Electronic Journal of Statistics
ISSN: 1935-7524
arXiv: arXiv:0000.0000

Adaptive estimation in the supremum norm for semiparametric mixtures of regressions

Heiko Werner[1,*] **Hajo Holzmann**[1,**] **and Pierre Vandekerkhove**[2]

[1] *Fachbereich Mathematik und Informatik, Philipps-Universität Marburg, Germany. e-mail:* [*] wernerh@Mathematik.Uni-Marburg.de; [**] holzmann@Mathematik.Uni-Marburg.de

[2] *Université Paris-Est LAMA (UMR 8050), UPEMLV F-77454, Marne-la-Vallée, France e-mail:* pierre.vandekerkhove@u-pem.fr

Abstract: We investigate a flexible two-component semiparametric mixture of regressions model, in which one of the conditional component distributions of the response given the covariate is unknown but assumed symmetric about a location parameter, while the other is specified up to a scale parameter. The location and scale parameters together with the proportion are allowed to depend nonparametrically on covariates. After settling identifiability, we provide local M-estimators for these parameters which converge in the sup-norm at the optimal rates over Hölder-smoothness classes. We also introduce an adaptive version of the estimators based on the Lepski-method. Sup-norm bounds show that the local M-estimator properly estimates the functions globally, and are the first step in the construction of useful inferential tools such as confidence bands. In our analysis we develop general results about rates of convergence in the sup-norm as well as adaptive estimation of local M-estimators which might be of some independent interest, and which can also be applied in various other settings. We investigate the finite-sample behaviour of our method in a simulation study, and give an illustration to a real data set from bioinformatics.

Keywords and phrases: adaptive estimation, M-estimation, switching regression, semiparametric mixture, uniform rates of convergence.

1. Introduction

Practitioners are frequently interested in modelling the effect of a d-dimensional explanatory vector X on a response random variable Y by using a regression model estimated from a random sample $(X_i, Y_i)_{1\leq i\leq n}$ of (X, Y). To allow varying parameters for different groups of observations, finite mixtures of regressions (FMRs) have been suggested in the literature. Statistical inference for parametric FMR models using a moment generating function method was first introduced by Quandt and Ramsey (1978). An approach based on the expectation-maximization (EM) algorithm was suggested by De Veaux (1989) in the two-component case. Zhu and Zhang (2004) established the asymptotic theory for testing for the number of components in parametric FMR models. More recently, Städler et al. (2010) proposed an ℓ_1-penalized method based on a Lasso-type estimator for a high-dimensional FMR model with $d \gg n$.

To gain further flexibility, various authors suggested the use of semiparametric FMR models. Hunter and Young (2012) studied the identifiability of an m-component semiparametric FMR model and numerically investigated an EM algorithm for estimating its parameters. Bordes et al. (2013) showed asymptotic normality of a semiparametric estimator in a two-component mixture of linear regressions. Huang and Yao (2012) and Huang et al. (2013) considered a semiparametric linear and nonlinear FMR model with Gaussian noise in which means, variances and mixing proportions depend on covariