, γ so that for any $t \in \{\pi_1, \dots, \pi_{m-1}, \mu_1, \dots, \mu_m, \sigma_1, \dots, \sigma_m\}$, $\theta \in \overline{\Xi}$, $x \in J$, $h > 0$, we have*

$$\mathbb{P}_\gamma\Big(\Big|\partial_t M_n(\theta, x; h) - \mathbb{E}_\gamma\big[\partial_t M_n(\theta, x; h)\big]\Big| \geq \omega\Big) \leq 2\exp\bigg(-\frac{\omega^2 n h^d}{(C_1 + C_2\omega)}\bigg) \,, \quad \omega > 0 \,.$$

Main proofs

Let us first prove that the polytope

$$\Lambda_m = \Big\{(\pi_1, \dots, \pi_{m-1}) \in [\pi_-, \pi_+]^{m-1} \Big| 1 - \textstyle\sum_{c=1}^{m-1} \pi_c \in [\pi_-, \pi_+]\Big\}$$

as defined in (5.2.3), fulfils the latter part of Assumption **(B1)** for any $0 < \pi_- < 1/m < \pi_+ < 1$, which in particular implies this property for

$$\Theta = \Big\{(\pi_1, \dots, \pi_{m-1}) \in U_\pi^{m-1} \Big| 1 - \textstyle\sum_{c=1}^{m-1} \pi_c \in U_\pi\Big\} \times U_\mu^m \times U_\sigma^m \,,$$

as U_π, U_μ, U_σ are compact intervals with $1/m \in \operatorname{int}(U_\pi)$ by Assumption **(N1)**.

Lemma 5.2.6.

(i) For every $\lambda_0 \in \partial\,\Lambda_m$ and any $\varepsilon > 0$, there are some $1 \leq l \leq m-1$, $\lambda_1, \dots, \lambda_l \in \Lambda_m$ so that

$$\lambda_{k+1} - \lambda_k = c_k e_{j_k} \,, \quad k = 0, \dots, l-1 \,, \quad \lambda_l \in \operatorname{int}(\Lambda_m)$$

for some unit vectors e_{j_k} and some coefficients $c_k \in \mathbb{R}$ with $\sum_{k=0}^{l-1} |c_k| \leq \varepsilon$.

(ii) For every $\lambda_0, \lambda' \in \Lambda_m$, $\lambda_0 \neq \lambda'$, there is some $l \in \mathbb{N}$, $\lambda_1, \dots, \lambda_l \in \Lambda_m$ so that with $\lambda_{l+1} = \lambda'$, we have

$$\lambda_{k+1} - \lambda_k = c_k e_{j_k} , \quad k = 0, \dots, l ,$$

for some unit vectors e_{j_k} and some coefficients $c_k \in \mathbb{R}$ with

$$\sum_{k=0}^{l} |c_k| \leq (2m-1)\|\lambda_0 - \lambda'\|_1 .$$

Proof. (i) Denote $\lambda_0 = (\pi_1, \dots, \pi_{m-1})^T$. As $\lambda_0 \in \partial\,\Lambda_m$, it holds that

$$\|\lambda_0\|_1 = \sum_{c=1}^{m-1} \pi_c \in \{1-\pi_+, 1-\pi_-\} \quad \vee \quad \exists\, c \in \{1, \dots, m-1\} :\ \pi_c \in \{\pi_-, \pi_+\} .$$

We will distinguish between the cases

$$(1)\ \ \|\lambda_0\|_1 = 1-\pi_+ , \quad (2)\ \ \|\lambda_0\|_1 = 1-\pi_- , \quad (3)\ \ \|\lambda_0\|_1 \in (1-\pi_+, 1-\pi_-) .$$

We first show that cases (1) and (2) can be reduced to case (3). This is because for vectors λ_0 fulfilling (1), there needs to be a component π_c that can be increased by a small amount without leaving Λ_m, resulting in a vector fulfilling (3). The reduction from (2) to (3) works analogously. In case λ_0 fulfils (3), sufficiently small changes to the components of λ_0 that are either π_- or π_+ results in a vector λ_l with $\|\lambda_l\|_1 \in (1-\pi_+, 1-\pi_-)$ and components in (π_-, π_+).

Let us prove this rigorously. First show the reduction from (1) to (3). Assume $\|\lambda_0\|_1 = 1-\pi_-$. Then, there is some $c \in \{1, \dots, m-1\}$ so that $\pi_c > \pi_-$. Because otherwise we would have

$$1-\pi_- = \|\lambda_0\|_1 = (m-1)\pi_- ,$$

a contradiction as $\pi_- < 1/m$. Define

$$\xi_c = -\min\left\{\frac{\varepsilon}{2}, \frac{\pi_c - \pi_-}{2}\right\} e_c$$

for the c-th unit vector $e_c \in \mathbb{R}^{m-1}$. Then, the vector

$$\lambda_1 = \lambda_0 + \xi_c \in \Lambda_m$$

fulfils

$$1-\pi_+ < \|\lambda_1\|_1 < 1-\pi_- , \quad \|\lambda_1 - \lambda_0\|_1 = \|\xi_c\|_1 \leq \varepsilon/2.$$

Let us consider case (3), i.e. $\|\lambda_0\|_1 \in (1-\pi_+, 1-\pi_-)$.

Define $\bar{\pi} = \frac{\pi_- + \pi_+}{2}$, the sets

$$\begin{aligned} \mathcal{J}_1 &= \left\{ c \in \{1, \dots, m-1\} : \pi_c \geq \bar{\pi} \right\} , \\ \mathcal{J}_2 &= \mathcal{J}_1^c , \\ J_i &= \#\mathcal{J}_i \quad i = 1, 2 \end{aligned}$$

and set $1/J_i = 0$ for $J_i = 0$, $i = 1, 2$. Furthermore, define

$$\bar{\varepsilon} = \min \left\{ \frac{\|\lambda_0\|_1 - (1 - \pi_+)}{2}, \frac{1 - \pi_- - \|\lambda_0\|_1}{2} \right\}$$

and subsequently the step width by

$$\varepsilon_i = \frac{1}{J_i} \min \left\{ \varepsilon/2, \bar{\varepsilon} \right\} , \quad i = 1, 2 .$$

The directional vectors are defined by

$$\xi_c = \left(\sum_{i=1}^{2} (-1)^i \varepsilon_i \mathbb{1}_{c \in \mathcal{J}_i} \right) e_c$$

for the c-th unit vector $e_c \in \mathbb{R}^{m-1}$. Then, the vectors

$$\lambda_k = \lambda_0 + \sum_{c=1}^{k} \xi_c , \quad k = 1, \dots, m-1$$

fulfil the postulated properties as

$$\lambda_k \in \Lambda_m , \quad k = 1, \dots, m-1 , \quad \lambda_{m-1} \in \operatorname{int}(\Lambda_m) , \quad \sum_{k=0}^{m-2} \|\lambda_{k+1} - \lambda_k\|_1 \leq \varepsilon .$$

(ii) First note that for $\lambda_0, \lambda' \in \partial\, \Lambda_m$, (i) gives appropriate finite sequences of line segments within Λ_m from λ_0, λ' to vectors $\bar{\lambda}_0, \bar{\lambda}' \in \operatorname{int}(\Lambda_m)$ with lengths of at most $\varepsilon = \frac{\|\lambda_0 - \lambda'\|_1}{2}$. In particular, we have

$$\|\bar{\lambda}_0 - \bar{\lambda}'\|_1 \leq \|\bar{\lambda}_0 - \lambda_0\|_1 + \|\lambda_0 - \lambda'\|_1 + \|\lambda' - \bar{\lambda}'\|_1 \leq 2\|\lambda_0 - \lambda'\|_1 .$$

This means, it is enough to show that for $\lambda_0, \lambda' \in \operatorname{int}(\Lambda_m)$, there are $\lambda_1, \dots, \lambda_l \in \operatorname{int}(\Lambda_m)$ so that with $\lambda_{l+1} = \lambda'$, we have

$$\lambda_{k+1} - \lambda_k = c_k e_{j_k} , \quad k = 0, \dots, l ,$$

for some unit vectors e_{j_k} and some coefficients $c_k \in \mathbb{R}$ with $\sum_{k=0}^{l} |c_k| \leq (m-1)\|\lambda_0 - \lambda'\|_1$.

Therefore, deduce that there is some $\varepsilon > 0$ so that the closed ε-neighbourhood $B_\varepsilon([\lambda_0, \lambda'])$ of the line segment $[\lambda_0, \lambda']$ is a subset of $\operatorname{int}(\Lambda_m)$. This is because $\operatorname{int}(\Lambda_m)$ is open and convex. Indeed, let $\varepsilon > 0$ so that $B_\varepsilon(\lambda_0), B_\varepsilon(\lambda') \subset \operatorname{int}(\Lambda_m)$. Then, any

$$v \in B_\varepsilon\big([\lambda_0, \lambda']\big) = \left\{ p\lambda_0 + (1-p)\lambda' + z : \ p \in [0,1], z \in \mathbb{R}^{m-1}, \|z\|_\infty \leq \varepsilon \right\}$$

must be an element of the line segment $[\lambda_0 + z, \lambda' + z] \subset \operatorname{int}(\Lambda_m)$ for a proper choice of z by convexity of $\operatorname{int}(\Lambda_m)$. This already proves the result by deducing that one can make appropriate changes of a length smaller than ε in single components of λ_0 without leaving the neighbourhood $B_\varepsilon\big([\lambda_0, \lambda']\big)$ so that after at most $m-1$ steps, the resulting vector again lies in the line segment. □

The following lemma gives the conditions that are easy to verify.

Lemma 5.2.7. *Under Assumptions* **(N1)**, **(N2)** *and* $\widetilde{\textbf{(N3)}}$, *Conditions* **(B1)**, **(B2)**, **(B3)** *and* **(B6)** *are fulfilled. In particular, the constant* $\bar{C}$ *in* **(B1)** *is given by* $\bar{C} = 4m - 1$.

Proof. Prove **(B1)**. The compactness of Θ and I is given by **(N1)**. The compactness of the parameter function sets $\Gamma(\alpha)$ is given by Remark 2.4.5 (ii) when using the Hölder norm $\|\cdot\|_{a/2}$. The fact that the parameter function sets $\Gamma(\alpha)$ are the intersections of all sets $\Gamma(\beta)$ with $\beta < \alpha$ follows from Remark 2.4.5 (iv). Note that Θ fulfils the latter part of **(B1)** for $\bar{C} = 4m - 1$ according to Lemma 5.2.6 (ii) and Remark 3.1.5 (i).

As the mixture density f_{mix} is invariant under relabeling, i.e. permutation of the components, so are the contrast functions M and M_n so that **(B2)** is naturally fulfilled.

Prove **(B3)**. The continuity of M in its first two arguments is obvious. Hence, it is enough to fix $\vartheta \in \overline{\Xi}$, $x \in I$ and examine a sequence $(\gamma_n)_{n\in\mathbb{N}} \subset \Gamma(a)$, $\gamma_n = (\theta_n(\cdot), \ell_n)$ with $\gamma_n \to \gamma = (\theta(\cdot), \ell) \in \Gamma(a)$. Especially, $\|\theta_n(\cdot) - \theta(\cdot)\|_\infty, \|\ell_n - \ell\|_\infty \to 0$. For those sequences, we deduce

$$\begin{aligned} |M(\vartheta, x; \gamma_n) - M(\vartheta, x; \gamma)| &\leq \int |g(y;\vartheta)| \cdot \big|f_{\text{mix}}\big(y; \theta_n(x)\big)\ell_n(x) - f_{\text{mix}}\big(y; \theta(x)\big)\ell(x)\big| \,\mathrm{d}y \\ &\leq \int |g(y;\vartheta)| \cdot \sup_{\theta\in\overline{\Xi}} f_{\text{mix}}(y;\theta)\,\mathrm{d}y \cdot \big\|\ell_n - \ell\big\|_\infty \qquad (5.2.7) \\ &\quad + \int |g(y;\vartheta)| \cdot \big|f_{\text{mix}}\big(y; \theta_n(x)\big) - f_{\text{mix}}\big(y; \theta(x)\big)\big| \,\mathrm{d}y \cdot \|\ell\|_\infty \,. \qquad (5.2.8) \end{aligned}$$

(5.2.7) is directly treated by Lemma 5.2.2 and the fact that $\|\ell_n - \ell\|_\infty \to 0$, whereas by Lemma 5.2.3 (i) and Lemma 5.2.2 once again, we deduce

$$(5.2.8) \leq \int \big|g(y;\vartheta)\big| c_* \phi(y|0, s_*^2)\,\mathrm{d}y \cdot \big\|\theta_n(\cdot) - \theta(\cdot)\|_\infty \cdot \|\ell\|_\infty \to 0 \,.$$

The coordinates of the parameter functions $\theta_*(\cdot)$ are Hölder-α-smooth, especially continuous so that **(B3)** is dealt with.

Condition **(B6)** holds because $\varepsilon^\Delta > 0$ is independent of x, γ and α.

□

Now, we treat conditions **(B4)**, **(B5)**, **(B7)**, **(B8)** and **(B9)** that are a bit more involved.

Lemma 5.2.8. *Let Assumption* **(N1)** *hold,* $0 < a \leq b < \infty$ *and* $J \subset \operatorname{int}(I)$ *be compact. Then (i)* **(B4)** *and (ii)* **(B5)** *hold. That is:*

(i) For all $x \in J$, $\alpha \in [a,b]$, $\gamma \in \Gamma(\alpha)$, *the Hessian matrix*

$$V_x\big(\theta_*(x);\gamma\big) = \partial^2_{\theta^2} M\big(\theta_*(x), x; \gamma\big)$$

is positive definite. Especially, the eigenvalues $\lambda^1_{x,\gamma} \geq \lambda^2_{x,\gamma} \geq \lambda^3_{x,\gamma}$ *of* $V_x(\theta_*(x);\gamma)$ *are positive.*

(ii) The Hessian matrices $V_x(\theta;\gamma)$ *are uniformly Lipschitz continuous in* θ, *i.e. for all* $\theta, \theta' \in \overline{\Xi}$, *we have*

$$\sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \sup_{x\in J} \big\|V_x(\theta;\gamma) - V_x(\theta';\gamma)\big\| \leq L_{\mathrm{Hess}} \|\theta - \theta'\|_1,$$

where L_{Hess} *depends only on* $\overline{\Xi}$, I *and* $\Gamma(a)$.

Proof of lemma 5.2.8. (i) According to Lemma 5.2.2, integration under the integral sign gives

$$\begin{aligned}
&\partial^2_{\theta^2} M\big(\theta_*(x), x; \gamma\big) \\
&= \mathbb{E}_\gamma\left[\frac{\partial^2_{\theta^2} f_{\mathrm{mix}}\big(Y;\theta_*(x)\big)}{f_{\mathrm{mix}}\big(Y;\theta_*(x)\big)} - \frac{\big(\partial_\theta f_{\mathrm{mix}}\big(Y;\theta_*(x)\big)\big)\big(\partial_\theta f_{\mathrm{mix}}\big(Y;\theta_*(x)\big)\big)^T}{f_{\mathrm{mix}}\big(Y;\theta_*(x)\big)^2}\middle| X = x\right]\cdot \ell(x) \\
&= \left(\int \frac{\partial^2_{\theta^2} f_{\mathrm{mix}}\big(y;\theta_*(x)\big)}{f_{\mathrm{mix}}\big(y;\theta_*(x)\big)} f_{\mathrm{mix}}\big(y;\theta_*(x)\big)\,\mathrm{d}y \right. \\
&\quad \left. - \mathbb{E}_\gamma\left[\frac{\big(\partial_\theta f_{\mathrm{mix}}\big(Y;\theta_*(x)\big)\big)\big(\partial_\theta f_{\mathrm{mix}}\big(Y;\theta_*(x)\big)\big)^T}{f_{\mathrm{mix}}\big(Y;\theta_*(x)\big)^2}\middle| X = x\right]\right)\cdot \ell(x) \\
&= \left(\partial^2_{\theta^2}\int f_{\mathrm{mix}}\big(y;\theta_*(x)\big)\,\mathrm{d}y - \mathbb{E}_\gamma\left[\frac{\big(\partial_\theta f_{\mathrm{mix}}\big(Y;\theta_*(x)\big)\big)\big(\partial_\theta f_{\mathrm{mix}}\big(Y;\theta_*(x)\big)\big)^T}{f_{\mathrm{mix}}\big(Y;\theta_*(x)\big)^2}\middle| X = x\right]\right)\cdot \ell(x) \\
&= -\mathbb{E}_\gamma\left[\frac{\big(\partial_\theta f_{\mathrm{mix}}\big(Y;\theta_*(x)\big)\big)\big(\partial_\theta f_{\mathrm{mix}}\big(Y;\theta_*(x)\big)\big)^T}{f_{\mathrm{mix}}\big(Y;\theta_*(x)\big)^2}\middle| X = x\right]\cdot \ell(x)\,.
\end{aligned}$$

Now assume there is a $v = (v_1, \ldots, v_{3m-1}) \in \mathbb{R}^{3m-1}$ so that $v^T \partial^2_{\theta^2} M\big(\theta_*(x), x; \gamma\big) v = 0$, which is equivalent to

$$\begin{aligned}
0 &= \mathbb{E}_\gamma\left[v^T \frac{\big(\partial_\theta f_{\mathrm{mix}}\big(Y;\theta_*(x)\big)\big)\big(\partial_\theta f_{\mathrm{mix}}\big(Y;\theta_*(x)\big)\big)^T}{f_{\mathrm{mix}}\big(Y;\theta_*(x)\big)^2} v \middle| X = x\right] \\
&= \mathbb{E}_\gamma\left[\left(\frac{\partial_\theta f_{\mathrm{mix}}\big(Y;\theta_*(x)\big)}{f_{\mathrm{mix}}\big(Y;\theta_*(x)\big)} v\right)^2 \middle| X = x\right]
\end{aligned}$$

implying that

$$\begin{aligned} 0 =& \mathbb{E}_\gamma \left[\left| \left(\frac{\partial_\theta f_{\text{mix}}\big(Y; \theta_*(x)\big)}{f_{\text{mix}}\big(Y; \theta_*(x)\big)} \right)^T v \right| \,\middle|\, X = x \right] \\ =& \int \left| \big(\partial_\theta f_{\text{mix}}\big(y; \theta_*(x)\big)\big)^T v \right| \mathrm{d}y \ . \end{aligned}$$

As the integrand is non-negative and continuous, we get that

$$\partial_\theta f_{\text{mix}}\big(y; \theta_*(x)\big)^T v = 0 \ , \qquad \forall y \in \mathbb{R} \ . \tag{5.2.9}$$

Let p_1, p_2 be any polynomials and $\sigma_c \neq \sigma_{c'}$, then

$$\begin{aligned} \frac{p_1(y)}{p_2(y)} \frac{\phi(y|\mu_c, \sigma_c^2)}{\phi(y|\mu_{c'}, \sigma_{c'}^2)} =& \frac{p_1(y)}{p_2(y)} \exp\left(\frac{(\sigma_c^2 - \sigma_{c'}^2)y^2 + 2(\sigma_{c'}^2 \mu_c - \sigma_c^2 \mu_{c'})y + \sigma_c^2 \mu_{c'}^2 - \sigma_{c'}^2 \mu_c^2}{\sigma_c^2 \sigma_{c'}^2} \right) \\ \xrightarrow[y \to \infty]{} & \begin{cases} \infty \ , & \sigma_c^2 > \sigma_{c'}^2 \\ 0 \ , & \sigma_c^2 < \sigma_{c'}^2 \end{cases} \ . \end{aligned} \tag{5.2.10}$$

Now, if p_1, p_2 are any polynomials and $\sigma_c = \sigma_{c'}$ as well as $\mu_c \neq \mu_{c'}$, then

$$\begin{aligned} \frac{p_1(y)}{p_2(y)} \frac{\phi(y|\mu_c, \sigma_c^2)}{\phi(y|\mu_{c'}, \sigma_c^2)} =& \frac{p_1(y)}{p_2(y)} \exp\left(\frac{2(\mu_c - \mu_{c'})y + \mu_{c'}^2 - \mu_c^2}{\sigma_c^2} \right) \\ \xrightarrow[y \to \infty]{} & \begin{cases} \infty \ , & \mu_c > \mu_{c'} \\ 0 \ , & \mu_c < \mu_{c'} \end{cases} \ . \end{aligned} \tag{5.2.11}$$

We give an iterative procedure to show $v = 0$. For notational simplicity insert a 0 between the $(m-1)$-th and m-th component of v. Let $\mathcal{C} \subset \{1, \dots, m\}$. If $\max_{c \in \mathcal{C}} \sigma_c^2$ is unique, define $\tilde{c} = \operatorname{argmax}_{c \in \mathcal{C}} \sigma_c^2$, otherwise define $\tilde{c} = \operatorname{argmax}_{c \in \mathcal{C}} \mu_c$, which is then unique by Assumption **(N1)**.

By (5.2.9), for all $y \in \mathbb{R}$, we have

$$\begin{aligned} &\sum_{c \in \mathcal{C} \setminus \{\tilde{c}\}} v_c \partial_{\pi_c} f_{\text{mix}}\big(y; \theta_*(x)\big) + v_{m+c} \partial_{\mu_c} f_{\text{mix}}\big(y; \theta_*(x)\big) + v_{2m+c} \partial_{\sigma_c} f_{\text{mix}}\big(y; \theta_*(x)\big) \\ =& - \Big(v_{\tilde{c}} \partial_{\pi_{\tilde{c}}} f_{\text{mix}}\big(y; \theta_*(x)\big) + v_{m+\tilde{c}} \partial_{\mu_{\tilde{c}}} f_{\text{mix}}\big(y; \theta_*(x)\big) + v_{2m+\tilde{c}} \partial_{\sigma_{\tilde{c}}} f_{\text{mix}}\big(y; \theta_*(x)\big) \Big) \ . \end{aligned}$$

As

$$\frac{\partial_{\pi_{\tilde{c}}} f_{\text{mix}}\big(y; \theta_*(x)\big)}{\partial_{\sigma_{\tilde{c}}} f_{\text{mix}}\big(y; \theta_*(x)\big)} \xrightarrow[y \to \infty]{} 0 \ , \quad \frac{\partial_{\mu_{\tilde{c}}} f_{\text{mix}}\big(y; \theta_*(x)\big)}{\partial_{\sigma_{\tilde{c}}} f_{\text{mix}}\big(y; \theta_*(x)\big)} \xrightarrow[y \to \infty]{} 0 \ , \quad \frac{\partial_{\pi_{\tilde{c}}} f_{\text{mix}}\big(y; \theta_*(x)\big)}{\partial_{\mu_{\tilde{c}}} f_{\text{mix}}\big(y; \theta_*(x)\big)} \xrightarrow[y \to \infty]{} 0 \ ,$$

$\pi_c(x) > 0$ for all c, x and because of (5.2.10) and (5.2.11), we get that

$$0 = \lim_{y \to \infty} \left\{ \sum_{c \in \mathcal{C} \setminus \{\tilde{c}\}} v_c \frac{\partial_{\pi_c} f_{\text{mix}}\big(y; \theta_*(x)\big)}{\partial_{\sigma_{\tilde{c}}} f_{\text{mix}}\big(y; \theta_*(x)\big)} + v_{m+c} \frac{\partial_{\mu_c} f_{\text{mix}}\big(y; \theta_*(x)\big)}{\partial_{\sigma_{\tilde{c}}} f_{\text{mix}}\big(y; \theta_*(x)\big)} + v_{2m+c} \frac{\partial_{\sigma_c} f_{\text{mix}}\big(y; \theta_*(x)\big)}{\partial_{\sigma_{\tilde{c}}} f_{\text{mix}}\big(y; \theta_*(x)\big)} \right.$$

$$+ v_{\tilde{c}} \frac{\partial_{\pi_{\tilde{c}}} f_{\mathrm{mix}}\big(y;\theta_*(x)\big)}{\partial_{\sigma_{\tilde{c}}} f_{\mathrm{mix}}\big(y;\theta_*(x)\big)} + v_{m+\tilde{c}} \frac{\partial_{\mu_{\tilde{c}}} f_{\mathrm{mix}}\big(y;\theta_*(x)\big)}{\partial_{\sigma_{\tilde{c}}} f_{\mathrm{mix}}\big(y;\theta_*(x)\big)} \Bigg\} = -v_{2m+\tilde{c}}$$

and subsequently

$$\begin{aligned}
0 &= \lim_{y\to\infty} \Bigg\{ \sum_{c\in\mathcal{C}\setminus\{\tilde{c}\}} v_c \frac{\partial_{\pi_c} f_{\mathrm{mix}}\big(y;\theta_*(x)\big)}{\partial_{\mu_{\tilde{c}}} f_{\mathrm{mix}}\big(y;\theta_*(x)\big)} + v_{m+c} \frac{\partial_{\mu_c} f_{\mathrm{mix}}\big(y;\theta_*(x)\big)}{\partial_{\mu_{\tilde{c}}} f_{\mathrm{mix}}\big(y;\theta_*(x)\big)} + v_{2m+c} \frac{\partial_{\mu_c} f_{\mathrm{mix}}\big(y;\theta_*(x)\big)}{\partial_{\sigma_{\tilde{c}}} f_{\mathrm{mix}}\big(y;\theta_*(x)\big)} \Bigg\} \\
&\quad + v_{\tilde{c}} \frac{\partial_{\pi_{\tilde{c}}} f_{\mathrm{mix}}\big(y;\theta_*(x)\big)}{\partial_{\mu_{\tilde{c}}} f_{\mathrm{mix}}\big(y;\theta_*(x)\big)} = -v_{m+\tilde{c}} \,, \\
0 &= \lim_{y\to\infty} \Bigg\{ \sum_{c\in\mathcal{C}\setminus\{\tilde{c}\}} v_c \frac{\partial_{\pi_c} f_{\mathrm{mix}}\big(y;\theta_*(x)\big)}{\partial_{\pi_{\tilde{c}}} f_{\mathrm{mix}}\big(y;\theta_*(x)\big)} + v_{m+c} \frac{\partial_{\mu_c} f_{\mathrm{mix}}\big(y;\theta_*(x)\big)}{\partial_{\pi_{\tilde{c}}} f_{\mathrm{mix}}\big(y;\theta_*(x)\big)} + v_{2m+c} \frac{\partial_{\mu_c} f_{\mathrm{mix}}\big(y;\theta_*(x)\big)}{\partial_{\pi_{\tilde{c}}} f_{\mathrm{mix}}\big(y;\theta_*(x)\big)} \Bigg\} \\
&= - v_{\tilde{c}} \,.
\end{aligned}$$

(ii) In order to prove Lipschitz continuity of the Hessian matrix, repeat the arguments used in the proof of Lemma 5.2.7, especially the proof of **(B3)** and use Lemma 5.2.3. According to Lemma 5.2.3 (iii), there is a function $F(\cdot, \overline{\Xi})$ that is integrable with respect to $\sup_{\theta\in\overline{\Xi}} f_{\mathrm{mix}}(y;\theta)$ so that

$$\begin{aligned}
&\sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \sup_{x\in J} \|V_x(\theta;\gamma) - V_x(\theta';\gamma)\| \\
&\leq \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \sup_{x\in J} \mathbb{E}_\gamma\Bigg[\Big\| \partial^2_{\theta^2} g(Y;\theta) - \partial^2_{\theta^2} g(Y;\tilde{\theta}) \Big\| \Big| X = x \Bigg] \ell(x) \\
&\leq \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \sup_{x\in J} \int F(y;\overline{\Xi}) f_{\mathrm{mix}}\big(y;\theta_*(x)\big)\,\mathrm{d}y \cdot \|\theta - \tilde{\theta}\| \ell(x) \\
&\leq \int F(y;\overline{\Xi}) \sup_{\theta\in\overline{\Xi}} f_{\mathrm{mix}}(y;\theta)\,\mathrm{d}y \cdot \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \|\ell\|_\infty \cdot \|\theta - \tilde{\theta}\| \\
&\lesssim \|\theta - \tilde{\theta}\| \,.
\end{aligned}$$

□

The following lemma shows that the deterministic and stochastic estimation errors of the empirical contrasts' gradients are of the usual non-parametric order.

Lemma 5.2.9. *Let $0 < a \leq b < \infty$. Under Assumption* **(N1)**, *for some kernel K fulfilling Assumptions* **(N2)** *and* **($\widetilde{\mathbf{N}}$3)** *and sequences of bandwidth parameters $(h_n(\alpha))_{n\in\mathbb{N}}$, $\alpha \in [a,b]$ so that*

$$\sup_{\alpha\in[a,b]} h_n(\alpha), \quad \sup_{\alpha\in[a,b]} \frac{\log n}{n h_n(\alpha)^d} \to 0 \,,$$

Conditions **(B7)** *and* **(B9)** *hold for any compact cuboid $J \subset \mathrm{int}(I)$ containing an open subset. To be specific on* **(B7)**, *for any compact cuboid $J \subset \mathrm{int}(I)$ containing an open subset, we have that*

$$\limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \sup_{x\in J,\theta\in\Theta} h_n(\alpha)^{-\alpha} \big\| \mathbb{E}_\gamma\big[S_n\big(\theta,x;h_n(\alpha)\big)\big] - S(\theta,x;\gamma) \big\| \leq C_* \,, \tag{5.2.12}$$

$$\limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(a)} \left(\frac{\log n}{n h_n(\alpha)^d}\right)^{-1} \mathbb{E}_\gamma\left[\sup_{x\in J,\theta\in\Theta} \left\|S_n\big(\theta,x;h_n(\alpha)\big) - \mathbb{E}_\gamma\big[S_n\big(\theta,x;h_n(\alpha)\big)\big]\right\|_1^2\right]$$
$$\leq C_{\mathrm{STOCH}} \,. \tag{5.2.13}$$

The constant $C_* > 0$ depends only on a, b, $\Gamma(a)$, Θ, I and K; the constant $C_{\mathrm{STOCH}} > 0$ depends only on $\Gamma(a)$, $\|K\|_\infty$, L_K, I, Θ. Particularly, for $h_n(\alpha) = \big(\frac{\log n}{n}\big)^{\frac{1}{2\alpha+d}}$, we have

$$\limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \left(\frac{\log n}{n}\right)^{-\frac{2\alpha}{2\alpha+d}} \mathbb{E}_\gamma\left[\sup_{x\in J,\theta\in\Theta} \left\|S_n\big(\theta,x;h_n(\alpha)\big) - S(\theta,x;\gamma)\right\|^2\right] < \infty \,.$$

Note that the assertion on the stochastic error (5.2.13) is stated uniformly over all $\gamma \in \Gamma(a)$. The respective degrees of smoothness of the true parameter curves only play a role in determining the convergence rates of the bias terms.

Proof of Lemma 5.2.9. By using the shape of the log likelihood's derivatives given in Remark 5.2.1 and differentiating under the integral sign, which is allowed according to Lemma 5.2.2, we get that for any $\theta \in \Theta$, $x \in J$, $\gamma \in \Gamma(a)$, $h \in (0,\infty)$, the gradients $S_n(\theta,x;h) = \partial_\theta M_n(\theta,x;h)$ and $S(\theta,x;\gamma) = \partial_\theta M(\theta,x;\gamma)$ have the components

$$\begin{aligned}
\partial_{\pi_{\tilde c}} M_n(\theta,x;h) &= \frac{1}{n}\sum_{k=1}^n \frac{\phi(Y_k|\mu_{\tilde c},\sigma_{\tilde c}^2) - \phi(Y_k|\mu_m,\sigma_m^2)}{\sum_{\nu=1}^m \pi_\nu \phi(Y_k|\mu_\nu,\sigma_\nu^2)} K_h(X_k - x) \,,\\
\partial_{\mu_c} M_n(\theta,x;h) &= \frac{1}{n}\sum_{k=1}^n \frac{\pi_c(\mu_c - Y_k)\sigma_c^{-2}\phi(Y_k|\mu_c,\sigma_c^2)}{\sum_{\nu=1}^m \pi_\nu \phi(Y_k|\mu_\nu,\sigma_\nu)} K_h(X_k - x) \,,\\
\partial_{\sigma_c} M_n(\theta,x;h) &= \frac{1}{n}\sum_{k=1}^n \frac{\pi_c\big((Y_k-\mu_c)^2 - \sigma_c^2\big)\sigma_c^{-3}\phi(Y_k|\mu_c,\sigma_c^2)}{\sum_{\nu=1}^m \pi_\nu \phi(Y_k|\mu_\nu,\sigma_\nu^2)} K_h(X_k - x) \,,\\
\partial_{\pi_{\tilde c}} M(\theta,x;\gamma) &= \mathbb{E}_\gamma\left[\frac{\phi(Y|\mu_{\tilde c},\sigma_{\tilde c}^2) - \phi(Y|\mu_m,\sigma_m^2)}{\sum_{\nu=1}^m \pi_\nu \phi(Y|\mu_\nu,\sigma_\nu^2)}\middle| X = x\right]\ell(x)\\
\partial_{\mu_c} M(\theta,x;\gamma) &= \mathbb{E}_\gamma\left[\frac{\pi_c(\mu_c - Y)\sigma_c^{-2}\phi(Y|\mu_c,\sigma_c^2)}{\sum_{\nu=1}^m \pi_\nu \phi(Y|\mu_\nu,\sigma_\nu^2)}\middle| X = x\right]\ell(x) \,,\\
\partial_{\sigma_c} M(\theta,x;\gamma) &= \mathbb{E}_\gamma\left[\frac{\pi_c\big((Y-\mu_c)^2 - \sigma_c^2\big)\sigma_c^{-3}\phi(Y|\mu_c,\sigma_c^2)}{\sum_{\nu=1}^m \pi_\nu \phi(Y|\mu_\nu,\sigma_\nu^2)}\middle| X = x\right]\ell(x) \,,
\end{aligned}$$

where $c \in \{1,\dots,m\}$, $\tilde c \in \{1,\dots,m-1\}$.

Prove (5.2.12). Note that the following calculations are independent of θ, x, γ and α. Write $h_n(\alpha) = h$. For some $t \in \{\pi_1,\dots,\pi_{m-1},\mu_1,\dots,\mu_m,\sigma_1,\dots,\sigma_m\}$, we have

$$
\begin{aligned}
&\mathbb{E}_\gamma\big[\partial_t M_n(\theta,x;h)\big] - \partial_t M(\theta,x;\gamma)\\
=\;& \int \frac{\partial_t f_{\mathrm{mix}}(y;\theta)}{f_{\mathrm{mix}}(y;\theta)} \cdot \Big(K_h * f_{\mathrm{mix}}\big(y;\theta_*(\cdot)\big)\ell\Big)(x)\,\mathrm{d}y - \int \frac{\partial_t f_{\mathrm{mix}}(y;\theta)}{f_{\mathrm{mix}}(y;\theta)} f_{\mathrm{mix}}\big(y;\theta_*(x)\big)\ell(x)\,\mathrm{d}y\\
=\;& \int \frac{\partial_t f_{\mathrm{mix}}(y;\theta)}{f_{\mathrm{mix}}(y;\theta)} \cdot \bigg\{ \Big(K_h * f_{\mathrm{mix}}\big(y;\theta_*(\cdot)\big)\ell\Big)(x) - \big(K_h*\ell\big)(x) f_{\mathrm{mix}}\big(y;\theta_*(x)\big)\\
&\qquad + \Big[\big(K_h*\ell\big)(x) - \ell(x)\Big] f_{\mathrm{mix}}\big(y;\theta_*(x)\big)\bigg\}\,\mathrm{d}y\,.
\end{aligned}
$$

Now,

$$
\frac{\partial_t f_{\mathrm{mix}}(y;\theta)}{f_{\mathrm{mix}}(y;\theta)} \sup_{\theta\in\Theta} f_{\mathrm{mix}}(y;\theta)
$$

is integrable according to Lemma 5.2.2 and we have

$$
\big|\big(K_h*\ell\big)(x) - \ell(x)\big| \lesssim h^\alpha\,.
$$

Thus, according to Lemma 3.2.1, it suffices to show that

$$
\begin{aligned}
&\Big|\Big(K_h * f_{\mathrm{mix}}\big(y;\theta_*(\cdot)\big)\ell\Big)(x) - \big(K_h*\ell\big)(x) f_{\mathrm{mix}}\big(y;\theta_*(x)\big)\Big|\\
=&\left|\int K_h(z)\ell(z+x)\Big(f_{\mathrm{mix}}\big(y;\theta_*(z+x)\big) - f_{\mathrm{mix}}\big(y;\theta_*(x)\big)\Big)\,\mathrm{d}z\right|\\
=&\int |K(z)|\ell(hz+x)\big|f_{\mathrm{mix}}\big(y;\theta_*(hz+x)\big) - f_{\mathrm{mix}}\big(y;\theta_*(x)\big)\big|\,\mathrm{d}z\\
\le& c_*\phi(y|0,s_*^2)\int |K(z)|\ell(hz+x)\\
&\sum_{c=1}^m \big\{\big\|\pi_c(hz+x)-\pi_c(x)\big\| + \big\|\sigma_c(hz+x)-\sigma_c(x)\big\| + \big\|\mu_c(hz+x)-\mu_c(x)\big\|\big\}\,\mathrm{d}z\\
\lesssim& \phi(y|0,s_*^2)\,h^\alpha\,,
\end{aligned}
$$

which is true according to Lemma 5.2.3 (i) and Lemma 3.2.1 once again because the parameter functions are Hölder-α-smooth. Now, Lemma 5.2.2 gives integrability of

$$
\frac{\partial_t f_{\mathrm{mix}}(y;\theta)}{f_{\mathrm{mix}}(y;\theta)}\,\phi(y|0,s_*^2)\,,
$$

concluding the proof of (5.2.12).

In order to prove (5.2.13) and **(B9)**, we only need to show that all assumptions of Theorem 3.2.5 are met for $\rho = 2$, because then (5.2.13) holds and **(B9)** is given by Lemma 3.2.9.

According to the model assumptions, it suffices to show the integrability and exponential deviation requirements for $\rho = 2$. (3.2.12) is given by Lemma 5.2.2; (3.2.13), (3.2.14)

by Lemma 5.2.3 (iii); and the exponential deviation inequality is given by Lemma 5.2.5. Hence, Theorem 3.2.5 gives for every $t \in \{\pi_1, \dots, \pi_{m-1}, \sigma_1, \dots, \sigma_m, \mu_1, \dots, \mu_m\}$ some constant $C_t < \infty$ so that

$$\limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(a)} \left(\frac{\log n}{n h_n(\alpha)^d}\right)^{-1} \mathbb{E}_\gamma\left[\sup_{x\in J}\sup_{\theta\in\Theta}\left|\partial_t M_n\big(\theta,x;h_n(\alpha)\big) - \mathbb{E}_\gamma\big[\partial_t M_n\big(\theta,x;h_n(\alpha)\big)\big]\right|^2\right] \le C_t < \infty \,.$$

Finally, (5.2.13) is given by the equivalence of norms on $\mathbb{R}^{3m-1}$, the triangular inequality and maximizing over t.

□

Lemma 5.2.10. *Let $0 < a \le b < \infty$ and $(h_n(\alpha))_{n\in\mathbb{N}}$ for $\alpha \in [a,b]$ be sequences of bandwidth parameters so that*

$$\sup_{\alpha\in[a,b]} h_n(\alpha), \quad \sup_{\alpha\in[a,b]} \frac{\log n}{n h_n(\alpha)^d} \to 0 \,.$$

Under Assumption **(N1)**, *for some kernel K fulfilling Assumptions* **(N2)** *and* **($\widetilde{\text{N3}}$)**, *Condition* **(B8)** *holds for any compact cuboid $J \subset \operatorname{int}(I)$ containing an open subset, i.e. the empirical contrast function M_n is uniformly consistent for the asymptotic contrast M. To be specific, for any compact $J \subset \operatorname{int}(I)$, and for all $\varepsilon > 0$, we have*

$$\sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(a)} \mathbb{P}_\gamma\left(\sup_{x\in J,\theta\in\Theta}\left|M_n(\theta,x;h_n(\alpha)) - M(\theta,x;\gamma)\right| \ge \varepsilon\right) = o(1) \,.$$

For the proof, one would typically want to decompose the estimation error into a stochastic and a non-stochastic term. However, note that we cannot use Theorem 3.2.5 for the stochastic error directly because (3.2.13) does not seem to be fulfilled. Luckily, (3.2.15) is fulfilled so that we can work around that.

Proof of Lemma 5.2.10. Fix any $\tilde\theta \in \Theta$ throughout the whole proof and define the function $T_n(\theta,x;h;\gamma) = M_n(\theta,x;h) - M(\theta,x;\gamma)$. For any $\theta \in \Theta$, **(B1)**, Lemma 5.2.6 and Remark 3.1.5 (i) give an $l_\theta \in \mathbb{N}_0$ and $\bar\theta_1(\theta), \dots, \bar\theta_{l_\theta}(\theta) \in \Theta$ so that with $\theta = \bar\theta_0(\theta)$, $\tilde\theta = \bar\theta_{l_\theta+1}(\theta)$, we have

$$\bar\theta_{k+1}(\theta) - \bar\theta_k(\theta) = c_k e_{j_k} \,, \quad k = 0, \dots, l_\Theta$$

for some unit vectors e_{j_k} and some coefficients $c_k \in \mathbb{R}$ with

$$\sum_{k=0}^{l_\theta} |c_k| \le (4m-1)\|\theta - \tilde\theta\| \le (4m-1)\operatorname{diam}(\Theta) \,.$$

Write

$$\bar\theta_k(\theta) = \big(\vartheta_1^{k,\theta}, \dots, \vartheta_{3m-1}^{k,\theta}\big)^T \,, \quad k = 0, \dots, l_\theta + 1 \,.$$

According to the fundamental theorem of calculus, using the notation $t = (t_1, \dots, t_{3m-1})$, we may write for any n, α, γ,

$$
\begin{aligned}
&\sup_{\theta\in\Theta, x\in J} \left|T_n\big(\theta, x; h_n(\alpha); \gamma\big)\right| \\
&\le \sup_{x\in J} \left|T_n\big(\tilde\theta, x; h_n(\alpha); \gamma\big)\right| \\
&\quad + \sup_{\theta\in\Theta, x\in J} \sum_{k=0}^{l_\theta} \left|T_n\big(\bar\theta_{k+1}(\theta), x; h_n(\alpha); \gamma\big) - T_n\big(\bar\theta_k(\theta), x; h_n(\alpha); \gamma\big)\right| \\
&= \sup_{x\in J} \left|T_n\big(\tilde\theta, x; h_n(\alpha); \gamma\big)\right| + \sup_{\theta\in\Theta, x\in J} \sum_{k=0}^{l_\theta} \left| \int_{\vartheta_{j_k}^{k,\theta}}^{\vartheta_{j_k}^{k+1,\theta}} \partial_{t_j} T_n\big(t, x; h_n(\alpha); \gamma\big)\, \mathrm{d}t_j \right| \\
&\le \sup_{x\in J} \left|T_n\big(\tilde\theta, x; h_n(\alpha); \gamma\big)\right| + (4m-1)\operatorname{diam}(\Theta) \sup_{\theta\in\Theta, x\in J, j=1,\dots,3m-1} \left|\partial_{\vartheta_j} T_n\big(\theta, x; h_n(\alpha); \gamma\big)\right| \\
&= \sup_{x\in J} \left|M_n\big(\tilde\theta, x; h_n(\alpha)\big) - M(\tilde\theta, x; \gamma)\right| \qquad (5.2.14) \\
&\quad + (4m-1)\operatorname{diam}(\Theta) \sup_{\theta\in\Theta, x\in J} \left\|S_n\big(\theta, x; h_n(\alpha)\big) - S(\theta, x; \gamma)\right\|_\infty . \qquad (5.2.15)
\end{aligned}
$$

Make a bias variance decomposition for both (5.2.14) and (5.2.15). For the bias term of (5.2.15), making use of the calculations proving (5.2.12) and Lemma 3.2.1 for the function class $H(a, L, U)$, we deduce

$$
\sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(a)} h_n(\alpha)^{-a} \sup_{\theta\in\Theta, x\in J} \left\|\mathbb{E}_\gamma\big[S_n\big(\theta, x; h_n(\alpha)\big)\big] - S(\theta, x; \gamma)\right\|_\infty = O(1) ,
$$

so that according to $\inf_{\alpha\in[a,b]} h_n(\alpha)^{-a} \to \infty$, we have

$$
\sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(a)} \sup_{\theta\in\Theta, x\in J} \left\|\mathbb{E}_\gamma\big[S_n\big(\theta, x; h_n(\alpha)\big)\big] - S(\theta, x; \gamma)\right\|_\infty = o(1) .
$$

The variance term of (5.2.15) is directly dealt with by (5.2.13) because according to Markov's inequality and Jensen's inequality, we have for any $\varepsilon > 0$

$$
\begin{aligned}
&\sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(a)} \left(\frac{\log n}{n h_n(\alpha)^d}\right)^{-\frac12} \mathbb{P}_\gamma\left(\sup_{x\in J, \theta\in\Theta} \left\|S_n\big(\theta, x; h_n(\alpha)\big) - \mathbb{E}_\gamma\big[S_n\big(\theta, x; h_n(\alpha)\big)\big]\right\|_1 \ge \varepsilon \right) \\
&\le \frac{1}{\varepsilon} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(a)} \left(\frac{\log n}{n h_n(\alpha)^d}\right)^{-\frac12} \mathbb{E}_\gamma\left[\sup_{x\in J, \theta\in\Theta} \left\|S_n\big(\theta, x; h_n(\alpha)\big) - \mathbb{E}_\gamma\big[S_n\big(\theta, x; h_n(\alpha)\big)\big]\right\|_1 \right] \\
&\le \frac{1}{\varepsilon} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(a)} \left(\frac{\log n}{n h_n(\alpha)^d}\right)^{-\frac12} \left\{ \mathbb{E}_\gamma\left[\sup_{x\in J, \theta\in\Theta} \left\|S_n\big(\theta, x; h_n(\alpha)\big) - \mathbb{E}_\gamma\big[S_n\big(\theta, x; h_n(\alpha)\big)\big]\right\|_1^2 \right] \right\}^{\frac12} \\
&= O(1)
\end{aligned}
$$

so that $\inf_{\alpha\in[a,b]} \left(\frac{\log n}{n h_n(\alpha)^d}\right)^{-\frac12} \to \infty$ leads to the conclusion

$$
\sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(a)} \mathbb{P}_\gamma\left(\sup_{x\in J, \theta\in\Theta} \left\|S_n\big(\theta, x; h_n(\alpha)\big) - \mathbb{E}_\gamma\big[S_n\big(\theta, x; h_n(\alpha)\big)\big]\right\|_1 \ge \varepsilon \right) = o(1) .
$$

Thus, it remains to show that (5.2.14) is $o_{\mathbb{P}_\gamma}(1)$ uniformly over all $\alpha \in [a,b]$, $\gamma \in \Gamma(a)$. Therefore, let us decompose

$$\begin{aligned}
&\sup_{x\in J}\Big|M_n\big(\tilde\theta,x;h_n(\alpha)\big)-M(\tilde\theta,x;\gamma)\Big|\\
&\le \sup_{x\in J}\Big|M_n\big(\tilde\theta,x;h_n(\alpha)\big)-\mathbb{E}_\gamma\big[M_n\big(\tilde\theta,x;h_n(\alpha)\big)\big]\Big|+\sup_{x\in J}\Big|\mathbb{E}_\gamma\big[M_n\big(\tilde\theta,x;h_n(\alpha)\big)\big]-M(\tilde\theta,x;\gamma)\Big|\,.
\end{aligned} \tag{5.2.16}$$

For the second summand in (5.2.16), deduce that for any α, γ, x, we have

$$\begin{aligned}
&\Big|\mathbb{E}_\gamma\big[M_n\big(\tilde\theta,x;h_n(\alpha)\big)\big]-M(\tilde\theta,x;\gamma)\Big|\\
&=\bigg|\int g(y;\tilde\theta)\cdot\Big(K_{h_n(\alpha)}*f_{\mathrm{mix}}\big(y;\theta_*(\cdot)\big)\ell\Big)(x)\,\mathrm{d}y-\int g(y;\tilde\theta)f_{\mathrm{mix}}\big(y;\theta_*(x)\big)\ell(x)\,\mathrm{d}y\bigg|\\
&\le\int|g(y;\tilde\theta)|\cdot\bigg|\Big(K_{h_n(\alpha)}*f_{\mathrm{mix}}\big(y;\theta_*(\cdot)\big)\ell\Big)(x)-\Big(K_{h_n(\alpha)}*\ell\Big)(x)f_{\mathrm{mix}}\big(y;\theta_*(x)\big)\bigg|\,\mathrm{d}y\\
&\quad+\int f_{\mathrm{mix}}\big(y;\theta_*(x)\big)\Big|\Big(K_{h_n(\alpha)}*\ell\Big)(x)-\ell(x)\Big|\,\mathrm{d}y
\end{aligned}$$

so that we once again deduce from Lemma 3.2.1 for the function class $H(a,L,U)$ and Lemma 5.2.3 (i) that

$$\sup_{\alpha\in[a,b]}\sup_{\gamma\in\Gamma(a)}h_n(\alpha)^{-a}\sup_{x\in J}\Big|\mathbb{E}_\gamma\big[M_n\big(\tilde\theta,x;h_n(\alpha)\big)\big]-M(\tilde\theta,x;\gamma)\Big|=O(1)$$

and by $\inf_{\alpha\in[a,b]}h_n(\alpha)^{-a}\to\infty$, we derive

$$\sup_{\alpha\in[a,b]}\sup_{\gamma\in\Gamma(a)}\sup_{x\in J}\Big|\mathbb{E}_\gamma\big[M_n\big(\tilde\theta,x;h_n(\alpha)\big)\big]-M(\tilde\theta,x;\gamma)\Big|=o(1)\,.$$

Treat the first term in (5.2.16) by Theorem 3.2.5 with $\Theta=\{\tilde\theta\}$. (3.2.13) is naturally fulfilled. Once we proved (3.2.15), Theorem 3.2.5 gives

$$\sup_{\alpha\in[a,b]}\sup_{\gamma\in\Gamma(a)}\bigg(\frac{\log n}{nh_n(\alpha)^d}\bigg)^{-\frac12}\mathbb{E}_\gamma\bigg[\sup_{x\in J}\Big|M_n\big(\tilde\theta,x;h_n(\alpha)\big)-\mathbb{E}_\gamma\big[M_n\big(\tilde\theta,x;h_n(\alpha)\big)\big]\Big|\bigg]=O(1)\,.$$

Since $\inf_{\alpha\in[a,b]}\big(\frac{\log n}{nh_n(\alpha)^d}\big)^{-\frac12}\to\infty$, Markov's inequality gives for any $\varepsilon>0$

$$\begin{aligned}
&\sup_{\alpha\in[a,b]}\sup_{\gamma\in\Gamma(a)}\mathbb{P}_\gamma\bigg(\sup_{x\in J}\Big|M_n\big(\tilde\theta,x;h_n(\alpha)\big)-\mathbb{E}_\gamma\big[M_n\big(\tilde\theta,x;h_n(\alpha)\big)\big]\Big|\ge\varepsilon\bigg)\\
&\le\frac{1}{\varepsilon}\sup_{\alpha\in[a,b]}\sup_{\gamma\in\Gamma(a)}\mathbb{E}_\gamma\bigg[\sup_{x\in J}\Big|M_n\big(\tilde\theta,x;h_n(\alpha)\big)-\mathbb{E}_\gamma\big[M_n\big(\tilde\theta,x;h_n(\alpha)\big)\big]\Big|\bigg]\\
&=o(1)\,.
\end{aligned}$$

In order to prove (3.2.15), we proceed similarly to the proof of Lemma 5.2.5 with

$$Z_{k,h} := \frac{1}{n} g(Y_k; \tilde{\theta}) K_h(X_k - x) \, .$$

We need to show that there are constants C_1, C_2 depending only on Θ, J, K, $[a,b]$ and $\Gamma(a)$ so that for all $q \geq 2$

$$\mathbb{E}_\gamma \Big[\big| Z_{k,h} - \mathbb{E}_\gamma \big[Z_{k,h} \big] \big|^q \Big] \leq q! \bigg(\frac{C_1}{nh^d} \bigg)^{q-2} \frac{C_2}{n^2 h^d}$$

because then Lemma 3.2.3 gives the existence of constants C_1, C_2 independent of n, α, γ so that

$$\mathbb{P}_\gamma \bigg(\big| M_n\big(\theta, x; h_n(\alpha)\big) - \mathbb{E}_\gamma \big[M_n\big(\theta, x; h_n(\alpha)\big) \big] \big| \geq \omega \bigg) \leq 2 \exp \bigg(- \frac{\omega^2 n h_n(\alpha)^d}{C_1 + C_2 \omega} \bigg) \, , \quad \omega > 0 \, .$$

By proceeding similarly, we see that

$$\begin{aligned}
\mathbb{E}_\gamma \Big[\big| Z_{k,h} - \mathbb{E}_\gamma \big[Z_{k,h} \big] \big|^q \Big] &\leq 2^q \, \mathbb{E}_\gamma \big[\big| Z_{k,h} \big|^q \big] \\
&\leq \frac{2^q}{n^q} \iint \Big| g(y; \tilde{\theta}) K_h(z - x) \Big|^q f_{Y|X}^{\theta_*(\cdot)}(y|z) \ell(z) \, \mathrm{d}y \mathrm{d}z \\
&\leq \frac{2^q}{n^q} \bigg(\int |g(y; \tilde{\theta})|^q \sup_{\theta \in \overline{\Xi}} f_{\mathrm{mix}}(y; \theta) \, \mathrm{d}y \bigg) \cdot \bigg(\int |K_h^q(z - x)| \ell(z) \, \mathrm{d}z \bigg) \\
&\leq \frac{2^q \|K\|_\infty^q \|\ell\|_\infty \operatorname{diam}(J)}{n^q h^{d(q-1)}} \int |g(y; \tilde{\theta})|^q \sup_{\theta \in \overline{\Xi}} f_{\mathrm{mix}}(y; \theta) \, \mathrm{d}y \, ,
\end{aligned}$$

where we can treat the integral by (5.2.19) and convexity of $y \mapsto |y|^q$, i.e. there is a constant $C_{\overline{\Xi}} > 0$ depending only on $\overline{\Xi}$ so that

$$\begin{aligned}
&\int |g(y; \tilde{\theta})|^q \sup_{\theta \in \overline{\Xi}} f_{\mathrm{mix}}(y; \theta) \, \mathrm{d}y \\
&\leq 3^{q-1} C_{\overline{\Xi}} \bigg\{ \int \big(t - \mu_+ \big)^{2q} \sup_{\theta \in \overline{\Xi}} f_{\mathrm{mix}}(y; \theta) \, \mathrm{d}y + \int \big(t - \mu_- \big)^{2q} \sup_{\theta \in \overline{\Xi}} f_{\mathrm{mix}}(y; \theta) \, \mathrm{d}y \\
&\qquad + \int \sup_{\theta \in \overline{\Xi}} f_{\mathrm{mix}}(y; \theta) \, \mathrm{d}y \bigg\} \, . \qquad (5.2.17)
\end{aligned}$$

By repeating the subsequent considerations in the proof of Lemma 5.2.5 as well as (5.2.30), we get

$$\begin{aligned}
(5.2.17) &\leq 3^{q-1} \, 2 \, C_{\overline{\Xi}} \, c_* \int \big| y \big|^{2q} c_* \phi(y | 0, s_*^2) \, \mathrm{d}y \\
&\quad + \, 3^{q-1} C_{\overline{\Xi}} \int \sup_{\theta \in \overline{\Xi}} f_{\mathrm{mix}}(y; \theta) \, \mathrm{d}y \leq 3^q C_{\overline{\Xi}} C_* s_*^{2q} c_* 2^q q!
\end{aligned}$$

for some constants $0 < c_*, s_*, C_* < \infty$.

□

5.2.3 Proofs for auxiliary results

Proof of Lemma 5.2.2. Note that (5.2.4) follows from (5.2.5) directly. So, let us only show the latter.

Note that there are points $y_1 < y_2$ depending only on the compact parameter set $\overline{\Xi}$ so that

$$\mathbb{1}_{y<y_1} \sup_{\vartheta\in\overline{\Xi}} f_{\mathrm{mix}}(y;\vartheta) \le \mathbb{1}_{y<y_1}\phi(y|\mu_-,\sigma_+^2)\ ,\ \ \mathbb{1}_{y>y_2} \sup_{\vartheta\in\overline{\Xi}} f_{\mathrm{mix}}(y;\vartheta) \le \mathbb{1}_{y>y_2}\phi(y|\mu_+,\sigma_+^2)\ ,$$
$$\mathbb{1}_{y\in[y_1,y_2]} \sup_{\vartheta\in\overline{\Xi}} f_{\mathrm{mix}}(y;\vartheta) \le \mathbb{1}_{y\in[y_1,y_2]} \frac{1}{\sqrt{2\pi}\sigma_-}\ .$$

so that in conclusion

$$\sup_{\vartheta\in\overline{\Xi}} f_{\mathrm{mix}}(y;\vartheta) \le \mathbb{1}_{y<y_1}\phi(y|\mu_-,\sigma_+^2) + \mathbb{1}_{y>y_2}\phi(y|\mu_+,\sigma_+^2) + \mathbb{1}_{y\in[y_1,y_2]}\frac{1}{\sqrt{2\pi}\sigma_-} =: f(y)\ . \tag{5.2.18}$$

First, consider $\tau = g$. For any $y \in \mathbb{R}$ we have

$$\begin{aligned}
\sup_{\theta\in\overline{\Xi}} \big|g(y;\theta)\big| &= \sup_{\theta\in\overline{\Xi}} \left| \log\left(\sum_{c=1}^m \pi_c \phi(y|\mu_c,\sigma_c^2)\right)\right| \\
&\le \sup_{\theta\in\overline{\Xi}} \max\Big\{ -\min_{c=1,\dots,m} \log\big(\phi(y|\mu_c,\sigma_c^2)\big), \max_{c=1,\dots,m} \log\big(\phi(y|\mu_c,\sigma_c^2)\big)\Big\} \\
&= \sup_{\theta\in\overline{\Xi}} \max\Big\{ \max_{c=1,\dots,m} \frac{(y-\mu_c)^2}{2\sigma_c^2} + \log(\sqrt{2\pi}\sigma_c), \max_{c=1,\dots,m} -\frac{(y-\mu_c)^2}{2\sigma_c^2} - \log(\sqrt{2\pi}\sigma_c)\Big\}\ ,
\end{aligned}$$

where we can treat the terms in the maximum by

$$\begin{aligned}
&\max_{c=1,\dots,m} \frac{(y-\mu_c)^2}{2\sigma_c^2} + \log(\sqrt{2\pi}\sigma_c) \\
\le &\max_{c=1,\dots,m} \frac{(y-\mu_c)^2}{2\sigma_-^2} + \log(\sqrt{2\pi}\sigma_+) \\
\le &\max_{\mu\in\{\mu_-,\mu_+\}} \frac{(y-\mu)^2}{2\sigma_-^2} + \log(\sqrt{2\pi}\sigma_+) \\
\le &\frac{(y-\mu_+)^2}{2\sigma_-^2}\mathbb{1}_{y<\mu_-} + \frac{(\mu_+-\mu_-)^2}{2\sigma_-^2}\mathbb{1}_{y\in[\mu_-,\mu_+]} + \frac{(y-\mu_-)^2}{2\sigma_-^2}\mathbb{1}_{y>\mu_+} + \log(\sqrt{2\pi}\sigma_+)
\end{aligned}$$

and

$$\max_{c=1,\dots,m} -\frac{(y-\mu_c)^2}{2\sigma_c^2} - \log(\sqrt{2\pi}\sigma_c) \le -\log(\sqrt{2\pi}\sigma_-)$$

Hence,

$$\sup_{\theta\in\overline{\Xi}} \big|g(y;\theta)\big|$$

$$\leq \frac{(y-\mu_+)^2}{2\sigma_-^2}\mathbb{1}_{y<\mu_-} + \frac{(\mu_+-\mu_-)^2}{2\sigma_-^2}\mathbb{1}_{y\in[\mu_-,\mu_+]} + \frac{(y-\mu_-)^2}{2\sigma_-^2}\mathbb{1}_{y>\mu_+} + \log(\sqrt{2\pi}\sigma_+) + |\log(\sqrt{2\pi}\sigma_-)|\,, \tag{5.2.19}$$

which is integrable with respect to $f(t)\,\mathrm{d}t$. As all moments of normal and uniform distributions exist, we deduce (5.2.5) for $\tau = g$.

Similar arguments including that for any $\theta \in \overline{\Xi}$, $y \in \mathbb{R}$

$$\left|\frac{\pi_c\phi(y|\mu_c,\sigma_c^2)}{\sum_{c=1}^m \pi_c\phi(y|\mu_c,\sigma_c^2)}\right| \leq 1$$
$$\left|\frac{\phi(y|\mu_c,\sigma_c^2)}{\sum_{c=1}^m \pi_c\phi(y|\mu_c,\sigma_c^2)}\right| \leq \frac{1}{\pi_c} \leq \frac{1}{\pi_-}$$

treat the other cases of τ.

(5.2.6) can be directly concluded from the considerations after (5.2.19).

□

Proof of Lemma 5.2.3. Show (i). By means of the triangular equality,

$$\left|f_{\mathrm{mix}}(y;\theta) - f_{\mathrm{mix}}(y;\tilde{\theta})\right| \leq \sum_{c=1}^m \left|\pi_c\phi(y|\mu_c,\sigma_c^2) - \tilde{\pi}_c\phi(y|\tilde{\mu}_c,\tilde{\sigma}_c^2)\right|,$$

where each of the summands is treated by

$$\begin{aligned}
&\left|\pi_c\phi(y|\mu_c,\sigma_c^2) - \tilde{\pi}_c\phi(y|\tilde{\mu}_c,\tilde{\sigma}_c^2)\right| \\
\leq&\left|\pi_c - \tilde{\pi}_c\right|\phi(y|\mu_c,\sigma_c^2) && (5.2.20)\\
&+ \frac{\tilde{\pi}_c}{\sqrt{2\pi}\sigma_c}\left|\exp\left(-\frac{(y-\mu_c)^2}{2\sigma_c^2}\right) - \exp\left(-\frac{(y-\tilde{\mu}_c)^2}{2\sigma_c^2}\right)\right| && (5.2.21)\\
&+ \tilde{\pi}_c\left|\frac{1}{\sqrt{2\pi}\sigma_c}\exp\left(-\frac{(y-\tilde{\mu}_c)^2}{2\sigma_c^2}\right) - \frac{1}{\sqrt{2\pi}\tilde{\sigma}_c}\exp\left(-\frac{(y-\tilde{\mu}_c)^2}{2\tilde{\sigma}_c^2}\right)\right|, && (5.2.22)
\end{aligned}$$

where $|\pi_m - \tilde{\pi}_m| \leq \sum_{c=1}^{m-1}|\pi_c - \tilde{\pi}_c|$ and the function $\sup_{\theta\in\overline{\Xi}}\phi(\cdot|\mu_c,\sigma_c^2)$ is integrable as $\overline{\Xi}$ is compact so that (5.2.20) is dealt with.

As the inequality $\exp(t) \geq 1+t$, $t \in \mathbb{R}$ implies $1-\exp(t) \leq -t$, we deduce for $x, y \in [0,\infty)$

$$\begin{aligned}
\big|\exp(-x) - \exp(-y)\big| &= \exp\big(-\min\{x,y\}\big)\Big(1 - \exp\big(\min\{x,y\} - \max\{x,y\}\big)\Big)\\
&\leq \exp\big(-\min\{x,y\}\big)\big(\max\{x,y\} - \min\{x,y\}\big)\\
&= \max\big\{\exp(-x),\exp(-y)\big\}\big|x-y\big|\,. && (5.2.23)
\end{aligned}$$

Hence, (5.2.21) is dealt with by

$$\begin{aligned}
&\left|\exp\left(-\frac{(y-\mu_c)^2}{2\sigma_c^2}\right)-\exp\left(-\frac{(y-\tilde{\mu}_c)^2}{2\sigma_c^2}\right)\right| \\
\leq &\max\left\{\exp\left(-\frac{(y-\mu_c)^2}{2\sigma_c^2}\right),\exp\left(-\frac{(y-\tilde{\mu}_c)^2}{2\sigma_c^2}\right)\right\}\left|\frac{(y-\mu_c)^2}{2\sigma_c^2}-\frac{(y-\tilde{\mu}_c)^2}{2\sigma_c^2}\right| \\
\leq &\max\left\{\exp\left(-\frac{(y-\mu_c)^2}{2\sigma_c^2}\right),\exp\left(-\frac{(y-\tilde{\mu}_c)^2}{2\sigma_c^2}\right)\right\}\frac{1}{2\sigma_c^2}\left|\mu_c^2-\tilde{\mu}_c^2+2y(\tilde{\mu}_c-\mu_c)\right| \\
\leq &\max\left\{\exp\left(-\frac{(y-\mu_c)^2}{2\sigma_c^2}\right),\exp\left(-\frac{(y-\tilde{\mu}_c)^2}{2\sigma_c^2}\right)\right\}\frac{1}{2\sigma_c^2}\Big(|\mu_c+\tilde{\mu}_c|+2y\Big)|\mu_c-\tilde{\mu}_c|\,,
\end{aligned}$$

where this first factor is uniformly integrable.

Similar arguments and

$$\begin{aligned}
\left|\frac{1}{\sigma_c^2}-\frac{1}{\tilde{\sigma}_c^2}\right| \leq &\frac{|\sigma_c^2-\tilde{\sigma}_c^2|}{\sigma_c^2\tilde{\sigma}_c^2} \\
\leq &\frac{2\sigma_+}{\sigma_-^4}|\sigma_c-\tilde{\sigma}_c|
\end{aligned}$$

by compactness of $\overline{\Xi}$ yield the corresponding bound for (5.2.22). Similar arguments to the ones used in the proof of Lemma 5.2.5 yield the constants c_* and s_*.

(ii) By noting that for Hölder-α-smooth functions F, we have $|F(x)-F(x')| \leq L\|x-x'\|^{\min\{1,\alpha\}}$ as they are continuously differentiable for $\alpha > 1$, we deduce (ii) directly from (i).

(iii) First note that for integrability with respect to $\sup_{\theta\in\overline{\Xi}} f_{\text{mix}}(y;\theta)$ it is enough to show boundedness on compacta and integrability with respect to any normal distribution outside of those compacta because according to (5.2.18), there are constants $y_1 < y_2$ so that

$$\sup_{\theta\in\overline{\Xi}} f_{\text{mix}}(y;\theta) \leq \mathbb{1}_{y<y_1}\phi(y|\mu_-,\sigma_+^2) + \mathbb{1}_{y>y_2}\phi(y|\mu_+,\sigma_+^2) + \mathbb{1}_{y\in[y_1,y_2]}\frac{1}{\sqrt{2\pi}\sigma_-}\,.$$

Let us only prove this result for the partial derivative $\partial_{\mu_c} g$. The arguments for the other functions work similarly.

Write

$$\theta = (\pi_1,\ldots,\pi_{m-1},\mu_1,\ldots,\mu_m,\sigma_1,\ldots,\sigma_m), \quad \tilde{\theta} = (\tilde{\pi}_1,\ldots,\tilde{\pi}_{m-1},\tilde{\mu}_1,\ldots,\tilde{\mu}_m,\tilde{\sigma}_1,\ldots,\tilde{\sigma}_m)$$

as well as for $r = 0,\ldots,m-2$

$$\begin{aligned}
\tilde{\theta}^r(\theta) &= (\pi_1,\ldots,\pi_r,\tilde{\pi}_{r+1},\ldots,\tilde{\pi}_{m-1},\mu_1,\ldots,\mu_r,\tilde{\mu}_{r+1},\ldots,\tilde{\mu}_m,\sigma_1,\ldots,\sigma_r,\tilde{\sigma}_{r+1},\ldots,\tilde{\sigma}_m)\,, \\
\tilde{\theta}^{m-1}(\theta) &= (\pi_1,\ldots,\pi_{m-1},\mu_1,\ldots,\mu_{m-1},\tilde{\mu}_m,\sigma_1,\ldots,\sigma_{m-1},\tilde{\sigma}_m)\,,
\end{aligned}$$

$$\tilde{\theta}^m(\theta) = \theta \ .$$

Observe

$$\left|\partial_{\mu_j} g(y;\theta) - \partial_{\mu_j} g(y;\tilde{\theta})\right| \leq \sum_{r=0}^{m-1} \left|\partial_{\mu_j} g(y;\tilde{\theta}^r(\theta)) - \partial_{\mu_j} g(y;\tilde{\theta}^{r+1}(\theta))\right| \ .$$

Without loss of generality let $r = j-1$, otherwise the calculations simplify. The main trick now is to insert terms of the shape $\min\left\{\phi(y|\tilde{\mu}_j,\tilde{\sigma}_j^2), \phi(y|\mu_j,\sigma_j^2)\right\}$. Observe

$$\begin{aligned}
&\left|\partial_{\mu_j} g(y;\tilde{\theta}^{j-1}(\theta)) - \partial_{\mu_j} g(y;\tilde{\theta}^{j}(\theta))\right| \\
=&\left|\frac{\tilde{\pi}_j(\tilde{\mu}_j - y)\tilde{\sigma}_j^{-2}\phi(y|\tilde{\mu}_j,\tilde{\sigma}_j^2)}{f_{\mathrm{mix}}(y;\tilde{\theta}^{j-1}(\theta))} - \frac{\pi_j(\mu_j - y)\sigma_j^{-2}\phi(y|\mu_j,\sigma_j^2)}{f_{\mathrm{mix}}(y;\tilde{\theta}^{j}(\theta))}\right| \\
=&\left|\frac{\tilde{\pi}_j(\tilde{\mu}_j - y)\tilde{\sigma}_j^{-2}\phi(y|\tilde{\mu}_j,\tilde{\sigma}_j^2) f_{\mathrm{mix}}(y;\tilde{\theta}^{j}(\theta)) - \pi_j(\mu_j - y)\sigma_j^{-2}\phi(y|\mu_j,\sigma_j^2) f_{\mathrm{mix}}(y;\tilde{\theta}^{j-1}(\theta))}{f_{\mathrm{mix}}(y;\tilde{\theta}^{j-1}(\theta)) f_{\mathrm{mix}}(y;\tilde{\theta}^{j}(\theta))}\right| \\
\leq&\left|\frac{\Big(\phi(y|\tilde{\mu}_j,\tilde{\sigma}_j^2) - \min\left\{\phi(y|\tilde{\mu}_j,\tilde{\sigma}_j^2), \phi(y|\mu_j,\sigma_j^2)\right\}\Big)\tilde{\pi}_j(\tilde{\mu}_j - y)\tilde{\sigma}_j^{-2} f_{\mathrm{mix}}(y;\tilde{\theta}^{j}(\theta))}{f_{\mathrm{mix}}(y;\tilde{\theta}^{j-1}(\theta)) f_{\mathrm{mix}}(y;\tilde{\theta}^{j}(\theta))}\right| \qquad (5.2.24) \\
&+\left|\frac{\Big(\tilde{\pi}_j(\tilde{\mu}_j - y)\tilde{\sigma}_j^{-2} - \pi_j(\mu_j - y)\sigma_j^{-2}\Big)\min\left\{\phi(y|\tilde{\mu}_j,\tilde{\sigma}_j^2), \phi(y|\mu_j,\sigma_j^2)\right\} f_{\mathrm{mix}}(y;\tilde{\theta}^{j}(\theta))}{f_{\mathrm{mix}}(y;\tilde{\theta}^{j-1}(\theta)) f_{\mathrm{mix}}(y;\tilde{\theta}^{j}(\theta))}\right| \qquad (5.2.25) \\
&+\left|\frac{\Big(f_{\mathrm{mix}}(y;\tilde{\theta}^{j}(\theta)) - f_{\mathrm{mix}}(y;\tilde{\theta}^{j-1}(\theta))\Big)\min\left\{\phi(y|\tilde{\mu}_j,\tilde{\sigma}_j^2), \phi(y|\mu_j,\sigma_j^2)\right\}\pi_j(\mu_j - y)\sigma_j^{-2}}{f_{\mathrm{mix}}(y;\tilde{\theta}^{j-1}(\theta)) f_{\mathrm{mix}}(y;\tilde{\theta}^{j}(\theta))}\right| \qquad (5.2.26) \\
&+\left|\frac{\Big(\min\left\{\phi(y|\tilde{\mu}_j,\tilde{\sigma}_j^2), \phi(y|\mu_j,\sigma_j^2)\right\} - \phi(y|\mu_j,\sigma_j^2)\Big)\pi_j(\mu_j - y)\sigma_j^{-2} f_{\mathrm{mix}}(y;\tilde{\theta}^{j-1}(\theta))}{f_{\mathrm{mix}}(y;\tilde{\theta}^{j-1}(\theta)) f_{\mathrm{mix}}(y;\tilde{\theta}^{j}(\theta))} \ .\right| \qquad (5.2.27)
\end{aligned}$$

Now, treat (5.2.24) by the fact that

$$\begin{aligned}
&\left|\phi(y|\tilde{\mu}_j,\tilde{\sigma}_j^2) - \min\left\{\phi(y|\tilde{\mu}_j,\tilde{\sigma}_j^2), \phi(y|\mu_j,\sigma_j^2)\right\}\right| \\
\leq&\begin{cases}\phi(y|\tilde{\mu}_j,\tilde{\sigma}_j^2) \ , & \text{if } \phi(y|\tilde{\mu}_j,\tilde{\sigma}_j^2) > \phi(y|\mu_j,\sigma_j^2) \\ 0, & \text{else}\end{cases}
\end{aligned}$$

and the fact that

$$\frac{\phi(y|\tilde{\mu}_j,\tilde{\sigma}_j^2)}{f_{\mathrm{mix}}(y;\tilde{\theta}^{j-1}(\theta))}$$

is uniformly bounded over all $\theta, \tilde{\theta}$, so that (5.2.24) is bounded on a compact set around 0 and bounded by a linear function on the compact set's complement. As such, this

function is integrable with respect to any normal density. The term (5.2.27) is treated accordingly.

For (5.2.25), we particularly use that

$$\frac{\min\left\{\phi(y|\tilde{\mu}_j,\tilde{\sigma}_j^2),\phi(y|\mu_j,\sigma_j^2)\right\}}{f_{\mathrm{mix}}(y;\tilde{\theta}^{j-1}(\theta))} \leq \frac{\phi(y|\tilde{\mu}_j,\tilde{\sigma}_j^2)}{f_{\mathrm{mix}}(y;\tilde{\theta}^{j-1}(\theta))} \,.$$

For (5.2.26), we also make use of the fact that

$$\begin{aligned}
&\left|f_{\mathrm{mix}}\big(y;\tilde{\theta}^{j}(\theta)\big) - f_{\mathrm{mix}}\big(y;\tilde{\theta}^{j-1}(\theta)\big)\right| \\
&= \left|\tilde{\pi}_j\phi(y|\tilde{\mu}_j,\tilde{\sigma}_j^2) - \pi_j\phi(y|\mu_j,\sigma_j^2)\right| \\
&\leq \left|\tilde{\pi}_j - \pi_j\right|\phi(y|\tilde{\mu}_j,\tilde{\sigma}_j^2) + \pi_j\left|\phi(y|\tilde{\mu}_j,\tilde{\sigma}_j^2) - \phi(y|\mu_j,\sigma_j^2)\right| \,,
\end{aligned}$$

where

$$\begin{aligned}
&\left|(\phi(y|\tilde{\mu}_j,\tilde{\sigma}_j^2) - \phi(y|\mu_j,\sigma_j^2)\right| \\
&\leq \left|\frac{1}{\sqrt{2\pi}\tilde{\sigma}_j} - \frac{1}{\sqrt{2\pi}\sigma_j}\right| \exp\left(-\frac{(y-\tilde{\mu}_j)^2}{2\tilde{\sigma}_j^2}\right) \\
&\quad + \frac{1}{\sqrt{2\pi}\sigma_j}\left|\exp\left(-\frac{(y-\tilde{\mu}_j)^2}{2\tilde{\sigma}_j^2}\right) - \exp\left(-\frac{(y-\mu_j)^2}{2\sigma_j^2}\right)\right|
\end{aligned}$$

and by (5.2.23), we deduce

$$\begin{aligned}
&\left|\exp\left(-\frac{(y-\tilde{\mu}_j)^2}{2\tilde{\sigma}_j^2}\right) - \exp\left(-\frac{(y-\mu_j)^2}{2\sigma_j^2}\right)\right| \\
&\leq \max\left\{\exp\left(-\frac{(y-\tilde{\mu}_j)^2}{2\tilde{\sigma}_j^2}\right), \exp\left(-\frac{(y-\mu_j)^2}{2\sigma_j^2}\right)\right\} \cdot \left|\frac{(y-\mu_j)^2}{2\sigma_j^2} - \frac{(y-\tilde{\mu}_j)^2}{2\tilde{\sigma}_j^2}\right| \\
&\lesssim \max\left\{\phi(y|\tilde{\mu}_j,\tilde{\sigma}_j^2),\phi(y|\mu_j,\sigma_j^2)\right\}\Big(\left|\mu_j-\tilde{\mu}_j\right| + \left|\sigma_j - \tilde{\sigma}_j\right|\Big)
\end{aligned}$$

so that

$$\begin{aligned}
&\left|\frac{\Big(f_{\mathrm{mix}}\big(y;\tilde{\theta}^{j}(\theta)\big) - f_{\mathrm{mix}}\big(y;\tilde{\theta}^{j-1}(\theta)\big)\Big)\min\left\{\phi(y|\tilde{\mu}_j,\tilde{\sigma}_j^2),\phi(y|\mu_j,\sigma_j^2)\right\}}{f_{\mathrm{mix}}\big(y;\tilde{\theta}^{j-1}(\theta)\big)f_{\mathrm{mix}}\big(y;\tilde{\theta}^{j}(\theta)\big)}\right| \\
&\lesssim \Big(\left|\tilde{\pi}_j - \pi_j\right| + \left|\mu_j - \tilde{\mu}_j\right| + \left|\sigma_j - \tilde{\sigma}_j\right|\Big) \, \frac{\phi(y|\tilde{\mu}_j,\tilde{\sigma}_j^2)\phi(y|\mu_j,\sigma_j^2)}{f_{\mathrm{mix}}\big(y;\tilde{\theta}^{j-1}(\theta)\big)f_{\mathrm{mix}}\big(y;\tilde{\theta}^{j}(\theta)\big)} \,,
\end{aligned}$$

where the right-hand factor is uniformly bounded on a compact set around zero due to the compactness of $\overline{\Xi}$ and uniformly integrable with respect to any normal density as well.

□

Proof of Lemma 5.2.4. Minding the integrability result (5.2.5) in Lemma 5.2.2, we get that independently of $\alpha \in [a,b]$, $\gamma \in \Gamma(a)$, for any $\tau \in \{g, \partial_{\pi_c} g, \partial_{\mu_c} g, \partial_{\sigma_c} g\}$, we have

$$
\begin{aligned}
&\sup_{x\in J,\theta\in\overline{\Xi}} \mathrm{Var}_\gamma\left(\frac{1}{n}\sum_{k=1}^n \tau(Y_k;\theta)K_h(X_k-x)\right)\\
=&\sup_{x\in J,\theta\in\overline{\Xi}} \frac{1}{n}\left\{\int \tau^2(y,\theta)\Big(K_h^2 * f_{Y|X}^{\theta_*(\cdot)}(y|\cdot)\ell\Big)(x)\,\mathrm{d}y - \left(\int \tau(y;\theta)\Big(K_h * f_{Y|X}^{\theta_*(\cdot)}(y|\cdot)\ell\Big)(x)\,\mathrm{d}y\right)^2\right\}\\
\leq&\frac{1}{n}\sup_{x\in J}\big(K_h^2*\ell\big)(x)\int \sup_{\theta\in\overline{\Xi}}\tau^2(y;\theta)\sup_{\vartheta\in\overline{\Xi}} f_{\mathrm{mix}}(y;\vartheta)\,\mathrm{d}y\\
&+\frac{1}{n}\sup_{x\in J}\Big(\big(K_h*\ell\big)(x)\Big)^2\left(\int \sup_{\theta\in\overline{\Xi}}|\tau(y;\theta)|\sup_{\vartheta\in\overline{\Xi}} f_{\mathrm{mix}}(y;\vartheta)\,\mathrm{d}y\right)^2\\
=&O(n^{-1}h^{-d})+O(n^{-1})\,.
\end{aligned}
$$

□

In order to prove Lemma 5.2.5, we first observe that the functions

$$
\begin{aligned}
(y,\theta)\mapsto \partial_{\mu_j} g(y;\theta) &= \frac{\pi_j(\mu_j-y)\sigma_j^{-2}\phi(y|\mu_j,\sigma_j^2)}{\sum_{c=1}^m \pi_c\phi(y|\mu_c,\sigma_c^2)}\,,\\
(y,\theta)\mapsto \partial_{\sigma_j} g(y;\theta) &= \frac{\pi_j\big((y-\mu_j)^2-\sigma_j^2\big)\sigma_j^{-3}\phi(y|\mu_j,\sigma_j^2)}{\sum_{c=1}^m \pi_c\phi(y|\mu_c,\sigma_c^2)}
\end{aligned}
$$

are bounded, whenever $\theta_{m+j}=\sigma_j<\max\{\sigma_c : c=1,\ldots,m\}$ because then

$$
\frac{\phi(y|\mu_j,\sigma_j^2)}{\sum_{c=1}^m \pi_c\phi(y|\mu_c,\sigma_c^2)}
$$

converges to zero exponentially fast, $|y|\to\infty$ and because $\overline{\Xi}$ is compact. For those θ, one could use Lemma 3.2.2.

However, if $\theta_{m+j}=\sigma_j>\max\big\{\sigma_c : c=1,\ldots,m\big\}\backslash\{\sigma_j\}$, then

$$
\frac{\phi(y|\mu_j,\sigma_j^2)}{\sum_{c=1}^m \pi_c\phi(y|\mu_c,\sigma_c^2)}\to\frac{1}{\pi_j}\,,\quad |y|\to\infty\,,
$$

so that the derivatives with respect to μ_j or σ_j are not bounded. That means that for those θ, one cannot use Lemma 3.2.2. Particularly, as the derivatives are continuous in y,θ, the exponential bound one gets for $\theta_{m+j}=\sigma_j<\max\{\sigma_c : c=1,\ldots,m\}$ must depend on θ in the term R, where then $\sup_\theta R=\infty$, which does not suffice.

Hence, we need to use an alternative approach, that is, using Bennett's inequality, cf. Lemma 3.2.3.

Proof of Lemma 5.2.5. Let us first treat the case when $t = \pi_c$ for some c. As for all $y \in \mathbb{R}$, $\theta \in \overline{\overline{\Xi}}$, we have

$$|\partial_{\pi_j} g(y;\theta)| = \left|\frac{\partial_{\pi_j} f_{\mathrm{mix}}(y;\theta)}{f_{\mathrm{mix}}(y;\theta)}\right| = \left|\frac{\phi(y|\mu_j,\sigma_j^2) - \phi(y|\mu_m,\sigma_m^2)}{\sum_{c=1}^m \pi_c \phi(y|\mu_c,\sigma_c^2)}\right| \leq \frac{2}{\pi_-}\,, \tag{5.2.28}$$

and because K is a bounded kernel, Lemma 5.2.4, the monotonicity of $\omega \mapsto \exp(-1/\omega)$ and Lemma 3.2.2 give for all $\theta \in \overline{\overline{\Xi}}$, $x \in J$, $\gamma \in \Gamma(a)$, $\omega > 0$,

$$\begin{aligned}
&\mathbb{P}_\gamma\big(\big|\partial_t M_n(\theta,x;h) - \mathbb{E}_\gamma[\partial_t M_n(\theta,x;h)\big| \geq \omega\big) \\
&\leq 2\exp\bigg(-\frac{\omega^2}{2\sup_{x\in J,\theta\in\overline{\overline{\Xi}}} \mathrm{Var}_\gamma\big(\frac{1}{n}\sum_{k=1}^n \partial_t g(Y_k;\theta)K_h(X_k-x)\big) + \frac{3\|K\|_\infty \omega}{2\pi_- nh^d}}\bigg) \\
&\leq 2\exp\bigg(-\frac{\omega^2 nh^d}{C_1 + C_2\omega}\bigg)
\end{aligned}$$

when C_1, C_2 are chosen large enough.

Now, treat the case $t = \mu_c$ for some c. Write

$$Z_{k,h} := \frac{1}{n}\bigg(\frac{\mu_j - Y_k}{\sigma_j^2}\bigg)\frac{\pi_j\phi(Y_k|\mu_j,\sigma_j^2)}{\sum_{c=1}^m \pi_c\phi(Y_k|\mu_c,\sigma_c^2)}K_h(X_k - x)\,.$$

We need to show that there are constants C_1, C_2 depending only on $\overline{\overline{\Xi}}$, J, K, $[a,b]$ and $\Gamma(a)$ so that for all $q \geq 2$

$$\mathbb{E}_\gamma\Big[\big|Z_{k,h} - \mathbb{E}_\gamma\big[Z_{k,h}\big]\big|^q\Big] \leq q!\bigg(\frac{C_1}{nh^d}\bigg)^{q-2}\frac{C_2}{n^2h^d}\,,$$

because then Lemma 3.2.3 gives

$$\mathbb{P}_\gamma\Big(\big|\partial_t M_n(\theta,x;h) - \mathbb{E}_\gamma\big[\partial_t M_n(\theta,x;h)\big]\big| \geq \omega\Big) \leq 2\exp\bigg(-\frac{1}{2}\frac{\omega^2 nh^d}{C_1 + C_2\omega}\bigg)\,, \quad \omega > 0\,.$$

As the function $y \mapsto |y|^q$ is convex, we have that $|y_1 - y_2|^q \leq 2^{q-1}|y_1|^q + 2^{q-1}|y_2|^q$, so that according to (5.2.28)

$$\begin{aligned}
&\mathbb{E}_\gamma\big[|Z_{k,h} - \mathbb{E}_\gamma[Z_{k,h}]|^q\big] \\
&\leq 2^{q-1}\mathbb{E}_\gamma\big[|Z_{k,h}|^q\big] + 2^{q-1}\big(\mathbb{E}_\gamma[Z_{k,h}]\big)^q \\
&\leq 2^q\mathbb{E}_\gamma\big[|Z_{k,h}|^q\big] \\
&\leq \frac{2^q}{n^q}\iint\bigg|\frac{\mu_j - y}{\sigma_j^2}\frac{\pi_j\phi(y|\mu_j,\sigma_j^2)}{\sum_{c=1}^m \pi_c\phi(y|\mu_c,\sigma_c^2)}K_h(z-x)\bigg|^q f_{Y|X}^{\theta_*(\cdot)}(y|z)\ell(z)\,\mathrm{d}y\mathrm{d}z \\
&\leq \frac{2^q}{n^q}\iint\bigg|\frac{\mu_j - y}{\sigma_j^2}K_h(z-x)\bigg|^q f_{Y|X}^{\theta_*(\cdot)}(y|z)\ell(z)\,\mathrm{d}y\mathrm{d}z
\end{aligned}$$

by Jensen's inequality and (5.2.28). Now, the integral can be treated by the transformation $y \mapsto \sigma_j^2 y + \mu_j$, i.e.

$$\begin{aligned}
&\iint \left| \frac{\mu_j - y}{\sigma_j^2} K_h(z-x) \right|^q f_{Y|X}^{\theta_*(\cdot)}(y|z)\ell(z)\,\mathrm{d}y\mathrm{d}z \\
=&\iint |y|^q \sigma_j^2 f_{Y|X}^{\theta_*(\cdot)}(\sigma_j^2 y + \mu_j|z) \big|K_h(z-x)\big|^q \ell(z)\,\mathrm{d}y\mathrm{d}z \\
\leq&\sigma_+^2 \int |y|^q c_* \phi(y|0,s_*^2)\,\mathrm{d}y \cdot \left(\frac{1}{h^d}\right)^{q-1} \int \frac{1}{h^d} \left| K\left(\frac{z-x}{h}\right)\right|^q \ell(z)\,\mathrm{d}z\,,
\end{aligned}$$

where c_* and s_* will be specified later. Now, as absolute moments of centred normal distributed variables N with variance s_*^2 are bounded by $\mathbb{E}[|N|^q] \leq (q-1)!!s_*^q$, we get that

$$\int |y|^q c_* \phi(y|0,s_*^2)\,\mathrm{d}y \leq s_*^q c_* q!\,,$$

furthermore, we have that

$$\sup_{x\in J} \left| \int \frac{1}{h^d} \left| K\left(\frac{z-x}{h}\right)\right|^q \ell(z)\,\mathrm{d}z \right| \leq \|K\|_\infty^q \|\ell\|_\infty \operatorname{diam}(J)$$

so that finally for all $q \geq 2$

$$\mathbb{E}_\gamma\Big[\big|Z_{k,h} - \mathbb{E}_\gamma\big[Z_{k,h}\big]\big|^q\Big] \leq q! \left(\frac{C_1}{nh^d}\right)^{q-2} \frac{C_2}{n^2 h^d}$$

with C_1, C_2 chosen appropriately.

In order to find c_*, s_*^2, we observe that for any $y \in \mathbb{R}$

$$\begin{aligned}
f_{Y|X}^{\theta_*(\cdot)}\big(\sigma_j^2 y + \mu_j|z\big) &\leq \sup_{\substack{z\in J, \mu_{c'}, \sigma_{c'} \\ c'=1,\dots,m}} f_{Y|X}^{\theta_*(\cdot)}\big(\sigma_{c'}^2 y + \mu_{c'}|z\big) \\
&= \sup_{\substack{z\in J, \mu_c, \sigma_c \\ c'=1,\dots,m}} \sum_{c=1}^m \pi_c(z) \phi\big(\sigma_{c'}^2 y + \mu_{c'}|\mu_c(z), \sigma_c^2(z)\big) \\
&\leq \sup_{\substack{\mu_c, \sigma_c, \mu_{c'}, \sigma_{c'} \\ c=1,\dots,m \\ c'=1,\dots,m}} \sigma_{c'}^2 \phi\left(y \middle| \frac{\mu_c - \mu_{c'}}{\sigma_{c'}^2}, \frac{\sigma_c^2}{\sigma_{c'}^4}\right). \qquad (5.2.29)
\end{aligned}$$

Define the compact sets

$$\begin{aligned}
\Sigma_1 :=& \left\{ \frac{\mu_c - \tilde{\mu}_{c'}}{\tilde{\sigma}_{c'}} : \theta, \tilde{\theta} \in \overline{\Xi}, c, c' = 1, \dots, m \right\}, \\
\Sigma_2 :=& \left\{ \frac{\sigma_c^2}{\tilde{\sigma}_{c'}^4} : \theta, \theta' \in \overline{\Xi}, c, c' = 1, \dots, m \right\}
\end{aligned}$$

as well as $\bar{\sigma}_+ = \operatorname{argmax}\{\sigma^2 \in \Sigma_2\}$, $\bar{\sigma}_- = \operatorname{argmin}\{\sigma^2 \in \Sigma_2\}$, $\bar{\mu}_+ = \max \Sigma_1$, $\bar{\mu}_- = \min \Sigma_1$ that exist due to the compactness of $\overline{\Xi}$. Then, by compactness of Σ_1, Σ_2 and (5.2.10) there is a constant $y_1 > 0$ so that for all $y \in \mathbb{R}$

$$(5.2.29) \leq \sigma_+^2 \left(\phi(y|\bar{\mu}_+, \bar{\sigma}_+^2) \mathbb{1}_{y>y_1} + \phi(y|\bar{\mu}_-, \bar{\sigma}_+^2) \mathbb{1}_{y<-y_1} + \frac{1}{\sqrt{2\pi}\bar{\sigma}_-} \mathbb{1}_{y\in[-y_1,y_1]} \right) .$$

Furthermore, if $s_* > \bar{\sigma}_+$, (5.2.10) gives

$$\lim_{|y|\to\infty} \frac{\phi(y|0, s_*^2)}{\phi(y|\bar{\mu}_+, \bar{\sigma}_+^2)} = \lim_{|y|\to\infty} \frac{\phi(y|0, s_*^2)}{\phi(y|\bar{\mu}_-, \bar{\sigma}_+^2)} = \infty ,$$

hence, there is a constant $y_2 > y_1 > 0$ so that for all $y \geq |y_2|$

$$\phi(y|0, s_*^2) \geq \max\{\phi(y|\bar{\mu}_+, \bar{\sigma}_+^2), \phi(y|\bar{\mu}_-, \bar{\sigma}_+^2)\} .$$

If we finally define

$$c_* := \max\left\{1, \frac{\bar{\sigma}_+^2}{\sqrt{2\pi}\bar{\sigma}_- \phi(c_2|0, s_*^2)}\right\} ,$$

we have that

$$f_{Y|X}^{\theta_*(\cdot)}\left(\sigma_j^2 y + \mu_j | z\right) \leq c_* \phi(y|0, s_*^2) .$$

For the derivatives with respect to some σ_j, the proof works analogously, except that one examines $2m$-absolute moments of centred normal variables, i.e.

$$\int |y|^{2q} c_* \phi(y|0, s_*^2) \, \mathrm{d}y \leq s_*^q c_* (2q-1)!! \leq s_*^{2q} c_* 2^q q! . \tag{5.2.30}$$

Conclude by choosing $C_1, C_2 < \infty$ large enough and noting that they can be chosen independently of θ, x, h, n, γ.

□

5.3 Proofs for Section 4.2

5.3.1 Proof for the identifiability result

The following simple lemma states that it is enough to prove $\mu = \mu_*$ in order to deduce $\vartheta = \vartheta_*$ under the assumptions in Theorem 4.2.4.

Lemma 5.3.1. *Let $\vartheta_1, \vartheta_2 \in \mathfrak{X}$, $\vartheta_i = (p_i, \sigma_i, \mu_i, f_i)^T$, $i = 1, 2$. If $p_1 \in (0, 1)$, $\mu_1 = \mu_2 \neq 0$, and $f_{\mathrm{mix}}(y; \vartheta_1) = f_{\mathrm{mix}}(y; \vartheta_2)$ for almost all $y \in \mathbb{R}$, then $(p_1, \sigma_1, \mu_1) = (p_2, \sigma_2, \mu_2)$ and $f_1 = f_2$ almost surely.*

The proof of this lemma is straightforward. Calculate the first three moments of the mixture densities $f_{\mathrm{mix}}(\cdot; \vartheta_i)$, $i = 1, 2$, which need to coincide. From this system of equations, equality of the parameters follows directly. A complete proof can be found in Werner (2015).

Proof of Theorem 4.2.4. Since $\bar{f} \in \mathcal{E}_3$ has a finite second-order moment, we may assume without loss of generality that it is normalized to one, i.e.

$$\int y^2 \bar{f}(y)\,\mathrm{d}y = 1.$$

Denote $\vartheta = (p, \sigma, \mu, f)^T$. Taking the Fourier transform in (4.2.1), using that the Fourier transforms of $\bar{f}$, f_*, f are real-valued because of their symmetry and considering real and imaginary part separately gives for all $t \in \mathbb{R}$ that

$$\begin{aligned}(1-p_*)\varphi_{\bar{f}}(\sigma_* t) - (1-p)\varphi_{\bar{f}}(\sigma t) + p_* \cos(\mu_* t)\varphi_{f_*}(t) &= p\cos(\mu t)\varphi_f(t)\,,\\ p_* \sin(\mu_* t)\varphi_{f_*}(t) &= p\sin(\mu t)\varphi_f(t)\,.\end{aligned} \tag{5.3.1}$$

Multiplying these equations by $\sin(\mu t)$ and $\cos(\mu t)$, respectively, and using the trigonometric addition formula

$$\sin\big(\mu_* t - \mu t\big) = \sin(\mu_* t)\cos(\mu t) - \cos(\mu_* t)\sin(\mu t)$$

yields

$$\big[(1-p_*)\varphi_{\bar{f}}(\sigma_* t) - (1-p)\varphi_{\bar{f}}(\sigma t)\big]\sin(\mu t) = p_*\varphi_{f_*}(t)\sin\big((\mu_* - \mu)t\big)\,, \qquad t \in \mathbb{R}\,. \tag{5.3.2}$$

As the first moments of $f_{\mathrm{mix}}(\cdot;\vartheta)$ and $f_{\mathrm{mix}}(\cdot;\vartheta_*)$ have to coincide, we have

$$p\mu = p_*\mu_*\,, \tag{5.3.3}$$

which gives $p, \mu \neq 0$. Hence, $t = \frac{\pi}{\mu}$ is a zero of the left-hand side of (5.3.2), giving $\sin\big(\frac{\mu_* - \mu}{\mu}\pi\big) = 0$ as $p_*, \varphi_{f_*} > 0$, so that $\frac{\mu_* - \mu}{\mu} \in \mathbb{Z}$. The latter is true if and only if there is a $k \in \mathbb{Z}$ so that $\mu_* = k\mu$. By (5.3.3), we have $kp_* = p$, particularly

$$1 \leq k \leq p_*^{-1} < 2$$

because $p_* > 1/2$ and $p \in (0, 1]$. Hence, $k = 1$ and we deduce $\mu = \mu_*$, concluding the proof by Lemma 5.3.1. □

5.3.2 Proofs for the estimation results

The proofs of Theorems 4.2.16 and 4.2.19 once again will be based on Theorems 3.1.6 and 3.1.13. Hence, it is enough to show that Assumption A.2.2 in the appendix is fulfilled. Again, the complete proof is given by the assembly of the lemmata and proofs in this section.

Fix some bounded and convex open $\Theta \subset \Xi \subset (1/2, 1) \times (0, \infty) \times \mathbb{R}\backslash\{0\}$ so that

$$\overline{\Xi} = [p_-, p_+] \times [\sigma_-, \sigma_+] \times [\mu_-, \mu_+] \subset (1/2, 1) \times (0, \infty) \times \mathbb{R}\backslash\{0\}$$

for some

$$1/2 < p_- < p_+ < 1\,, \quad 0 < \sigma_- < \sigma_+ < \infty\,, \quad \mu_- < \mu_+.$$

Note that for fixed x, γ, h the contrast functions $M(\cdot, x; \gamma)$ and $M_n(\cdot, x; h)$ are defined on $[0,1] \times (0,\infty) \times \mathbb{R}$, which is a superset of $\overline{\Xi}$. In particular, the model is identifiable over $\overline{\Xi}$.

Throughout the proofs in this section, let us use the notation $a_n \lesssim b_n$ only if $a_n \leq C b_n$ for $n \geq n_0$ and C depends only on I, $\overline{\Xi}$, U, U_p, U_σ, U_μ, U_ℓ, L, $[a,b]$, $\bar{f}$ the Hölder constant function $L(\cdot)$ of $(f_x^*(\cdot))_{x\in I}$ or is universal. In particular, the constant C then is independent of specific θ, x, h, α or γ.

Let us start by proving the contrast property of the function H.

Proof of Proposition 4.2.13. The fact that for all $t \in \mathbb{R}$

$$\mathbb{E}_{\vartheta_*}\big[H(Y,t,\theta_*)\big] = 0$$

is clear as stated in (4.2.5). Now, let $\theta \in [0,1] \times (0,\infty) \times \mathbb{R}$ so that for all $t \in \mathbb{R}$

$$0 = \mathbb{E}_{\vartheta_*}\big[H(Y,t,\theta)\big] = \mathbb{E}_{\vartheta_*}\big[\sin\big((Y-\mu)\,t\big)\big] + (1-p)\,\varphi_{\bar{f}}(\sigma\,t)\,\sin(t\mu)\;.$$

By using

$$\mathbb{E}_{\vartheta_*}\big[\sin\big((Y-\mu)t\big)\big] = \Im\bigg(\int e^{it(y-\mu)} f_{\mathrm{mix}}(y;\vartheta_*)\,\mathrm{d}y\bigg) = \Im\big(\varphi_{f_{\mathrm{mix}}(\cdot+\mu;\vartheta_*)}(t)\big)$$

and

$$\Im\left(\varphi_{\frac{1}{\sigma}\bar{f}\left(\frac{\cdot+\mu}{\sigma}\right)}(t)\right) = \Im\bigg(\int e^{it(\sigma y-\mu)}\bar{f}(y)\,\mathrm{d}y\bigg) = \varphi_{\bar{f}}(\sigma t)\sin(-t\mu)\;,$$

we conclude that for all $t \in \mathbb{R}$

$$0 = \mathbb{E}_{\vartheta_*}\big[H(Y,t,\theta)\big] = \Im\left(\varphi_{f_{\mathrm{mix}}(\cdot+\mu;\vartheta_*)-\frac{1-p}{\sigma}\bar{f}\left(\frac{\cdot+\mu}{\sigma}\right)}(t)\right)\;.$$

Hence, the function

$$\tau(\cdot;\theta|\vartheta_*) := f_{\mathrm{mix}}(\cdot+\mu;\vartheta_*) - \frac{1-p}{\sigma}\bar{f}\left(\frac{\cdot+\mu}{\sigma}\right)$$

is symmetric about zero. Taking the Fourier transforms on both sides of

$$\frac{1-p}{\sigma}\bar{f}\left(\frac{\cdot}{\sigma}\right) + \tau(\cdot-\mu;\theta|\vartheta_*) = f_{\mathrm{mix}}(\cdot;\vartheta_*)$$

once again yields equation (5.3.1), i.e.

$$\begin{aligned}(1-p_*)\varphi_{\bar{f}}(\sigma_* t) - (1-p)\varphi_{\bar{f}}(\sigma t) + p_*\cos(\mu_* t)\varphi_{f_*}(t) &= p\cos(\mu t)\varphi_{\tau(\cdot;\theta|\vartheta_*)}(t)\;,\\ p_*\sin(\mu_* t)\varphi_{f_*}(t) &= p\sin(\mu t)\varphi_{\tau(\cdot;\theta|\vartheta_*)}(t)\;.\end{aligned}$$

Multiplying the first equation by $\sin(\mu t)$ and the second one by $\cos(\mu t)$ once again gives

$$\big[(1-p_*)\varphi_{\bar{f}}(\sigma_* t) - (1-p)\varphi_{\bar{f}}(\sigma t)\big]\sin(\mu t) = p_*\varphi_{f_*}(t)\sin\big((\mu_*-\mu)t\big)\;.$$

As Assumption 4.2.3 is fulfilled we can repeat the proof of Theorem 4.2.4 starting after (5.3.2). Note that we cannot use Theorem 4.2.4 to confirm the result because $\tau(\cdot;\theta|\vartheta_*)$ does not have to be a density. □

The same procedure works under Assumption 4.2.9.

Auxiliary results

Let us examine the contrast functions M and M_n on a basic level. For theoretical purposes, we first deduce an alternative form of H by

$$\begin{aligned}
&\mathbb{E}_\gamma\Big[\sin\big((Y-\mu)t\big)\Big|X=x\Big]\\
&=\int \Im\Big(\exp\big(i(y-\mu)t\big)\Big)\Big(\frac{1-p_*(x)}{\sigma_*(x)}\bar f\Big(\frac{y}{\sigma_*(x)}\Big)+p_*(x)f_x^*\big(y-\mu_*(x)\big)\Big)\,\mathrm{d}y\\
&=\big(1-p_*(x)\big)\sin(-\mu t)\varphi_{\bar f}\big(\sigma_*(x)t\big)+p_*(x)\sin\big((\mu_*(x)-\mu)t\big)\varphi_{f_x^*}(t)\,, \qquad (5.3.4)
\end{aligned}$$

so that

$$\begin{aligned}
&\mathbb{E}_\gamma\Big[H(Y,t,\theta)\Big|X=x\Big]\\
&=\sin(\mu t)\Big((1-p)\varphi_{\bar f}(\sigma t)-\big(1-p_*(x)\big)\varphi_{\bar f}\big(\sigma_*(x)t\big)\Big)+p_*(x)\sin\big((\mu_*(x)-\mu_*)t\big)\varphi_{f_x^*}(t)
\end{aligned} \qquad (5.3.5)$$

yielding an alternative form of M by definition.

Let us use representation (5.3.5) in order to differentiate the function H, directly yielding the derivatives of the contrast M by differentiating under the integral sign, i.e.

$$\begin{aligned}
\partial_\theta M(\theta,x;\gamma) =&2\int \mathbb{E}_\gamma\Big[H(Y,t,\theta)\Big|X=x\Big]\cdot\mathbb{E}_\gamma\Big[\partial_\theta H(Y,t,\theta)\Big|X=x\Big]q(t)\,\mathrm{d}t\cdot\ell^2(x)\,,\\
V_x(\theta;\gamma) =&2\int\Big(\mathbb{E}_\gamma\Big[\partial_\theta H(Y,t,\theta)\Big|X=x\Big]^T\mathbb{E}_\gamma\Big[\partial_\theta H(Y,t,\theta)\Big|X=x\Big]\\
&+\mathbb{E}_\gamma\Big[H(Y,t,\theta)\Big|X=x\Big]\cdot\mathbb{E}_\gamma\Big[\partial^2_{\theta^2}H(Y,t,\theta)\Big|X=x\Big]\Big)q(t)\,\mathrm{d}t\cdot\ell^2(x)\,,
\end{aligned}$$

where $V_x(\theta;\gamma)$ denotes the Hessian matrix of $M(\cdot,x;\gamma)$ evaluated at θ.

The derivatives of the function H are easily computed and given by

$$\begin{aligned}
\partial_p H(y,t,\theta) &= -\varphi_{\bar f}(\sigma t)\sin(\mu t)\,,\\
\partial_\sigma H(y,t,\theta) &= t(1-p)\partial\varphi_{\bar f}(\sigma t)\sin(\mu t)\,,\\
\partial_\mu H(y,t,\theta) &= -t\cos\big((y-\mu)t\big)+t(1-p)\varphi_{\bar f}(\sigma t)\cos(\mu t)\,,\\
\mathbb{E}_\gamma\Big[\partial_p H(Y,t,\theta)\Big|X=x\Big] &= -\varphi_{\bar f}(\sigma t)\sin(\mu t)\,, \qquad (5.3.6)\\
\mathbb{E}_\gamma\Big[\partial_\sigma H(Y,t,\theta)\Big|X=x\Big] &= t(1-p)\partial\varphi_{\bar f}(\sigma t)\sin(\mu t)\,,\\
\mathbb{E}_\gamma\Big[\partial_\mu H(Y,t,\theta)\Big|X=x\Big] &= -t\,\mathbb{E}_\gamma\Big[\cos\big((Y-\mu)t\big)\Big|X=x\Big]+t(1-p)\varphi_{\bar f}(\sigma t)\cos(\mu t)\\
&= t\cos(\mu t)\Big((1-p)\varphi_{\bar f}(\sigma t)-\big(1-p_*(x)\big)\varphi_{\bar f}\big(\sigma_*(x)t\big)\Big)\\
&\quad -p_*(x)t\varphi_{f_x^*}(t)\cos\big((\mu_*(x)-\mu)t\big)\,, \qquad (5.3.7)
\end{aligned}$$

$$
\begin{aligned}
\mathbb{E}_\gamma\Big[\partial_{p^2}^2 H(Y,t,\theta)\Big|X=x\Big] &= 0\,,\\
\mathbb{E}_\gamma\Big[\partial_p\partial_\sigma H(Y,t,\theta)\Big|X=x\Big] &= -\,t\partial\varphi_{\bar f}(\sigma t)\sin(\mu t)\,,\\
\mathbb{E}_\gamma\Big[\partial_p\partial_\mu H(Y,t,\theta)\Big|X=x\Big] &= -\,t\varphi_{\bar f}(\sigma t)\cos(\mu t)\,,\\
\mathbb{E}_\gamma\Big[\partial_{\sigma^2}^2 H(Y,t,\theta)\Big|X=x\Big] &= t^2(1-p)\partial^2\varphi_{\bar f}(\sigma t)\sin(\mu t)\,,\\
\mathbb{E}_\gamma\Big[\partial_\sigma\partial_\mu H(Y,t,\theta)\Big|X=x\Big] &= t^2(1-p)\partial\varphi_{\bar f}(\sigma t)\cos(\mu t)\,,\\
\mathbb{E}_\gamma\Big[\partial_{\mu^2}^2 H(Y,t,\theta)\Big|X=x\Big] &= -\,t^2\mathbb{E}_\gamma\Big[\sin\big((Y-\mu)t\big)\Big|X=x\Big] - t^2(1-p)\varphi_{\bar f}(\sigma t)\sin(\mu t)\,.
\end{aligned}
$$

The following lemma gives basic results on the function H.

Lemma 5.3.2. *Under Assumptions* **(R1)** *and* **(R2)**, *there is a constant $C>0$ depending only on $\overline{\Xi}$ and $\bar f$ so that for all $t\in\mathbb{R}$, $\theta,\tilde\theta\in\overline{\Xi}$ we have*

(i) $\sup_{y,t\in\mathbb{R}}\sup_{\theta\in\overline{\Xi}}|H(y,t,\theta)|\le C$,

(ii) $\sup_{y\in\mathbb{R}}\sup_{\theta\in\overline{\Xi}}\big\|\partial_\theta H(y,t,\theta)\big\|\le C\big(1+|t|\big)$,

(iii) $\sup_{y\in\mathbb{R}}\sup_{\theta\in\overline{\Xi}}\big\|\partial_{\theta^2}^2 H(y,t,\theta)\big\|\le C\big(1+t^2\big)$,

(iv) $\sup_{y\in\mathbb{R}}\big|H(y,t,\theta)-H(y,t,\tilde\theta)\big|\le C\big(1+|t|\big)\,\|\theta-\tilde\theta\|$,

(v) $\sup_{y\in\mathbb{R}}\big\|\partial_\theta H(y,t,\theta)-\partial_{\tilde\theta}H(y,t,\tilde\theta)\big\|\le C\big(1+t^2\big)\|\theta-\tilde\theta\|$,

(vi) $\sup_{y\in\mathbb{R}}\big\|\partial_{\theta^2}^2 H(y,t,\theta)-\partial_{\tilde\theta^2}^2 H(y,t,\tilde\theta)\big\|\le C\big(1+|t|^3\big)\|\theta-\tilde\theta\|$.

Lemma 5.3.3. *Let $0<a\le b<\infty$, K be a kernel fulfilling Assumptions* **(R3)** *and* **($\widetilde{\mathbf{R}}$4)**. *Then, under Assumptions* **(R1)**, **(R2)**, *for any compact $J\subset\operatorname{int}(I)$, there is some constant $C>0$ so that*

$$
\begin{aligned}
&\sup_{*}\ h^{-\alpha}\sup_{x\in J,\theta\in\Theta}\Big|\mathbb{E}_\gamma\big[H(Y,t,\theta)\big|X=x\big]-\Big(\mathbb{E}_\gamma\big[H(Y,t,\theta)\big|X=\cdot\big]*K_h\Big)(x)\Big|\le C\big(1+|t|\big)\,,\\
&\sup_{*}\ h^{-\alpha}\sup_{x\in J,\theta\in\Theta}\Big|\mathbb{E}_\gamma\big[\partial_\theta H(Y,t,\theta)\big|X=x\big]-\Big(\mathbb{E}_\gamma\big[\partial_\theta H(Y,t,\theta)\big|X=\cdot\big]*K_h\Big)(x)\Big|\le C\big(1+t^2\big)\,,
\end{aligned}
$$

where the suprema are taken over $\alpha\in[a,b]$, $\gamma\in\Gamma(\alpha)$, $h\in(0,\infty)$.

Main proofs

The proofs of Theorems 4.2.16 and 4.2.19 once again will be based on Theorems 3.1.6 and 3.1.13. Hence, it is enough to show that Assumption A.2.2 in the appendix is fulfilled.

Lemma 5.3.4. *Let $0<a\le b<\infty$. Under Assumptions* **(R1)**, **(R2)**, **(R5)**, *Conditions (i)* **($\widetilde{\mathbf{B}}$3)** *and (ii)* **($\widetilde{\mathbf{B}}$4)** *hold for any compact $J\subset\operatorname{int}(I)$. That is:*

(i) For all $x\in J$, $\alpha\in[a,b]$, $\gamma\in\Gamma(\alpha)$, the matrix $V_x(\theta_*(x);\gamma)$ is positive definite.

(ii) The Hessian matrices V_x are uniformly Lipschitz continuous in θ, i.e. for all $\theta, \theta' \in \overline{\overline{\Xi}}$, we have

$$\sup_{\alpha\in[a,b]}\sup_{\gamma\in\Gamma(\alpha)}\sup_{x\in J}\|V_x(\theta;\gamma)-V_x(\theta';\gamma)\| \le C\|\theta-\theta'\|_1 ,$$

where C depends only on $\overline{\overline{\Xi}}$, I and q.

Proof of Lemma 5.3.4. (i) Let us start by showing that for all $x \in J$, the Hessian matrix $V_x(\theta_*(x);\gamma)$ is positive definite.

Because $\mathbb{E}_\gamma\big[H(Y,t,\theta_*(x))\big|X=x\big]=0$ for all t, the Hessian matrix $V_x(\theta_*(x);\gamma)$ reduces to

$$\begin{aligned}&V_x(\theta_*(x);\gamma)\\ &=2\int \mathbb{E}_\gamma\Big[\partial_\theta H\big(Y,t,\theta_*(x)\big)\Big|X=x\Big]^T\mathbb{E}_\gamma\Big[\partial_\theta H\big(Y,t,\theta_*(x)\big)\Big|X=x\Big]q(t)\,\mathrm{d}t\cdot\ell^2(x) .\end{aligned}$$

When inserting the true parameter $\theta_*(x)$, the derivatives (5.3.6) - (5.3.7) reduce to

$$\begin{aligned}\mathbb{E}_\gamma\Big[\partial_p H\big(Y,t,\theta_*(x)\big)\Big|X=x\Big] &= -\varphi_{\bar f}\big(\sigma_*(x)t\big)\sin\big(\mu_*(x)t\big) ,\\ \mathbb{E}_\gamma\Big[\partial_\sigma H\big(Y,t,\theta_*(x)\big)\Big|X=x\Big] &= t\big(1-p_*(x)\big)\partial\varphi_{\bar f}\big(\sigma_*(x)t\big)\sin\big(\mu_*(x)t\big) ,\\ \mathbb{E}_\gamma\Big[\partial_\mu H\big(Y,t,\theta_*(x)\big)\Big|X=x\Big] &= -tp_*(x)\varphi_{f_x^*}(t) .\end{aligned}$$

Since $M(\cdot,x;\gamma)$ attains a minimum at $\theta_*(x)$, the Hessian matrix $V_x(\theta_*(x);\gamma)$ is positive semidefinite. So assume there is a $v^T=(v_1,v_2,v_3)\in\mathbb{R}^3$ so that

$$0=v^TV_x(\theta_*(x);\gamma)v=2\int\Big(\mathbb{E}_\gamma\Big[\partial_\theta H\big(Y,t,\theta_*(x)\big)\Big|X=x\Big]v\Big)^2q(t)\,\mathrm{d}t\ell^2(x) .$$

Since $q,\ell>0$ and the function $t\mapsto\mathbb{E}_\gamma\Big[\partial_\theta H(Y,t,\theta_*(x))\Big|X=x\Big]$ is continuous, we conclude

$$\begin{aligned}0=&\,v_1\mathbb{E}_\gamma\Big[\partial_p H\big(Y,t,\theta_*(x)\big)\Big|X=x\Big]+v_2\mathbb{E}_\gamma\Big[\partial_\sigma H\big(Y,t,\theta_*(x)\big)\Big|X=x\Big]\\ &+v_3\mathbb{E}_\gamma\Big[\partial_\mu H\big(Y,t,\theta_*(x)\big)\Big|X=x\Big]\\ =&-v_1\varphi_{\bar f}\big(\sigma_*(x)t\big)\sin\big(\mu_*(x)t\big)+v_2t\big(1-p_*(x)\big)\partial\varphi_{\bar f}\big(\sigma_*(x)t\big)\sin\big(\mu_*(x)t\big)-v_3tp_*(x)\varphi_{f_x^*}(t)\\ =:&\,g(t)\end{aligned}\tag{5.3.8}$$

for all $t\in\mathbb{R}$. It remains to show that $v=0$.

First note that the first and second summand in (5.3.8) are zero for $t\in\frac{\pi}{\mu_*(x)}\mathbb{Z}$. Hence, we have $v_3=0$ as $\varphi_{f_x^*},p_*(x)>0$. Since g is zero on $\mathbb{R}$, so is its first derivative, which

exists as $\bar{f}$ and f_x^* have finite third moments. Now let us differentiate g at $t = 0$. The derivative is determined by

$$\begin{aligned}
\partial_t\Big(- \varphi_{\bar{f}}\big(\sigma_*(x)t\big)\sin\big(\mu_*(x)t\big)\Big)\Big|_{t=0} &= -\mu_*(x)\varphi_{\bar{f}}(0)\cos(0) = -\mu_*(x)\ , \\
\partial_t\Big(t\big(1-p_*(x)\big)\partial\varphi_{\bar{f}}\big(\sigma_*(x)t\big)\sin\big(\mu_*(x)t\big)\Big)\Big|_{t=0} &= 0\ , \\
\partial_t\Big(- tp_*(x)\varphi_{f_x^*}(t)\Big)\Big|_{t=0} &= -p_*(x)\ ,
\end{aligned}$$

giving

$$v_1 = -\frac{p_*(x)}{\mu_*(x)}v_3\ ,$$

because $\mu_*(x), p_*(x) \neq 0$. Minding $v_3 = 0$, we derive $v_1 = 0$. And since the function

$$t \mapsto t\big(1-p_*(x)\big)\partial\varphi_{\bar{f}}\big(\sigma_*(x)t\big)\sin\big(\mu_*(x)t\big)$$

is non-zero in a neighbourhood around 0 excluding 0, we get $v_2 = 0$ by (5.3.8), so that the matrix $V_x(\theta_*(x);\gamma)$ is indeed positive definite.

Note that under the identifiability Assumption 4.2.9, the arguments would work analogously.

(ii) By using the identity

$$\begin{aligned}
V_x(\theta;\gamma) =& 2\int\Big(\mathbb{E}_\gamma\Big[\partial_\theta H(Y,t,\theta)\Big|X=x\Big]^T\mathbb{E}_\gamma\Big[\partial_\theta H(Y,t,\theta)\Big|X=x\Big] \\
&+ \mathbb{E}_\gamma\Big[H(Y,t,\theta)\Big|X=x\Big]\cdot\mathbb{E}_\gamma\Big[\partial_{\theta^2}^2 H(Y,t,\theta)\Big|X=x\Big]\Big)q(t)\,\mathrm{d}t\cdot\ell^2(x)\ ,
\end{aligned}$$

Lemma 5.3.2 and minding the fact, that q has finite moments of order up to 3, we deduce the result.

□

The following lemma shows that the deterministic and stochastic estimation errors of the empirical contrast are of the usual non-parametric order.

Lemma 5.3.5. *Let $0 < a \leq b < \infty$. Under Assumptions* **(R1)**, **(R2)**, **(R5)**, *for some kernel K fulfilling Assumptions* **(R3)** *and* **($\widetilde{\mathbf{R}}$4)** *and sequences of bandwidth parameters $h_n(\alpha)$, $\alpha \in [a,b]$ so that*

$$\sup_{\alpha\in[a,b]} h_n(\alpha), \quad \sup_{\alpha\in[a,b]}\frac{\log n}{nh_n(\alpha)^d} \to 0\ ,$$

Conditions **($\widetilde{\mathbf{B}}$5)**, **($\widetilde{\mathbf{B}}$6)** *and* **($\widetilde{\mathbf{B}}$7)** *hold for any compact cuboid $J \subset \mathrm{int}(I)$ containing an open subset. To be specific on* **($\widetilde{\mathbf{B}}$5)**, *we have for any compact cuboid $J \subset \mathrm{int}(I)$ containing an open subset*

$$\limsup_{n\to\infty}\sup_{\alpha\in[a,b]}\sup_{\gamma\in\Gamma(\alpha)}\sup_{x\in J,\theta\in\Theta} h_n(\alpha)^{-\alpha}\big\|\mathbb{E}_\gamma\big[S_n\big(\theta,x;h_n(\alpha)\big)\big] - S(\theta,x;\gamma)\big\| \leq C_*\ , \tag{5.3.9}$$

$$\limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(a)} \left(\frac{\log n}{nh_n(\alpha)^d}\right)^{-1} \mathbb{E}_\gamma\Big[\sup_{x\in J,\theta\in\Theta} \big\|S_n\big(\theta,x;h_n(\alpha)\big) - \mathbb{E}_\gamma\big[S_n\big(\theta,x;h_n(\alpha)\big)\big]\big\|^2\Big] \tag{5.3.10}$$

$$\leq C_{\mathrm{STOCH}} \,.$$

The constant $C_* > 0$ depends only on a, b, the function classes $\Gamma(\alpha)$, Θ, I, q and K; the constant $C_{\mathrm{STOCH}} > 0$ depends only on $\|K\|_\infty$, L_K, U_ℓ, I, Θ but is free from a and b.

Particularly, when $h_n(\alpha) = \big(\frac{\log n}{n}\big)^{\frac{1}{2\alpha+d}}$, there is a constant $C > 0$ so that

$$\limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \left(\frac{\log n}{n}\right)^{-\frac{2\alpha}{2\alpha+d}} \mathbb{E}_\gamma\Big[\sup_{x\in J,\theta\in\Theta} \big\|S_n\big(\theta,x;h_n(\alpha)\big) - S(\theta,x;\gamma)\big\|^2\Big] \leq C \,. \tag{5.3.11}$$

Proof of Lemma 5.3.5. First, let us prove (5.3.9). We will show that for all θ, x, α, γ, h, we have

$$\begin{aligned}
&\big\|\mathbb{E}_\gamma\big[S_n(\theta,x;h)\big] - S(\theta,x;\gamma)\big\| \\
&\leq 2\int \Big\|\Big(\big(\mathbb{E}_\gamma\big[H(Y,t,\theta)\big|X=\cdot\big]\ell\big) * K_h\Big)(x) \cdot \Big(\Big(\mathbb{E}_\gamma\big[\partial_\theta H(Y,t,\theta)\big|X=\cdot\big]\ell\Big) * K_h\Big)(x) \\
&\qquad - \ell^2(x)\mathbb{E}_\gamma\big[H(Y,t,\theta)\big|X=x\big]\cdot\mathbb{E}_\gamma\big[\partial_\theta H(Y,t,\theta)\big|X=x\big]\Big\| q(t)\,\mathrm{d}t \,. \qquad (5.3.12)\\
&\lesssim h^\alpha
\end{aligned}$$

Let us make a zero addition of the term

$$\ell(x)\mathbb{E}_\gamma\big[H(Y,t,\theta)\big|X=x\big]\cdot\Big(\Big(\mathbb{E}_\gamma\big[\partial_\theta H(Y,t,\theta)\big|X=\cdot\big]\ell\Big) * K_h\Big)(x)$$

within the norm in (5.3.12). Since ℓ is bounded by $\sup U_\ell$ and the functions H and $\partial_\theta H(\cdot,t,\cdot)/(1+|t|)$ are uniformly bounded according to Lemma 5.3.2 (i) and (ii), it is enough to examine occurring differences.

First deduce that

$$\begin{aligned}
&\Big|\Big(\big(\mathbb{E}_\gamma\big[H(Y,t,\theta)\big|X=\cdot\big]\ell\big) * K_h\Big)(x) - \ell(x)\mathbb{E}_\gamma\big[H(Y,t,\theta)\big|X=x\big]\Big| \qquad (5.3.13)\\
&\leq \Big|\Big(\big(\mathbb{E}_\gamma\big[H(Y,t,\theta)\big|X=\cdot\big]\ell\big) * K_h\Big)(x) - \ell(x)\Big(\big(\mathbb{E}_\gamma\big[H(Y,t,\theta)\big|X=\cdot\big]\big) * K_h\Big)(x)\Big| \\
&\quad + \Big|\ell(x)\Big(\big(\mathbb{E}_\gamma\big[H(Y,t,\theta)\big|X=\cdot\big]\big) * K_h\Big)(x) - \ell(x)\mathbb{E}_\gamma\big[H(Y,t,\theta)\big|X=x\big]\Big| \,,
\end{aligned}$$

where the first summand is treated by Lemma 5.3.2 (i) and the fact that ℓ is Hölder-α-smooth, in particular,

$$\big|\big(\ell * K_h\big)(x) - \ell(x)\big| \lesssim h^\alpha \,.$$

The second summand is dealt with by Lemma 5.3.3 directly so that in conclusion

$$(5.3.13) \lesssim h^{\alpha}\big(1+|t|\big) \ .$$

Analogously, we derive that

$$\begin{aligned}
&\left\|\Big(\big(\mathbb{E}_{\gamma}\big[\partial_{\theta}H(Y,t,\theta)\big|X=\cdot\big]\ell\big)*K_h\Big)(x)-\ell(x)\mathbb{E}_{\gamma}\big[\partial_{\theta}H(Y,t,\theta)\big|X=x\big]\right\| \\
\leq&\left\|\Big(\big(\mathbb{E}_{\gamma}\big[\partial_{\theta}H(Y,t,\theta)\big|X=\cdot\big]\ell\big)*K_h\Big)(x)-\ell(x)\Big(\big(\mathbb{E}_{\gamma}\big[\partial_{\theta}H(Y,t,\theta)\big|X=\cdot\big]\big)*K_h\Big)(x)\right\| \\
&+\left\|\ell(x)\Big(\big(\mathbb{E}_{\gamma}\big[\partial_{\theta}H(Y,t,\theta)\big|X=\cdot\big]\big)*K_h\Big)(x)-\ell(x)\mathbb{E}_{\gamma}\big[\partial_{\theta}H(Y,t,\theta)\big|X=x\big]\right\| \\
\lesssim& h^{\alpha}\big(1+t^2\big) \qquad (5.3.14)
\end{aligned}$$

and since q has finite third moments, conclude the bias examination.

In order to prove (5.3.10) and **$(\widetilde{\mathbf{B}}\mathbf{7})$**, we only need to show that all assumptions of Theorem 3.2.7 are fulfilled. The gradient of the empirical contrast M_n is given by

$$S_n(\theta,x;h)=\frac{2}{n(n-1)}\sum_{\substack{j,k=1\\ j\neq k}}^{n}\int H(Y_j,t,\theta)\partial_{\theta}H(Y_k,t,\theta)q(t)\,\mathrm{d}tK_h(X_j-x)K_h(X_k-x) \ .$$

According to Lemma 5.3.2 (i), (ii), (iv) and (v), each of the coordinates of the function

$$\theta\mapsto\int H(Y_j,t,\theta)\partial_{\theta}H(Y_k,t,\theta)q(t)\,\mathrm{d}t$$

fulfils all of the assumptions postulated on the function τ in Theorem 3.2.7, so that application of Theorem 3.2.7 concludes the proof of **$(\widetilde{\mathbf{B}}\mathbf{5})$**.

$(\widetilde{\mathbf{B}}\mathbf{7})$ is subsequently given directly by Lemma 3.2.10.

In order to prove **$(\widetilde{\mathbf{B}}\mathbf{6})$**, make a bias variance decomposition for the contrast's estimation error $\sup_{\theta\in\Theta,x\in J}|M_n(\theta,x;h)-M(\theta,x;\gamma)|$ and repeat the arguments above. In particular, by using Lemma 5.3.3 as well as Lemma 3.2.1, deduce that

$$\sup_{\alpha\in[a,b]}\sup_{\gamma\in\Gamma(a)}h_n(\alpha)^{-a}\sup_{\theta\in\Theta,x\in J}\Big|\mathbb{E}_{\gamma}\big[M_n\big(\theta,x;h_n(\alpha)\big)\big]-M(\theta,x;\gamma)\Big|=O(1)$$

yielding

$$\sup_{\alpha\in[a,b]}\sup_{\gamma\in\Gamma(a)}\sup_{\theta\in\Theta,x\in J}\Big|\mathbb{E}_{\gamma}\big[M_n\big(\theta,x;h_n(\alpha)\big)\big]-M(\theta,x;\gamma)\Big|=o(1) \ .$$

Use Theorem 3.2.7 for the empirical contrast M_n in order to deduce

$$\sup_{\alpha\in[a,b]}\sup_{\gamma\in\Gamma(a)}\left(\frac{\log n}{nh_n(\alpha)^d}\right)^{-\frac{1}{2}}\mathbb{E}_{\gamma}\left[\sup_{\theta\in\Theta,x\in J}\Big|M_n\big(\theta,x;h_n(\alpha)\big)-\mathbb{E}_{\gamma}\big[M_n\big(\theta,x;h_n(\alpha)\big)\big]\Big|\right]=O(1) \ ,$$

which gives

$$\sup_{\alpha\in[a,b]}\sup_{\gamma\in\Gamma(a)}\mathbb{P}_\gamma\Bigg(\sup_{\theta\in\Theta,x\in J}\Big|M_n\big(\theta,x;h_n(\alpha)\big)-\mathbb{E}_\gamma\big[M_n\big(\theta,x;h_n(\alpha)\big)\big]\Big|\geq\varepsilon\Bigg)=o(1)\;,\quad\varepsilon>0$$

by Markov's inequality. □

5.3.3 Proofs for auxiliary results

Proof of Lemma 5.3.2. The proofs of (i)-(iii) are deduced from the fact that the functions sin, cos, $\varphi_{\bar f}$, $\partial\varphi_{\bar f}$ and $\partial^2\varphi_{\bar f}$ are bounded.

For (iv)-(vi), we additionally use the Lipschitz continuity of sin, cos, $\varphi_{\bar f}$, $\partial\varphi_{\bar f}$ and $\partial^2\varphi_{\bar f}$. In particular, the Lipschitz continuity of $t\mapsto\exp(it)$ with Lipschitz constant 1 yields

$$\begin{aligned}\big|\partial^k\varphi_{\bar f}(\sigma t)-\partial^k\varphi_{\bar f}(\sigma' t)\big|&\leq\int\big|\exp(i\sigma ty)-\exp(i\sigma' ty)\big|\cdot\big|i^k y^k\bar f(y)\big|\mathrm{d}y\\&\leq|t|\,|\sigma-\sigma'|\int|y|^{k+1}\bar f(y)\,\mathrm{d}y\;,\quad k=0,1,2\;.\end{aligned}\tag{5.3.15}$$

□

Proof of Lemma 5.3.3. (i) Fix some $\alpha\in[a,b]$, $\gamma\in\Gamma(\alpha)$, $h\in(0,\infty)$, $x\in J$. Use the alternative representation (5.3.4), the fact that characteristic functions are bounded by 1, boundedness and Lipschitz continuity of sin, cos, the compactness of $\overline{\Xi}$ and (5.3.15) for $k=0$ in order to deduce that there is some $\bar C>0$ independent of α, γ, h, x, t so that

$$\begin{aligned}&\Big|\mathbb{E}_\gamma\big[H(Y,t,\theta)\big|X=x\big]-\Big(\mathbb{E}_\gamma\big[H(Y,t,\theta)\big|X=\cdot\big]*K_h\Big)(x)\Big|\\&=\Big|\int\big(\mathbb{E}_\gamma\big[H(Y,t,\theta)\big|X=x\big]-\mathbb{E}_\gamma\big[H(Y,t,\theta)\big|X=z+x\big]\big)K_h(z)\,\mathrm{d}z\Big|\\&=\Big|\int\Big(\mathbb{E}_\gamma\big[\sin\big((Y-\mu)t\big)\big|X=x\big]-\mathbb{E}_\gamma\big[\sin\big((Y-\mu)t\big)\big|X=z+x\big]\Big)K_h(z)\,\mathrm{d}z\Big|\\&=\Big|\int\Big(\sin(-\mu t)\big(1-p_*(x)\big)\varphi_{\bar f}\big(\sigma_*(x)t\big)-\sin(-\mu t)\big(1-p_*(z+x)\big)\varphi_{\bar f}\big(\sigma_*(z+x)t\big)\\&\quad+\sin\big((\mu_*(x)-\mu)t\big)p_*(x)\varphi_{f_x^*}(t)-\sin\big((\mu_*(z+x)-\mu)t\big)p_*(z+x)\varphi_{f_{z+x}^*}(t)\Big)K_h(z)\,\mathrm{d}z\Big|\\&\leq\bar C\big(1+|t|\big)\Big|\int\Big(\big(p_*(x)-p_*(z+x)\big)+\big(\sigma_*(x)-\sigma_*(z+x)\big)+\big(\mu_*(x)-\mu_*(z+x)\big)\Big)K_h(z)\,\mathrm{d}z\Big|\end{aligned}\tag{5.3.16}$$

$$+\bar C\Big|\int\Big(\varphi_{f_x^*}(t)-\varphi_{f_{z+x}^*}(t)\Big)K_h(z)\,\mathrm{d}z\Big|\;.\tag{5.3.17}$$

The term (5.3.16) is directly treated by Lemma 3.2.1. The term (5.3.17) is handled by the fact that $x \mapsto f_x^*(y)$ is Hölder-α-smooth with Hölder constant $L(y)$ that is integrable in y so that Hölder-α-smoothness extends to the family of characteristic functions $(\varphi_{f_x^*})_{x\in I}$. Note that the k-th partial derivatives of $f_\cdot^*(y)$ are bounded by $L(y)$, $|k| \le \lfloor\alpha\rfloor$ so that Lemma 3.2.1 is applicable again.

(ii) Since

$$\begin{aligned}
&\mathbb{E}_\gamma\big[\cos\big((Y-\mu)t\big)\big|X=x\big]\\
&=\int \Re\big(\exp(i(y-\mu)t)\big)\left(\frac{1-p_*(x)}{\sigma_*(x)}\bar f\Big(\frac{y}{\sigma_*(x)}\Big)+p_*(x)f_x^*(y-\mu_*(x))\right)\mathrm{d}y\\
&=\big(1-p_*(x)\big)\cos(-\mu t)\varphi_{\bar f}\big(\sigma_*(x)t\big)+p_*(x)\cos\big((\mu_*(x)-\mu)t\big)\varphi_{f_x^*}(t)\ ,
\end{aligned}$$

we can proceed just like we did in (i). □

6 Simulations

In this section, we illustrate the asymptotic results of the estimator $\hat{\theta}_n(\cdot; h)$, cf. (4.1.4), for the mixture of normal regressions model, cf. Section 4.1. Let us conduct a simulation series for two different models of this kind.

Regime 1: Consider the two-component mixture of regressions model with conditional model densities

$$f_1(\cdot|x) = \pi(x)\phi\big(\,\cdot\,\big|\mu_1(x), \sigma_1^2(x)\big) + \big(1-\pi(x)\big)\phi\big(\,\cdot\,\big|\mu_2(x), \sigma_2^2(x)\big)\ ,$$

where the parameter functions are given by

$$\begin{aligned}
\mu_1(x) &= 3 + 0.5\sin(0.2x)\ , & \mu_2(x) &= -0.5\sin(x)\ ,\\
\sigma_1(x) &= 0.9 + 0.2\sin(0.2x)\ , & \sigma_2(x) &= 0.8 + 0.2\sin(0.3x)\ ,\\
\pi(x) &= 0.3 + 0.03\sin(0.5x)\ .
\end{aligned}$$

Regime 2: Consider the three-component mixture of regressions model with conditional model densities

$$f_2(\cdot|x) = \sum_{c=1}^{3} p_c(x)\phi\big(\,\cdot\,\big|\lambda_c(x), \delta_c^2(x)\big)\ ,$$

where the parameter functions are given by

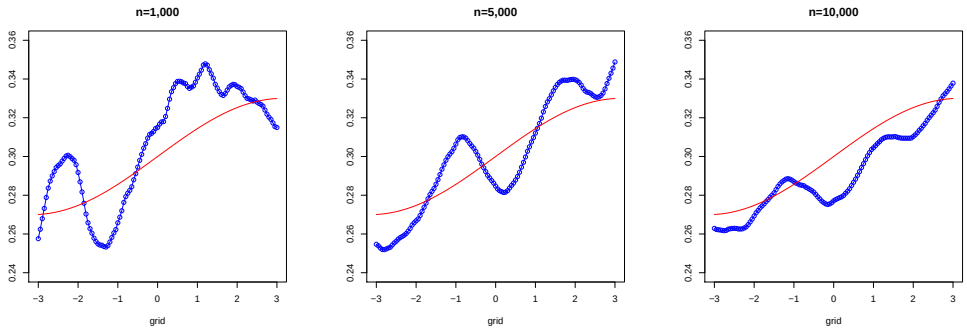
$$\begin{aligned}
\lambda_1(x) &= \mu_1(x)\ , & \lambda_2(x) &= \mu_2(x)\ , & \lambda_3(x) &= -5 + 0.25\exp(0.4x)\ ,\\
\delta_1(x) &= \sigma_1(x)\ , & \delta_2(x) &= \sigma_2(x)\ , & \delta_3(x) &= 0.9 + 0.1\exp(0.1x)\ ,\\
p_1(x) &= \pi(x)\ , & p_2(x) &= 0.25 + 0.02\cos(0.25x)\ , & p_3(x) &= 1 - p_1(x) - p_2(x)\ .
\end{aligned}$$

In both regimes, the univariate covariate values are drawn from a $\mathcal{U}(-4, 4)$ distribution. For each regime, we generate 400 datasets with $n = 1\,000$, $n = 5\,000$, $n = 10\,000$ observations to which the estimation method is applied.

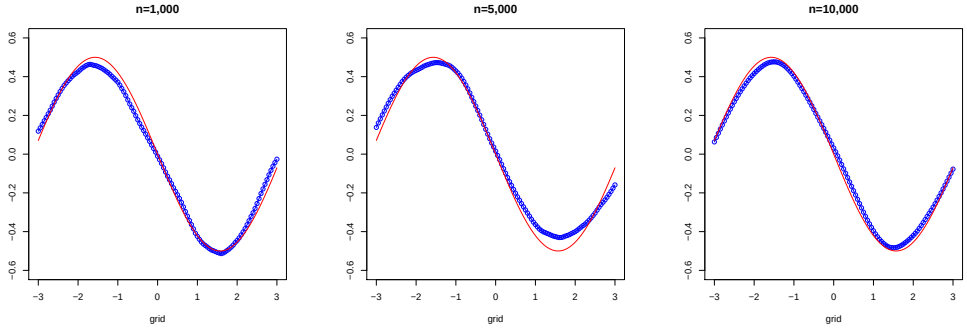
For the estimation procedure, we assume that the number of components m is known beforehand. Let us use the triangular kernel $K : \mathbb{R} \to \mathbb{R}$, $K(x) = (1 - |x|)\mathbb{1}_{|x|\leq 1}$. We propose bandwidth parameters to come from the set $\{0.1, 0.2, \ldots, 1\}$ in order to demonstrate the influence of the bandwidth on performance of the method. The uniform asymptotic results are illustrated by estimating all parameter functions on the grid $\mathcal{G} = [-3, 3] \cap (0.05 \cdot \mathbb{Z})$.

In order to approximate the maxima of the local log-likelihood functions, we apply the EM algorithm proposed by Huang et al. (2013) using perturbed true values as initial values. The estimation results for both regimes are given in the Tables A.1-A.3 in which mean and standard deviation of the estimators' supremum errors are displayed. Figure 6.2 displays boxplots for the empirical distribution of the supremum errors of the estimators $\hat{\pi}(\cdot)$, $\hat{\mu}_2(\cdot)$, $\hat{\sigma}_2(\cdot)$ in Regime 1. The influence of the bandwidth parameter on the estimation results can be seen quite clearly. Figure 6.1 displays typical estimated parameter curves.

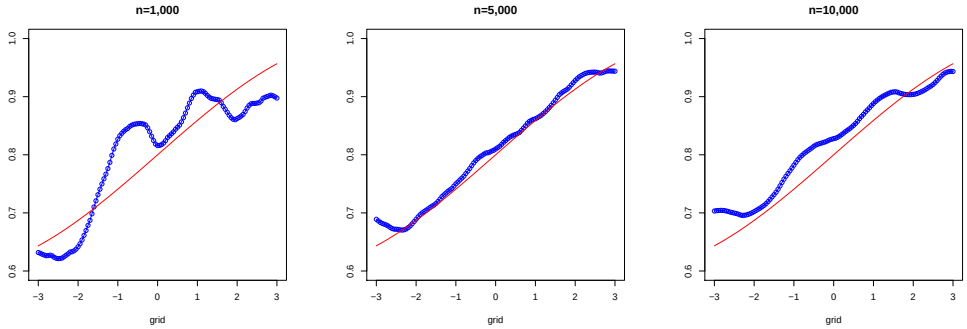
We conclude that the method works quite well for a variety of sample sizes and bandwidth parameters. In particular, the means of the empirical L^∞-errors approach zero with increasing sample size, while the standard deviations also decrease. This affirms the theoretical results stated in Section 4.1.3.



$\hat{\pi}(\cdot)$ in circled blue, $\pi(\cdot)$ in red

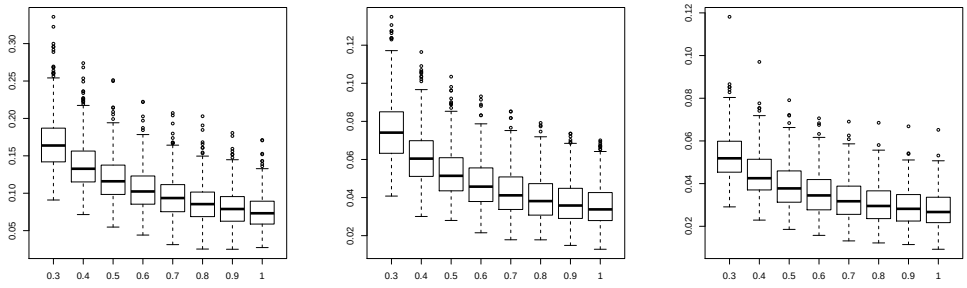


$\hat{\mu}_2(\cdot)$ in circled blue, $\mu_2(\cdot)$ in red

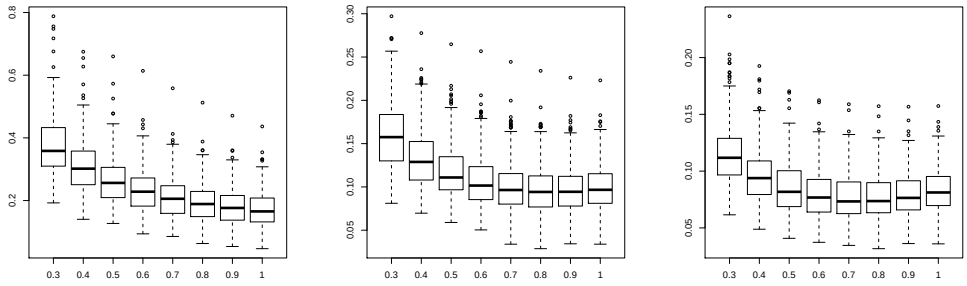


$\hat{\sigma}_2(\cdot)$ in circled blue, $\sigma_2(\cdot)$ in red

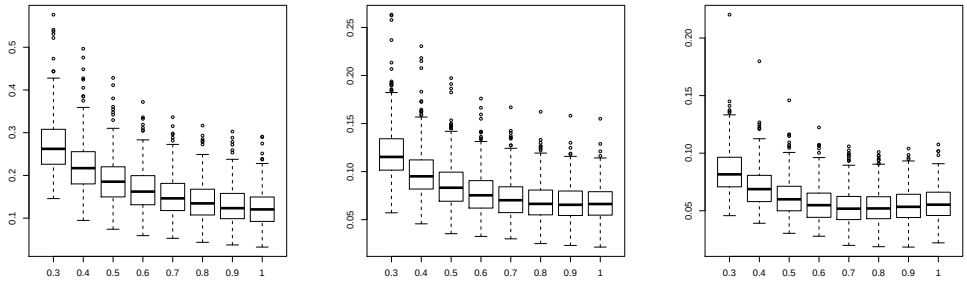
Figure 6.1: Typical estimated parameter curves (blue circled lines) of parameter curves (red solid lines) in Regime 1 with bandwidth parameter $h = 1$.



Empirical distribution of $\sup_{x \in \mathcal{G}} |\hat{\pi}(x) - \pi(x)|$



Empirical distribution of $\sup_{x \in \mathcal{G}} |\hat{\mu}_2(x) - \mu_2(x)|$



Empirical distribution of $\sup_{x \in \mathcal{G}} |\hat{\sigma}_2(x) - \sigma_2(x)|$

Figure 6.2: Boxplots of the supremum error of estimators in Regime 1 for $n = 1,000$, $n = 5,000$, $n = 10,000$ observations. In each graph, the values on the horizontal axis correspond to the bandwidth parameters h being used.

Bibliography

Bennett, G., 1962. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45.

Bordes, L. and P. Vandekerkhove, 2010. Semiparametric two-component mixture model with a known component: an asymptotically normal estimator. *Mathematical Methods of Statistics*, 19(1):22–41.

Bordes, L., C. Delmas, and P. Vandekerkhove, 2006a. Semiparametric estimation of a two-component mixture model where one component is known. *Scandinavian Journal of Statistics*, 33(4):733–752.

Bordes, L., S. Mottelet, P. Vandekerkhove, et al., 2006b. Semiparametric estimation of a two-component mixture model. *The Annals of Statistics*, 34(3):1204–1232.

Bordes, L., I. Kojadinovic, and P. Vandekerkhove, 2013. Semiparametric estimation of a mixture of two linear regressions in which one component is known. *Preprint*.

Butucea, C. and P. Vandekerkhove, 2014. Semiparametric mixtures of symmetric distributions. *Scandinavian Journal of Statistics*, 41(1):227–239.

Butucea, C., R. N. Tzoumpe, P. Vandekerkhove, et al., 2017. Semiparametric topographical mixture models with symmetric errors. *Bernoulli*, 23(2):825–862.

DeSarbo, W. S., M. Wedel, M. Vriens, and V. Ramaswamy, 1992. Latent class metric conjoint analysis. *Marketing Letters*, 3(3):273–288.

Driver, B. K., June 2003. Analysis tools with applications. *Lecture Notes*.

Fan, J., M. Farmen, and I. Gijbels, 1998. Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3):591–608.

Fisher, R. A., 1922. On the mathematical foundations of theoretical statistics. *Phil. Trans. R. Soc. Lond. A*, 222(594-604):309–368.

Giné, E. and A. Guillou, 2002. Rates of strong uniform consistency for multivariate kernel density estimators. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 38(6):907 – 921.

Giné, E., R. Latała, and J. Zinn, 2000. Exponential and moment inequalities for U-statistics. In *High Dimensional Probability II*, volume 47, pages 13–38. Springer.

Goldfeld, S. M. and R. E. Quandt, 1973. A markov model for switching regressions. *Journal of econometrics*, 1(1):3–15.

Helsen, K., K. Jedidi, and W. S. DeSarbo, 1993. A new approach to country segmentation utilizing multinational diffusion patterns. *The Journal of marketing*, 57(4):60–71.

Hohmann, D. and H. Holzmann, 2013a. Semiparametric location mixtures with distinct components. *Statistics*, 47(2):348–362.

Hohmann, D. and H. Holzmann, 2013b. Two-component mixtures with independent coordinates as conditional mixtures: Nonparametric identification and estimation. *Electronic Journal of Statistics*, 7:859–880.

Huang, M., R. Li, and S. Wang, 2013. Nonparametric mixture of regression models. *Journal of the American Statistical Association*, 108(503):929–941.

Huang, M. and W. Yao, 2012. Mixture of regression models with varying mixing proportions: a semiparametric approach. *Journal of the American Statistical Association*, 107(498):711–724.

Huber, P. J. et al., 1967. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233. Berkeley, CA.

Hunter, D. R. and D. S. Young, 2012. Semiparametric mixtures of regressions. *Journal of Nonparametric Statistics*, 24(1):19–38.

Hunter, D. R., S. Wang, and T. P. Hettmansperger, 2007. Inference for mixtures of symmetric distributions. *The Annals of Statistics*, 35(1):224–251.

Hurn, M., A. Justel, and C. P. Robert, 2003. Estimating mixtures of regressions. *Journal of Computational and Graphical Statistics*, 12(1):55–79.

Jordan, M. I. and L. Xu, 1995. Convergence results for the EM approach to mixtures of experts architectures. *Neural networks*, 8(9):1409–1431.

Kasahara, H. and K. Shimotsu, 2009. Nonparametric identification of finite mixture models of dynamic discrete choices. *Econometrica*, 77(1):135–175.

Lepskii, O., 1991. On a problem of adaptive estimation in Gaussian white noise. *Theory of Probability & Its Applications*, 35(3):454–466.

Lepskii, O., 1992. Asymptotically minimax adaptive estimation. i: Upper bounds. optimally adaptive estimates. *Theory of Probability & Its Applications*, 36(4):682–697.

Hall, P. and X.-H. Zhou, 2003. Nonparametric estimation of component distributions in a multivariate mixture. *The Annals of statistics*, 31(1):201–224.

McLachlan, G. and D. Peel, 2004. *Finite Mixture Models*. Wiley series in probability and statistics: Applied probability and statistics. Wiley.

Newcomb, S., 1886. A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics*, 8(4):343–366.

Parzen, E., 1962. On estimation of a probability density function and mode. *The Annals of mathematical statistics*, 33(3):1065–1076.

Pearson, K., 1894. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110.

Pollard, D., 2012. *Convergence of stochastic processes*. Springer Science & Business Media.

Quandt, R. E., 1958. The estimation of the parameters of a linear regression system obeying two separate regimes. *Journal of the american statistical association*, 53(284): 873–880.

Quandt, R. E., 1972. A new approach to estimating switching regressions. *Journal of the American statistical association*, 67(338):306–310.

Quandt, R. E. and J. B. Ramsey, 1978. Estimating mixtures of normal distributions and switching regressions. *Journal of the American statistical Association*, 73(364): 730–738.

Ramaswamy, V., W. S. DeSarbo, D. J. Reibstein, and W. T. Robinson, 1993. An empirical pooling approach for estimating marketing mix elasticities with pims data. *Marketing Science*, 12(1):103–124.

Rosenblatt, M., 1956. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837.

Scott, D. W., 2015. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Silverman, B. W., 1978. Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *The Annals of Statistics*, 6(1):177–184.

Silverman, B. W., 1981. Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society. Series B (Methodological)*, 43(1):97–99.

Silverman, B. W., 2018. *Density estimation for statistics and data analysis*. Routledge.

Teicher, H., 1963. Identifiability of finite mixtures. *The Annals of Mathematical statistics*, 34(4):1265–1269.

Titterington, D., A. Smith, and U. Makov, 1985. *Statistical analysis of finite mixture distributions*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley.

Tsybakov, A., 2008. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York.

van der Vaart, A. W., 2000. *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A. and J. A. Wellner, 1996. *Weak convergence and empirical processes. With applications to statistics*. New York, NY: Springer.

Vandekerkhove, P., 2013. Estimation of a semiparametric mixture of regressions model. *Journal of Nonparametric Statistics*, 25(1):181–208.

Viele, K. and B. Tong, 2002. Modeling with mixtures of linear regressions. *Statistics and Computing*, 12(4):315–330.

Werner, H., 2015. Nonparametric identification and estimation in two-component mixtures and mixtures of regressions. Master's thesis, Philipps-Universität Marburg.

Young, D. S. and D. R. Hunter, 2010. Mixtures of regressions with predictor-dependent mixing proportions. *Computational Statistics & Data Analysis*, 54(10):2253–2266.

Appendix

A.1 Notation

Let us briefly discuss notation used throughout this thesis.

We will use the notation $\|\cdot\|$ for norms on any normed space. The context should make clear on which space it lives. Sometimes, we will use specific norms, such as

$$\|\theta\|_1 = \sum_{j=1}^{m} |\vartheta_j| \,, \quad \|\theta\|_\infty = \max_{j=1,\dots,m} |\vartheta_j| \,, \quad \theta = (\vartheta_1, \dots, \vartheta_m)^T \in \mathbb{R}^m \,.$$

Note that norms on $\mathbb{R}^m$ are equivalent in the sense that for any norms $\|\cdot\|$, $\|\cdot\|_*$ on $\mathbb{R}^m$, there are constants $c_1, c_2 > 0$ so that

$$c_1\|\theta\| \leq \|\theta\|_* \leq c_2\|\theta\| \,, \quad \theta \in \mathbb{R}^m \,.$$

When using a norm $\|\cdot\|$ for matrices $A \in \mathbb{R}^{m\times m}$, we assume that it is compatible with the norm being used on $\mathbb{R}^m$, i.e.

$$\|A\theta\| \leq \|A\| \cdot \|\theta\| \,, \quad \theta \in \mathbb{R}^m \,.$$

For functions $f : \Omega \to \mathbb{R}^m$ for some non-empty set Ω, we will use the notation

$$\|f\|_\infty = \begin{cases} \sup_{\omega\in\Omega} |f(\omega)| \,, & m = 1 \,, \\ \sup_{\omega\in\Omega} \|f(\omega)\|_\infty \,, & m > 1 \,, \end{cases}$$

which is a norm on the set of bounded $\mathbb{R}^m$-valued functions on Ω.

For a vector $\theta = (\vartheta_1, \dots, \vartheta_m)^T \in \mathbb{R}^m$ and a permutation ζ on $\{1, \dots, m\}$, write

$$\zeta(\theta) = (\vartheta_{\zeta(1)}, \dots, \vartheta_{\zeta(m)})^T \,.$$

For vectors $\theta, \theta' \in \mathbb{R}^m$ denote the line segment between θ and θ' by

$$[\theta, \theta'] = \big\{\lambda\theta + (1-\lambda)\theta' \big| \lambda \in [0,1]\big\}$$

and for $\varepsilon > 0$, the ε-ball with respect to $\|\cdot\|_\infty$ around θ by

$$B_\varepsilon(\theta) = \big\{\theta' \in \mathbb{R}^m : \|\theta - \theta'\|_\infty \leq \varepsilon\big\} \,.$$

Moreover, for sets $A \subset \mathbb{R}^m$, define the ε-neighbourhood with respect to $\|\cdot\|_\infty$ around A by

$$B_\varepsilon(A) = \left\{\theta \in \mathbb{R}^m : \operatorname{dist}(\theta, A) \leq \varepsilon\right\} ,$$

where the metric that is induced by $\|\cdot\|_\infty$ is being used.

For complex numbers $z = ai + b \in \mathbb{C}$ denote the imaginary part of z by $\Im(z) = a$ and the real part of z by $\Re(z) = b$.

For a metric space (Ω, d) and sets $A, B \subset \Omega$, we use the notation

$$\begin{aligned} \operatorname{dist}(A, B) &= \inf_{x \in A, y \in B} d(x, y) , \\ \operatorname{dist}(x, B) &= \inf_{y \in B} d(x, y) , \quad x \in \Omega , \\ \operatorname{diam} B = \operatorname{diam}(B) &= \sup_{x, y \in B} d(x, y) . \end{aligned}$$

Further denote $\overline{B}$ the closure, $\operatorname{int}(I)$ the interior and ∂I the boundary of B.

For a vector $k = (k_1, \ldots, k_m) \in \mathbb{N}_0^m$ with $\sum_{j=1}^m k_j = n$ denote $|k| = n$ as well as $k! = k_1! \cdot \ldots \cdot k_m!$. For a vector $\theta = (\vartheta_1, \ldots, \vartheta_m)^T \in \mathbb{R}^m$ write $\theta^k = \vartheta_1^{k_1} \cdot \ldots \cdot \vartheta_m^{k_m}$.

For partially differentiable functions $f : U \to \mathbb{R}$, $U \subset \mathbb{R}^m$ open, denote the partial derivative of f with respect to ϑ_k at the point $\theta = (\vartheta_1, \ldots, \vartheta_m)$ by $\partial_{\vartheta_k} f(\theta)$. For a vector $z \in \mathbb{R}^m$, $\|z\| = 1$, the directional derivative along z is denoted by $\partial_z f$ if it exists. When $m = 1$, write ∂f for the derivative of f. Now, let f be n-times partially differentiable with continuous n-order partial derivatives. For a vector $k = (k_1, \ldots, k_m) \in \mathbb{N}_0^m$ with $|k| = n$ denote the k-th partial derivative of f by

$$\partial^k f(\theta) = \partial_{\vartheta_1}^{k_1} \ldots \partial_{\vartheta_m}^{k_m} f(\theta) , \quad \theta \in U ,$$

where order of differentiation is arbitrary. Furthermore, if existent, denote the gradient and Hessian matrix of f by

$$\begin{aligned} \partial_\theta f(\cdot) &= \left(\partial_{\vartheta_1} f(\cdot), \ldots, \partial_{\vartheta_m} f(\cdot)\right)^T , \\ \partial_{\theta^2}^2 f(\cdot) &= \left(\partial_{\vartheta_i} \partial_{\vartheta_j} f(\cdot)\right)_{i,j=1,\ldots,m} \in \mathbb{R}^{m \times m} . \end{aligned}$$

For an integrable function $f : \mathbb{R} \to \mathbb{R}$ denote its Fourier transform by

$$\varphi_f(t) = \int \exp(itz) f(z) \, \mathrm{d}z , \quad t \in \mathbb{R} .$$

Whenever f is a density, we shall also call φ_f the characteristic function of f.

For $\alpha \in (0, \infty)$, we define

$$\lfloor \alpha \rfloor = \max\{n \in \mathbb{N}_0 : n < \alpha\} , \quad \lceil \alpha \rceil = \min\{n \in \mathbb{N} : n > \alpha\} .$$

Note that our definition differs from the predominant one for $\lfloor \alpha \rfloor$ as we demand it to be the largest natural number that is strictly smaller than α not smaller or equal to α.

For sets Ω, $\#\Omega$ denotes the number of elements of Ω.

For sequences $(a_n)_{n\in\mathbb{N}}, (b_n)_{n\in\mathbb{N}} \subset (0,\infty)$, we use the notation

$$a_n \lesssim b_n \quad \text{or} \quad a_n = O(b_n)$$

when there is an $n_0 \in \mathbb{N}$ and a constant $C > 0$ so that for all $n \geq n_0$, we have $a_n \leq C\, b_n$. We also use the notation $a_n = o(b_n)$ when $a_n/b_n \to 0$, $n \to \infty$. For sequences of random variables $(X_n)_{n\in\mathbb{N}}$, use the notation $X_n = O_{\mathbb{P}}(b_n)$ when X_n/b_n is tight and $X_n = o_{\mathbb{P}}(b_n)$ when X_n/b_n converges to 0 in probability.

For random variables X, Y, Z we write $X \perp\!\!\!\perp Y | Z$ when X is conditionally independent of Y given Z.

Denote the density of the univariate normal distribution $\mathcal{N}(\mu, \sigma^2)$ with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$ by $\phi(\cdot|\mu, \sigma^2)$.

A.2 Sets of alternative assumptions

The following assumptions are an alteration of Assumption 3.1.4 for deriving uniform convergence rates when the model is identifiable.

Assumption A.2.1. Let $\Theta \subset \Xi \subset \mathbb{R}^m$, $I \subset \mathbb{R}^d$, $(\Gamma, \|\cdot\|)$ be a normed space, $M : \Xi \times I \times \Gamma \to \mathbb{R}$ be a deterministic function, $M_n : \Xi \times I \to \mathbb{R}$ be random functions, $(r_n)_{n\in\mathbb{N}} \subset (0,\infty)$ be sequences with $r_n \to \infty$.

($\widetilde{\mathbf{A}}$1) Assume that Θ is compact and convex with $\Theta = \overline{\operatorname{int}(\Theta)}$, Ξ is open and convex, I is compact and $(\Gamma, \|\cdot\|)$ is a compact normed space. Furthermore, there is a constant $\bar{C} < \infty$ so that for all $\theta, \theta' \in \Theta$, there is an $l \in \mathbb{N}_0$ and $\bar{\theta}_1, \dots, \bar{\theta}_l \in \Theta$ so that with $\theta = \bar{\theta}_0$, $\theta' = \bar{\theta}_{l+1}$, we have

$$\bar{\theta}_{k+1} - \bar{\theta}_k = c_k e_{j_k} , \quad k = 0, \dots, l$$

for some unit vectors e_{j_k} and some coefficients $c_k \in \mathbb{R}$ with $\sum_{k=0}^{l} |c_k| \le \bar{C}\|\theta - \theta'\|$.

($\widetilde{\mathbf{A}}$2) The function M is continuous, i.e. the map

$$(\theta, x; \gamma) \mapsto M(\theta, x; \gamma)$$

is continuous. For every $x \in I$, $\gamma \in \Gamma$, the contrast $M(\cdot, x; \gamma)$ attains a unique minimum at $\theta_*(x;\gamma)$, where $(x;\gamma) \mapsto \theta_*(x;\gamma)$ is continuous.

($\widetilde{\mathbf{A}}$3) For all $x \in I, \gamma \in \Gamma$, the function $M(\cdot, x; \gamma)$ is twice continuously differentiable in its first argument and the Hessian matrix

$$V_x\big(\theta_*(x;\gamma);\gamma\big) := \partial^2_{\theta^2} M\big(\theta_*(x;\gamma), x; \gamma\big)$$

is positive definite. Particularly, the eigenvalues $\lambda^1_{x,\gamma} \ge \dots \ge \lambda^m_{x,\gamma}$ of $V_x\big(\theta_*(x;\gamma);\gamma\big)$ are positive. Furthermore, the map $(x,\gamma) \mapsto V_x\big(\theta_*(x;\gamma);\gamma\big)$ is continuous.

($\widetilde{\mathbf{A}}$4) The Hessian matrices $V_x(\cdot;\gamma)$ are uniformly Lipschitz continuous in θ, i.e. for all $\theta, \theta' \in \Xi$, we have

$$\sup_{\gamma\in\Gamma} \sup_{x\in I} \big\|V_x(\theta;\gamma) - V_x(\theta';\gamma)\big\| \le L_{\mathrm{Hess}} \|\theta - \theta'\| ,$$

where L_{Hess} depends only on Ξ, I and Γ.

($\widetilde{\mathbf{A}}$5) The empirical contrast is continuously differentiable in its first argument and for the gradients

$$S_n(\theta, x) := \partial_\theta M_n(\theta, x) , \quad S(\theta, x; \gamma) := \partial_\theta M(\theta, x; \gamma)$$

it holds that

$$\limsup_{n\to\infty} \sup_{\gamma\in\Gamma} r_n \mathbb{E}_\gamma \Big[\sup_{x\in I, \theta\in\Theta} \big\|S_n(\theta, x; \gamma) - S(\theta, x; \gamma)\big\| \Big] < \infty .$$

($\widetilde{\mathbf{A}}$6) The empirical contrast M_n is uniformly consistent for M, i.e.

$$\lim_{n\to\infty} \sup_{\gamma\in\Gamma} \mathbb{P}_\gamma\Big(\sup_{x\in I,\theta\in\Theta} \big|M_n(\theta,x) - M(\theta,x;\gamma)\big| \geq \varepsilon \Big) = 0 \,, \quad \varepsilon > 0 \,.$$

The following list of assumptions is the simplified version of Assumption 3.1.11 for identified models.

Assumption A.2.2. Let $0 < a < b < \infty$, $\Theta \subset \Xi \subset \mathbb{R}^m$, $I \subset \mathbb{R}^d$, $(\Gamma(\alpha), \|\cdot\|_\alpha)$ be normed spaces, $\alpha \in [a,b]$, $M : \Xi \times I \times \bigcup_{\alpha\in[a,b]} \big(\Gamma(\alpha)\times\{\alpha\}\big) \to \mathbb{R}$ be a deterministic function, $M_n : \Xi \times I \times [a,b] \to \mathbb{R}$ be random functions; β_k, $r(\alpha)$, $\hat{k}$ be defined as in (3.1.6), (3.1.7) and (3.1.8), respectively; $\hat{\alpha}_n = \beta_{\hat{k}}$. Continuity of functions taking γ as arguments is to be understood with respect to the maximum of norms in the other arguments and $\|\cdot\|_a$.

($\widetilde{\mathbf{B}}$1) Assume that Θ is compact and convex with $\Theta = \overline{\mathrm{int}(\Theta)}$, Ξ is open and convex, I is compact, and $(\Gamma(\alpha), \|\cdot\|_\alpha)$ are compactly nested spaces, i.e. $\Gamma(\alpha) \subset \Gamma(\alpha')$ and $\Gamma(\alpha)$ is compact with respect to $\|\cdot\|_{\alpha'}$ whenever $\alpha' < \alpha$. Furthermore, $\Gamma(\alpha)$ is closed with respect to $\|\cdot\|_a$. Additionally, for any α, $\alpha_n \nearrow \alpha$, it holds that

$$\bigcap_{n\in\mathbb{N}} \Gamma(\alpha_n) = \Gamma(\alpha) \,.$$

Moreover, there is a constant $\bar{C} < \infty$ so that for all $\theta, \theta' \in \Theta$, there is an $l \in \mathbb{N}_0$ and $\bar{\theta}_1, \ldots, \bar{\theta}_l \in \Theta$ so that with $\theta = \bar{\theta}_0$, $\theta' = \bar{\theta}_{l+1}$, we have

$$\bar{\theta}_{k+1} - \bar{\theta}_k = c_k e_{j_k} \,, \quad k = 0, \ldots, l$$

for some unit vectors e_{j_k} and some coefficients $c_k \in \mathbb{R}$ with $\sum_{k=0}^{l} |c_k| \leq \bar{C}\|\theta - \theta'\|$.

($\widetilde{\mathbf{B}}$2) The function M is continuous, i.e. the map

$$(\theta, x; \gamma; \alpha) \mapsto M(\theta, x; \gamma; \alpha)$$

is continuous. For every $x \in I$, $\alpha \in [a,b]$, $\gamma \in \Gamma(a)$, the contrast $M(\cdot, x; \gamma; \alpha)$ attains a unique minimum at $\theta_*(x; \gamma; \alpha)$, where the map $(x; \gamma; \alpha) \mapsto \theta_*(x; \gamma; \alpha)$ is continuous.

($\widetilde{\mathbf{B}}$3) For all $x \in I$, $\alpha \in [a,b]$, $\gamma \in \Gamma(a)$, the function $M(\cdot, x; \gamma; \alpha)$ is twice continuously differentiable in its first argument and the Hessian matrix

$$V_x\big(\theta_*(x;\gamma;\alpha);\gamma;\alpha\big) := \partial^2_{\theta^2} M\big(\theta_*(x;\gamma;\alpha), x; \gamma; \alpha\big)$$

is positive definite. Particularly, the eigenvalues $\lambda^1_{x,\gamma;\alpha} \geq \cdots \geq \lambda^m_{x,\gamma;\alpha}$ of the matrices $V_x\big(\theta_*(x;\gamma;\alpha);\gamma;\alpha\big)$ are positive. Furthermore, the map

$$(x; \gamma; \alpha) \mapsto V_x(\theta_*(x; \gamma; \alpha); \gamma; \alpha)$$

is continuous.

($\widetilde{\mathbf{B}}$4) The Hessian matrices $V_x(\cdot;\gamma;\alpha)$ are uniformly Lipschitz continuous in θ, i.e. for all $\theta, \theta' \in \Xi$, we have

$$\sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(a)} \sup_{x\in I} \big\|V_x(\theta;\gamma;\alpha) - V_x(\theta';\gamma;\alpha)\big\| \le L_{\mathrm{Hess}} \|\theta - \theta'\| \,,$$

where L_{Hess} depends only on Ξ, I, a, b and $\Gamma(a)$.

($\widetilde{\mathbf{B}}$5) The empirical contrast is continuously differentiable in its first argument and for the gradients

$$S_n(\theta, x;\alpha) := \partial_\theta M_n(\theta, x;\alpha) \,, \quad S(\theta, x;\gamma;\alpha) := \partial_\theta M(\theta, x;\gamma;\alpha)$$

it holds that

$$\limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} r(\alpha)^{-2} \mathbb{E}_\gamma \Big[\sup_{x\in I,\theta\in\Theta} \big\|S_n(\theta, x;\alpha) - S(\theta, x;\gamma;\alpha)\big\|^2 \Big] \le C^{**} < \infty.$$

($\widetilde{\mathbf{B}}$6) The empirical contrast M_n is uniformly consistent for M, i.e.

$$\lim_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(a)} \mathbb{P}_\gamma \Big(\sup_{x\in I,\theta\in\Theta} \big|M_n(\theta, x;\alpha) - M(\theta, x;\gamma;a)\big| \ge \varepsilon \Big) = 0 \,, \quad \varepsilon > 0 \,.$$

($\widetilde{\mathbf{B}}$7) There is a constant $C_- > 0$ and a monotone function $u : [C_-, \infty) \to \big(1, \infty\big)$ with $u(t) \to \infty$, $t \to \infty$ so that for every $C_{\mathrm{Lep}} \ge C_-$,

$$\limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \sup_{0\le l\le j\le k_n(\alpha)} n^{u(C_{\mathrm{Lep}})} p_{lj} < \infty \,,$$

where p_{lj} is defined in (3.1.12) and $0 \le k_n(\alpha) \le N-1$ with $N = \lceil \log n \rceil$ is chosen so that $\beta_{k_n(\alpha)} \le \alpha \le \beta_{k_n(\alpha)} + 1$.

A.3 Tables

In this section, we give tables from the simulation study in Chapter 6.

Table A.1: Means and standard deviations of the supremum error of proportion estimators in both regimes. The first column gives the respective bandwidth parameter, the second row gives the respective sample size.

		$\sup_{x\in\mathcal{G}} \lvert\hat{\pi}(x) - \pi(x)\rvert$			$\sup_{x\in\mathcal{G}} \lvert\hat{p}_1(x) - p_1(x)\rvert$			$\sup_{x\in\mathcal{G}} \lvert\hat{p}_2(x) - p_2(x)\rvert$		
		1 000	5 000	10 000	1 000	5 000	10 000	1 000	5 000	10 000
0.1	mean	0.329	0.149	0.106	0.306	0.142	0.101	0.315	0.142	0.100
	sd	0.067	0.027	0.018	0.055	0.024	0.016	0.058	0.026	0.018
0.2	mean	0.220	0.099	0.070	0.205	0.093	0.066	0.210	0.094	0.066
	sd	0.050	0.022	0.013	0.037	0.018	0.012	0.042	0.019	0.013
0.3	mean	0.168	0.076	0.053	0.162	0.072	0.051	0.165	0.071	0.051
	sd	0.039	0.017	0.012	0.033	0.016	0.010	0.035	0.016	0.011
0.4	mean	0.139	0.062	0.044	0.136	0.060	0.043	0.138	0.059	0.042
	sd	0.035	0.015	0.011	0.029	0.014	0.009	0.033	0.014	0.009
0.5	mean	0.120	0.053	0.039	0.117	0.052	0.037	0.118	0.051	0.036
	sd	0.031	0.014	0.010	0.027	0.013	0.009	0.029	0.013	0.008
0.6	mean	0.106	0.048	0.035	0.103	0.046	0.033	0.104	0.045	0.032
	sd	0.029	0.013	0.009	0.026	0.012	0.008	0.027	0.012	0.008
0.7	mean	0.096	0.043	0.032	0.093	0.042	0.030	0.093	0.041	0.029
	sd	0.027	0.013	0.009	0.024	0.012	0.008	0.025	0.011	0.008
0.8	mean	0.088	0.040	0.030	0.084	0.038	0.028	0.084	0.037	0.027
	sd	0.026	0.012	0.009	0.023	0.011	0.007	0.024	0.011	0.007
0.9	mean	0.081	0.038	0.029	0.078	0.036	0.026	0.077	0.035	0.025
	sd	0.025	0.012	0.009	0.022	0.011	0.007	0.023	0.010	0.007
1	mean	0.076	0.036	0.028	0.072	0.033	0.024	0.072	0.033	0.023
	sd	0.024	0.011	0.008	0.021	0.010	0.007	0.022	0.010	0.007

Table A.2: Means and standard deviations of the supremum error of location estimators in both regimes. The first column gives the respective bandwidth parameter, the second row gives the respective sample size.

		$\sup_{x\in\mathcal{G}} \lvert\hat{\mu}_1(x)-\mu_1(x)\rvert$			$\sup_{x\in\mathcal{G}} \lvert\hat{\mu}_2(x)-\mu_2(x)\rvert$			$\sup_{x\in\mathcal{G}} \lvert\hat{\lambda}_1(x)-\lambda_1(x)\rvert$			$\sup_{x\in\mathcal{G}} \lvert\hat{\lambda}_2(x)-\lambda_2(x)\rvert$			$\sup_{x\in\mathcal{G}} \lvert\hat{\lambda}_3(x)-\lambda_3(x)\rvert$		
		1 000	5 000	10 000	1 000	5 000	10 000	1 000	5 000	10 000	1 000	5 000	10 000	1 000	5 000	10 000
0.1	mean	1.547	0.611	0.422	0.785	0.325	0.228	1.461	0.576	0.400	1.748	0.609	0.402	1.196	0.464	0.324
	sd	0.437	0.127	0.080	0.165	0.063	0.043	0.425	0.115	0.069	0.616	0.152	0.084	0.311	0.079	0.058
0.2	mean	0.919	0.395	0.276	0.494	0.209	0.149	0.873	0.377	0.264	0.992	0.382	0.259	0.716	0.297	0.211
	sd	0.233	0.098	0.057	0.115	0.043	0.031	0.202	0.083	0.054	0.325	0.096	0.060	0.166	0.057	0.041
0.3	mean	0.693	0.304	0.214	0.375	0.160	0.115	0.666	0.290	0.203	0.705	0.288	0.198	0.532	0.229	0.162
	sd	0.191	0.072	0.050	0.095	0.036	0.026	0.166	0.069	0.046	0.213	0.073	0.048	0.129	0.047	0.034
0.4	mean	0.567	0.252	0.179	0.309	0.133	0.096	0.554	0.239	0.169	0.565	0.235	0.164	0.436	0.189	0.133
	sd	0.164	0.063	0.047	0.084	0.033	0.023	0.151	0.059	0.041	0.181	0.062	0.043	0.108	0.042	0.030
0.5	mean	0.487	0.219	0.156	0.264	0.117	0.085	0.475	0.207	0.146	0.476	0.202	0.143	0.375	0.162	0.114
	sd	0.149	0.060	0.045	0.076	0.031	0.022	0.132	0.055	0.038	0.159	0.055	0.040	0.097	0.038	0.027
0.6	mean	0.430	0.196	0.142	0.232	0.107	0.079	0.416	0.183	0.130	0.416	0.180	0.131	0.330	0.143	0.100
	sd	0.138	0.057	0.043	0.069	0.030	0.021	0.119	0.051	0.036	0.141	0.050	0.038	0.090	0.036	0.025
0.7	mean	0.387	0.180	0.133	0.209	0.101	0.077	0.372	0.166	0.118	0.373	0.166	0.124	0.296	0.128	0.090
	sd	0.129	0.055	0.042	0.065	0.029	0.020	0.109	0.049	0.034	0.129	0.047	0.037	0.084	0.034	0.024
0.8	mean	0.356	0.168	0.126	0.193	0.097	0.077	0.338	0.152	0.109	0.343	0.158	0.122	0.269	0.117	0.082
	sd	0.123	0.053	0.040	0.062	0.028	0.019	0.102	0.047	0.033	0.121	0.045	0.036	0.079	0.033	0.023
0.9	mean	0.330	0.159	0.122	0.181	0.097	0.079	0.311	0.142	0.103	0.319	0.154	0.123	0.248	0.108	0.076
	sd	0.114	0.051	0.039	0.059	0.027	0.019	0.097	0.045	0.031	0.114	0.043	0.036	0.074	0.031	0.022
1	mean	0.309	0.153	0.118	0.173	0.099	0.083	0.289	0.134	0.098	0.302	0.152	0.126	0.230	0.100	0.071
	sd	0.109	0.050	0.038	0.057	0.027	0.019	0.094	0.044	0.030	0.109	0.042	0.035	0.071	0.030	0.021

Table A.3: Means and standard deviations of the supremum error of scale estimators in both regimes. The first column gives the respective bandwidth parameter, the second row gives the respective sample size.

		$\sup_{x\in\mathcal{G}} \lvert\hat{\sigma}_1(x) - \sigma_1(x)\rvert$			$\sup_{x\in\mathcal{G}} \lvert\hat{\sigma}_2(x) - \sigma_2(x)\rvert$			$\sup_{x\in\mathcal{G}} \lvert\hat{\delta}_1(x) - \delta_1(x)\rvert$			$\sup_{x\in\mathcal{G}} \lvert\hat{\delta}_2(x) - \delta_2(x)\rvert$			$\sup_{x\in\mathcal{G}} \lvert\hat{\delta}_3(x) - \delta_3(x)\rvert$		
		1 000	5 000	10 000	1 000	5 000	10 000	1 000	5 000	10 000	1 000	5 000	10 000	1 000	5 000	10 000
0.1	mean	0.863	0.411	0.291	0.537	0.239	0.168	0.845	0.406	0.283	0.962	0.493	0.331	0.775	0.338	0.237
	sd	0.145	0.076	0.053	0.103	0.049	0.033	0.138	0.084	0.047	0.311	0.134	0.083	0.135	0.058	0.041
0.2	mean	0.601	0.269	0.195	0.354	0.157	0.110	0.575	0.266	0.187	0.695	0.314	0.214	0.510	0.220	0.154
	sd	0.136	0.062	0.039	0.076	0.038	0.024	0.098	0.056	0.035	0.226	0.092	0.058	0.117	0.042	0.030
0.3	mean	0.470	0.206	0.152	0.271	0.120	0.085	0.457	0.205	0.145	0.545	0.236	0.163	0.390	0.168	0.119
	sd	0.120	0.047	0.034	0.064	0.031	0.020	0.093	0.045	0.030	0.166	0.073	0.048	0.096	0.035	0.025
0.4	mean	0.393	0.172	0.127	0.222	0.100	0.071	0.379	0.169	0.119	0.458	0.192	0.134	0.324	0.138	0.097
	sd	0.110	0.041	0.031	0.058	0.027	0.017	0.086	0.040	0.028	0.169	0.062	0.042	0.089	0.031	0.022
0.5	mean	0.342	0.150	0.111	0.190	0.087	0.062	0.326	0.145	0.103	0.395	0.164	0.115	0.278	0.119	0.083
	sd	0.102	0.037	0.028	0.053	0.025	0.016	0.080	0.037	0.026	0.157	0.054	0.037	0.079	0.028	0.021
0.6	mean	0.303	0.134	0.100	0.168	0.078	0.057	0.287	0.127	0.091	0.347	0.146	0.103	0.244	0.105	0.074
	sd	0.094	0.034	0.027	0.050	0.022	0.015	0.075	0.034	0.024	0.142	0.049	0.033	0.072	0.027	0.019
0.7	mean	0.272	0.123	0.092	0.151	0.072	0.054	0.257	0.115	0.083	0.309	0.134	0.094	0.217	0.094	0.066
	sd	0.083	0.032	0.026	0.047	0.021	0.015	0.070	0.032	0.022	0.125	0.046	0.031	0.064	0.025	0.018
0.8	mean	0.249	0.114	0.087	0.139	0.069	0.053	0.234	0.105	0.076	0.280	0.125	0.089	0.197	0.086	0.060
	sd	0.078	0.031	0.025	0.045	0.020	0.015	0.066	0.031	0.021	0.115	0.044	0.029	0.060	0.024	0.017
0.9	mean	0.229	0.107	0.082	0.130	0.067	0.054	0.216	0.097	0.071	0.257	0.118	0.086	0.180	0.079	0.056
	sd	0.073	0.030	0.024	0.044	0.019	0.015	0.063	0.029	0.020	0.109	0.042	0.027	0.057	0.023	0.016
1	mean	0.213	0.101	0.079	0.123	0.067	0.057	0.200	0.091	0.068	0.239	0.114	0.085	0.166	0.074	0.052
	sd	0.069	0.029	0.024	0.043	0.019	0.015	0.061	0.028	0.020	0.104	0.041	0.027	0.054	0.023	0.016

A.4 Proof of the alternative identifiability result

The proof for Theorem 4.2.10 is given in Werner (2015) and presented here with minor changes for completion purposes.

Proof of Theorem 4.2.10. Denote $\vartheta = (p, \sigma, \mu, f)^T$. We may repeat the proof of Theorem 4.2.4 up until (5.3.2), i.e.

$$\left[(1-p_*)\varphi_{\bar{f}}(\sigma_* t) - (1-p)\varphi_{\bar{f}}(\sigma t)\right] \sin(\mu t) = p_* \varphi_{f_*}(t) \sin\left((\mu_* - \mu)t\right) , \qquad t \in \mathbb{R} . \quad \text{(A.4.1)}$$

We further note the first moment equation

$$p\mu = p_* \mu_* , \tag{A.4.2}$$

which directly implies $p, \mu \neq 0$.

First suppose that Condition 4.2.5 holds. Assume t to be so large that $\varphi_{f_*}(t) \neq 0$ holds. Dividing (A.4.1) by $\varphi_{f_*}(t)$ and taking limits in t gives

$$\lim_{t\to\infty} p_* \sin\left((\mu_* - \mu)t\right) = 0$$

according to Condition 4.2.5. As $p_* > 0$ and sin is periodic, it follows $\mu_* = \mu$ and since $\mu_* \neq 0$, we obtain $\vartheta_* = \vartheta$ by Lemma 5.3.1.

Since for $\mu_* = \mu \neq 0$ identification follows directly by Lemma 5.3.1, we assume $\mu_* \neq \mu$ and derive a contradiction to show identification under the other conditions.

Suppose that Condition 4.2.6 holds. We need to consider three cases.

Case 1: $\sigma = \sigma_*$. If we divide (A.4.1) by $\varphi_{\bar{f}}(\sigma_* t)$ and let $t \to \infty$, the right-hand side tends to 0 and hence

$$\lim_{t\to\infty} \left((1-p_*) - (1-p)\right) \sin(\mu t) = 0 .$$

As $\mu \neq 0$, this is only possible if $p = p_*$, in which case (A.4.2) implies $\mu = \mu_*$, a contradiction.

Case 2: $\sigma < \sigma_*$. If we divide (A.4.1) by $\varphi_{\bar{f}}(\sigma t)$ and let $t \to \infty$, we obtain

$$\lim_{t\to\infty} (1-p) \sin(\mu t) = 0 .$$

It follows that $p = 1$ because $\mu \neq 0$, so that (A.4.1) reduces to

$$(1-p_*)\varphi_{\bar{f}}(\sigma_* t) \sin(\mu t) = p_* \varphi_{f_*}(t) \sin\left((\mu_* - \mu)t\right) , \quad t \in \mathbb{R} . \tag{A.4.3}$$

Dividing by $\varphi_{\bar{f}}(\sigma_* t)$ and letting $t \to \infty$ gives

$$\lim_{t\to\infty} (1-p_*) \sin(\mu t) = 0 ,$$

thus $\mu = 0$ or $p_* = 1$, a contradiction.

Case 3: $\sigma > \sigma_*$. If we divide (A.4.1) by $\varphi_{\bar{f}}(\sigma_* t)$ and let $t \to \infty$, we get

$$\lim_{t\to\infty} (1 - p_*) \sin(\mu t) = 0 \,,$$

a contradiction as above.

Now suppose that Condition 4.2.7 holds. If $\sigma_0 < \sigma, \sigma_*$, the arguments used under Condition 4.2.5 apply, while if $\sigma_0 > \sigma, \sigma_*$, so do those under Condition 4.2.6. So let us consider the following cases.

Case 1: $\sigma_* < \sigma_0 < \sigma$. We divide (A.4.1) by $\varphi_{\bar{f}}(\sigma_* t)$ and take limits to conclude

$$0 = \lim_{t\to\infty} \left(1 - p_* - (1-p)\frac{\varphi_{\bar{f}}(\sigma t)}{\varphi_{\bar{f}}(\sigma_* t)}\right) \sin(\mu t) \,,$$

giving

$$0 = \lim_{t\to\infty} 1 - p_* - (1-p)\frac{\varphi_{\bar{f}}(\sigma t)}{\varphi_{\bar{f}}(\sigma_* t)}$$

as $\mu \neq 0$ and sin is periodic. According to (4.2.3) we conclude $p_* = 1$, a contradiction.

Case 2: $\sigma < \sigma_0 < \sigma_*$, we divide (A.4.1) by $\varphi_{\bar{f}}(\sigma t)$ to deduce $p = 1$ so that (A.4.1) reduces to (A.4.3) again. Dividing by $\varphi_{f_*}(t)$ and letting $t \to \infty$ gives $\mu = \mu_*$, a contradiction.

If $c = 0$ or $c = \infty$, the cases $\sigma = \sigma_0$ or $\sigma_* = \sigma_0$ or both may be dealt with similarly. Hence, suppose that $c \notin \{0, \infty\}$.

Case 3: $\sigma_* < \sigma = \sigma_0$. Divide (A.4.1) by $\varphi_{\bar{f}}(\sigma_* t)$, giving $p_* = 1$ like above, a contradiction.

Case 4: $\sigma < \sigma_* = \sigma_0$ yields $p = 1$ by dividing by $\varphi_{\bar{f}}(\sigma t)$, which leads to a contradiction like before.

Case 5: $\sigma_* > \sigma = \sigma_0$. Divide (A.4.1) by $\varphi_{f_*}(t)$ to conclude that

$$\lim_{t\to\infty} (1-p)\frac{\varphi_{\bar{f}}(\sigma_0 t)}{\varphi_{f_*}(t)} \sin(\mu t) + p_* \sin\big((\mu_* - \mu)t\big) = 0.$$

Letting $g(t) = (1-p)c\sin(\mu t) + p_* \sin((\mu_* - \mu)t)$, an almost periodic function, we also have

$$\left|(1-p)\frac{\varphi_{\bar{f}}(\sigma_0 t)}{\varphi_{f_*}(t)} \sin(\mu t) + p_* \sin\big((\mu_* - \mu)t\big) - g(t)\right| \to 0, \quad t \to \infty,$$

so that in particular $g(t) \to 0$, which is only possible for an almost periodic function if $g(t) = 0$ for all $t \in \mathbb{R}$, so that

$$(1-p)c\sin(\mu t) + p_* \sin\big((\mu_* - \mu)t\big) = 0 \ , \quad t \in \mathbb{R} \ .$$

If $p = 1$, we get $\mu = \mu_*$, yielding $\sigma = \sigma_*$ by Lemma 5.3.1, a contradiction. Thus assume $p < 1$. Then the zeros of $\sin(\mu t)$ must coincide with those of $\sin((\mu_* - \mu)t)$, so that $\mu_* = 2\mu$ and we may cancel $\sin(\mu t)$ and find $c(1-p) + p_* = 0$, so that $c = -p_*/(1-p)$ would be negative, contrary to our assumption.

Case 6: $\sigma > \sigma_* = \sigma_0$. This case works just like case 5.

Case 7: $\sigma = \sigma_* = \sigma_0$. Dividing (A.4.1) by $\varphi_{f_*}(t)$ and arguing as above gives

$$c\sin(\mu t)\big((1-p) - (1-p_*)\big) + p_* \sin((\mu_* - \mu)t) = 0 \ , \quad t \in \mathbb{R} \ .$$

Thus, either $p = p_*$, and we conclude $\mu = \mu_*$ by (A.4.2), a contradiction, or $p \neq p_*$ and the zeros of $\sin(\mu t)$ must coincide with those of $\sin\big((\mu_* - \mu)t\big)$. This implies $\mu_* = 2\mu$ so that $2p_* = p$ by (A.4.2). Therefore, $-cp_* + p_* = 0$, so that $c = 1$, contrary to our assumption. □

A.5 Component density considerations

Let us briefly discuss how one can approach modelling the function families $(f_x^*)_{x\in I}$ by a compact function class. We will restrict our considerations to the case that $\alpha \leq 1$ but conjecture that this can be extended to higher degrees of smoothness by arguments similar to the ones in the proof of Driver (2003, Theorem 5.14).

Lemma A.5.1. *Let $I \subset \mathbb{R}^d$ be a compact cuboid containing an open subset, $g_1, g_2 : \mathbb{R} \times I \to (0,\infty)$ be continuous bounded functions so that for all $x \in I$ the map $y \mapsto g_i(y,x)$ is a density, that for all $\varepsilon > 0$, there are constants $C_\varepsilon > 0$ so that $\sup_{\|(y,x)\|\geq C_\varepsilon} g_i(y,x) \leq \varepsilon$ and that for some $c_1, c_2 > 0$ and for all x, y, we have $c_2 g_2(y,x) - c_1 g_1(y,x) > 0$. Further let $g : \mathbb{R} \to (0,1)$ be a function. Let $U \subset [0,\infty)$ be bounded, $L(\cdot)$ be a positive bounded function of y and $L^* > 0$ be a constant. Then, for any $\alpha \in (0,1]$, the function class*

$$
\begin{aligned}
&\mathcal{F}(\alpha, \bar{f}, L(\cdot), U, L^*, g_1, g_2, c_1, c_2)\\
=&\big\{ f.(\cdot) : I \times \mathbb{R} \to U \big| \forall x \in I : f_x(\cdot) \in \mathcal{E}_3, f_x(\cdot) \text{ is Lipschitz continuous with constant } L^*,\\
&\quad \varphi_{f_x} \geq g,\ \forall y \in \mathbb{R} : f.(y) \in H(\alpha, L(y), U),\\
&\quad \forall x, y : c_1 g_1(y,x) \leq f_x(y) \leq c_2 g_2(y,x) \big\} ,
\end{aligned}
$$

is compact with respect to the supremum metric.

Note that every $f.(\cdot) \in \mathcal{F}(\alpha, \bar{f}, L(\cdot), U, L^*, g_1, g_2, c_1, c_2)$ fulfils Assumption **(I3)**. The functions g_1, g_2 and g can be chosen in a way that $\mathcal{F}(\alpha, \bar{f}, L(\cdot), U, L^*, g_1, g_2, c_1, c_2)$ contains a large variety of density families. Consider for example light-tailed densities like double exponential densities for g_1 and heavy tailed densities like Student-t densities for g_2. Furthermore note that this result also holds when tailoring the function class $\mathcal{F}(\alpha, \bar{f}, L(\cdot), U, L^*, g_1, g_2, c_1, c_2)$ to Assumption 4.2.9. Minor adjustments need to be made.

Proof of Lemma A.5.1. Write $\mathcal{F}(\alpha, \bar{f}, L(\cdot), U, L^*, g_1, g_2, c_1, c_2) = \mathcal{F}$. First we note that the conditions that define the function class $\mathcal{F}$ are limit invariant, i.e. if every member of a sequence of functions $f_n \in \mathcal{F}$ fulfils those properties and if $\|f_n - f\|_\infty \to 0$ for some continuous function f, then f fulfils those properties as well. Hence, $\mathcal{F}$ is closed with respect to the supremum metric.

Furthermore, $\mathcal{F}$ is complete because it is a closed subset of the set of bounded continuous functions on $I \times \mathbb{R}$, which is complete. It remains to show that $\mathcal{F}$ is totally bounded, i.e. we need to show that for any $\varepsilon > 0$, there are open balls $M_1, \ldots, M_n$ of radius ε so that $\mathcal{F} \subset \bigcup_{i=1}^n M_i$.

Fix $\varepsilon > 0$. Then for any $\|(y,x)\| \geq C_{\frac{\varepsilon}{\max\{c_1,c_2\}}}$, we have $\max\{c_1,c_2\} g_i(y,x) \leq \varepsilon$ so that

$$
A_\varepsilon = \big\{ (y,x) : |c_2 g_2(y,x) - c_1 g_1(y,x)| \geq \varepsilon \big\}
$$

is bounded and further closed as $(y, x) \mapsto |c_2 g_2(y, x) - c_1 g_1(y, x)|$ is continuous. Thus, A_ε is compact. Since for any function $f_1, f_2 \in \mathcal{F}$, we have $\left\| f_1|_{A_\varepsilon^\complement} - f_2|_{A_\varepsilon^\complement} \right\|_\infty \leq \varepsilon$ it is enough to show that

$$\mathcal{F}\big|_{A_\varepsilon} = \left\{ f|_{A_\varepsilon} \middle| f \in \mathcal{F} \right\}$$

can be covered by such open balls M_i because then the open balls of the corresponding functions in $\mathcal{F}$ cover $\mathcal{F}$. We will prove this by showing a stronger property, i.e. that $\mathcal{F}\big|_{A_\varepsilon}$ is compact. Therefore, we use the well-known Arzelà-Ascoli theorem. The set $\mathcal{F}\big|_{A_\varepsilon}$ is closed as are $\mathcal{F}$ and A_ε. Furthermore, $\mathcal{F}\big|_{A_\varepsilon}$ is bounded, as is U. Finally, $\mathcal{F}\big|_{A_\varepsilon}$ is equicontinuous as for any $f \in \mathcal{F}\big|_{A_\varepsilon}, y_1, y_2, x_1, x_2$, we have

$$\begin{aligned} |f(y_1, x_1) - f(y_2, x_2)| &\leq |f(y_1, x_1) - f(y_1, x_2)| + |f(y_1, x_2) - f(y_2, x_2)| \\ &\leq L(y_1)\|x_1 - x_2\|^\alpha + L^*|y_1 - y_2| \\ &\leq \|L(\cdot)\|_\infty \|x_1 - x_2\|^\alpha + L^*|y_1 - y_2| \,. \end{aligned}$$

□

A.6 Zusammenfassung auf Deutsch

Mischungsmodelle spielen in vielen statistischen Anwendungen eine wichtige Rolle, da sie einen natürlichen Ansatz darstellen, Heterogenität zu modellieren. Insbesondere in Zusammenhang mit Regressionsmodellen wurden Mischungsmodelle in den letzten Jahrzehnten intensiv studiert.

In dieser Arbeit werden zunächst theoretische Mittel erarbeitet, die zur Untersuchung von Schätzern in nichtparametrischen Regressionsmodellen auf gleichmäßige Konvergenzraten und gleichmäßige Adaptivität bezüglich der Glattheit der Parameterfunktionen dienen. Diese werden später auf nichtparametrische Regressionsmischungsmodelle angewendet.

Dazu seien Γ, I beliebige Mengen und Θ ein normierter Parameterraum. Für jedes $\gamma \in \Gamma$ sei eine deterministische Kontrastfunktion $M(\cdot, \cdot; \gamma) : \Theta \times I \to \mathbb{R}$ definiert, welche für jedes $x \in I$, $\gamma \in \Gamma$ ausschließlich von Parametern aus einer endlichen Menge $\mathfrak{S}_{x;\gamma}$ minimiert wird. Um diese Parametermengen zu schätzen, seien zufällige Kontrastfunktionen $M_n(\cdot, \cdot) : \Theta \times I \to \mathbb{R}$ gegeben, deren Minimierer im ersten Argument $\hat{\theta}_n(x)$ als Schätzer dienen. Hierbei kann man I als Kovariablenwerte und Γ als Menge aller Modellparameter interpretieren. Typischerweise sind $\gamma \in \Gamma$ Tupel aus Parameterfunktionen, Kovariablendichten und möglicherweise weiteren Modellparametern. Das Einführen der Menge $\mathfrak{S}_{x;\gamma}$ erlaubt die Untersuchung von Schätzern auf asymptotische Eigenschaften in Modellen, die nicht vollständig identifiziert sind. Dies ist ein häufiges Problem in Mischungsmodellen, da typischerweise das Umlabeln der Komponenten keine Veränderung der Verteilung zur Folge hat. In identifizierbaren Modellen sind die Mengen $\mathfrak{S}_{x;\gamma}$ typischerweise einelementig.

In Kapitel 3 verallgemeinern wir ein typisches Konsistenzresultat von M-Schätzern, vgl. van der Vaart (2000, Theorem 5.7), siehe Theorem 3.1.2, sodass es Aussagen über gleichmäßige Konsistenz über I, gleichmäßig über den Modellparametern Γ zulässt, auch wenn das Modell nicht vollständig identifiziert ist. Des Weiteren verallgemeinern wir ein klassisches Konvergenzratenresultat, vgl. van der Vaart und Wellner (1996, Theorem 3.2.5), siehe Theorem 3.1.3. Die Voraussetzungen des zweiten Resultats beinhalten gleichmäßige Konsistenz der zufälligen Kontrastfunktion, gleichmäßige Wegbeschränktheit der deterministischen Kontrastfunktion vom Minimum außerhalb Umgebungen um die Minimierer, und eine Lipschitzeigenschaft von erwarteten Schätzfehlerinkrementen

$$M_n(\cdot, x) - M(\cdot, x; \gamma) \ .$$

Insbesondere die letzten beiden Voraussetzungen sind mitunter schwer nachzuweisen. Deshalb formulieren wir im Anschluss speziellere Annahmen an Modelle, welche die Voraussetzungen von Theorem 3.1.3 implizieren, siehe Annahme 3.1.4. Diese Annahmen beinhalten im Wesentlichen Glattheitsannahmen an die Kontrastfunktionen, deren Ableitungen und der Parameterfunktionen, die positive Definitheit der Hesse Matrix von M am wahren Parameter, die gleichmäßige L^1 Konvergenz des Gradienten von M_n gegen den Gradienten von M mit entsprechender Rate und die gleichmäßige Konsistenz

der zufälligen Kontrastfunktion M_n.

Im Weiteren geben wir ein auf Lepskii (1992) basierendes Verfahren zur gleichmäßgen adaptiven Schätzung der Parameterfunktionen an, für den Fall, dass ein unbekannter Störparameter wie die Glattheit α von Hölder-α-glatten Funktionen vorliegt. Dieses Verfahren schätzt adaptiv an den unbekannten Störparameter α. Genauer gesagt wird angenommen, dass $\alpha \in [a, b]$, für beliebige Intervallgrenzen $0 < a < b < \infty$. Auf das Intervall $[a, b]$ wird ein in der Anzahl der Beobachtungen logarithmisch wachsendes äquidistantes Gitter gelegt, aus dem ein Punkt α_k gewählt wird, welcher zu einer adaptiven Wahl des Störparameters führt. Zu beachten ist, dass die adaptive Wahl nicht wie üblich basierend auf dem Verhalten der Schätzer für verschiedene Störparameter gewählt wird, da für diese keine exponentielle Fehlerungleichung zur Verfügung steht. Alternativ können die Gradienten der empirischen Kontrastfunktionen mit verschiedenen Störparametern zu Vergleichszwecken herangezogen werden. Für diese stehen derartige Ungleichungen zur Verfügung, siehe Lemmata 3.2.2, 3.2.3 und 3.2.4.

Im Anschluss erweitern wir Annahme 3.1.4, sodass diese gleichmäßige Adaptivität eines Schätzers implizieren, siehe Annahme 3.1.11. Insbesondere muss für jeden Störparameter der Gradient der zufälligen Kontrastfunktion den Gradienten der entsprechenden deterministischen Kontrastfunktion mit korrekter Rate gleichmäßig in L^2 schätzen. Außerdem müssen wir polynomielles Abklingen der Tail-Wahrscheinlichkeiten des gleichmäßgen Bias des Gradienten von M_n unter verschiedenen Störparametern annehmen.

Des Weiteren geben wir theoretische Mittel, mit denen man für gewisse Typen von zufälligen Kontrastfunktionen obere Schranken für den stochastischen gleichmäßigen L^p-Fehler der Kontrastfunktionen bestimmen kann. Bei diesen Typen handelt es sich um lineare und U-Statistik Schätzer, also

$$M_n(\theta, x; h) = \frac{1}{n} \sum_{k=1}^{n} \tau(Y_k, \theta) K_h(X_k - x) \ , \quad \text{oder}$$

$$M_n(\theta, x; h) = \frac{1}{n(n-1)} \sum_{\substack{j,k=1 \\ j \neq k}}^{n} \tau(Y_j, Y_k, \theta) K_h(X_j - x) K_h(X_k - x) \ ,$$

wobei τ jeweils eine deterministische Funktion, $K : \mathbb{R}^d \to \mathbb{R}$ ein Kern und $h \in (0, \infty)$ eine Bandbreite ist.

In Kapitel 4 stellen wir zwei Regressionsmischungsmodelle vor und wenden die vorher erarbeiteten theoretischen Mittel an.

Bei dem ersten Modell, siehe Abschnitt 4.1, handelt es sich um eine Mischung von Gauß'schen Regressionen. Das heißt, zwischen beobachteten Zufallsvariablen Y und

stetigen Regressoren X mit Träger $I \subset \mathbb{R}^d$ besteht der funktionale Zusammenhang

$$Y = \sum_{c=1}^{m} \mathbb{1}_{\Pi=c}\big(\sigma_c(X)\varepsilon_c + \mu_c(X)\big) \ ,$$

wobei Π eine latente Zufallsvariable mit Werten in $\{1, \dots, m\}$ ist, sodass $\mathbb{P}(\Pi = c|X = x) = \pi_c(x)$ für Mischungsfunktionen $\pi_c : I \to (0,1)$ mit $\sum_{c=1}^{m} \pi_c = 1$; die Zufallsvariablen ε_c bedingt auf $X = x$ unabhängig von Π und standardnormalverteilt sind; und $\mu_c : I \to \mathbb{R}$, $\sigma_c : I \to (0, \infty)$ Lokations- bzw. Skalenfunktionen sind. Anders ausgedrückt, die bedingte Verteilung von Y auf $X = x$ ist eine Mischung von Normalverteilungen, wobei Mischungs-, Lokations- und Skalenparameter von den Kovariablen abhängen. Dieses Modell wurde bereits von Huang et al. (2013) untersucht. Die Autoren konnten nichtparametrische Identifizierbarkeit mit univariaten Regressoren unter Differenzierbarkeitsannahme der Lokations- und Skalenfunktionen mit Transversalitätsargumenten zeigen. Ferner waren sie in der Lage, mit der lokalen Log-Likelihood Methode punktweise asymptotisch normale Schätzungen für zwei-mal differenzierbare Parameterfunktionen zu etablieren.

In Abschnitt 4.1.1 werden verchiedene Identifizierbarkeitsresultate gegeben, von denen eines die Grundlage für gleichmäßig konsistentes Schätzen in Abschnitt 4.1.2 bildet. Wir untersuchen Schätzungen von Hölder-α-glatten Parameterfunktionen, vgl. Abschnitt 2.4, mit der von Huang et al. (2013) vorgestellten Log-Likelihood Methode. Dabei können wir die gleichmäßige Konsistenz des Schätzers über Kompakta $J \subset \operatorname{int}(I)$ mit Rate $\big(\frac{\log n}{n}\big)^{\frac{\alpha}{2\alpha+d}}$ und Adaptivität mit einer der Lepski-Methode entsprechenden Bandbreitenwahl etablieren, vgl. Abschnitte 4.1.3, 4.1.5.

Bei der Anwendung der in Kapitel 3 erarbeiteten Methoden ist zu beachten, dass dieses Modell in der Tat nur bis auf Umlabeln der Komponenten identifizierbar ist. Entsprechend ergeben sich in der Praxis zunächst nur punktweise Schätzungen auf endlichen Gittern im Kovariablenträger, deren Komponentenlabel nicht zwingend übereinstimmen. Um eine sinnvolle Schätzung der Parameterfunktionen zu erhalten, muss man sicherstellen, dass die Label an verschiedenen Kovariablenstellen korrekt zugeordnet werden. Dadurch, dass der Schätzer gleichmäßig konsistent ist, können wir eine Methode angeben, mit der man mit gegen eins strebender Wahrscheinlichkeit die Label der Schätzer an benachbarten Gitterpunkten korrekt zuordnet, vgl. Abschnitt 4.1.4.

Bei dem anderen Modell, siehe Abschnitt 4.2, handelt es sich um eine Mischung zweier Regressionen. Zwischen beobachteten Zufallsvariablen Y und stetigen Regressoren X mit Träger $I \subset \mathbb{R}^d$ besteht der funktionale Zusammenhang

$$Y = W\big(\mu(X) + \varepsilon_1\big) + \big(1 - W\big)\sigma(X)\varepsilon_2 \ ,$$

wobei W eine latente Zufallsvariable mit bedingter Verteilung $W|X = x \sim \operatorname{Ber}(p(x))$, mit $p : I \to (0,1)$ ist; die Zufallsvariablen ε_1 und ε_2 bedingt auf $X = x$ unabhängig von W sind und symmetrische bedingte Dichten haben, wobei $\varepsilon_1|X = x \sim f_x$ eine unbekannte Dichte und $\varepsilon_2|X = x \sim \bar{f}$ eine bekannte Dichte ist; und $\mu : I \to \mathbb{R}$, $\sigma : I \to (0, \infty)$

eine Lokations- bzw. Skalenfunktion ist.

Ein ähnliches Modell wurde bereits von Butucea et al. (2017) untersucht. Die Autoren konnten punktweise Konsistenz und asymptotische Normalität unter Undersmoothing eines auf der Symmetrie der Komponenten basierenden Schätzers zeigen.

Da sich im Modell mit zwei Komponenten die beiden symmetrischen Komponentendichten unterscheiden und eine Komponente skaliert und die andere translatiert wird, ergibt sich in diesem Modell nicht das Problem des Umlabelns wie zum Beispiel im Gauß'schen Mischungsmodell oder im Modell von Butucea et al. (2017). Entsprechend erhalten wir stärkere Identifizierbarkeitsresultate. Ein Resultat basiert auf der Idee, die charakteristischen Funktionen beider Komponentendichten in den Tails unterscheidbar zu machen. Ein anderes Resultat basiert darauf, den Wertebereich der Mischungsfunktion p auf $(1/2, 1)$ einzuschränken.

In Abschnitt 4.2.2 konstruieren wir für das Mischungsmodell mit zwei Komponenten basierend auf der Symmetrie der Komponentendichten einen asymptotischen Kontrast M, vgl. (4.2.8), welcher nicht-negativ ist und ausschließlich am wahren Parameter null wird. Dieser Kontrast wird empirisch durch eine Funktion M_n, vgl. (4.2.9) geschätzt, deren Minimierer als Schätzer für die Hölder-α-glatten Modellparameterfunktionen dient. Dieser Schätzer ist ebenfalls auf jedem Kompaktum $J \subset \mathrm{int}(I)$ gleichmäßig konsistent und konvergiert gleichmäßig mit Rate $\left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+d}}$, siehe Abschnitt 4.2.3. Außerdem können wir auch hier die Lepski-Methode anwenden, um eine adaptive Bandbreitenwahl zu erhalten, siehe Abschnitt 4.2.4.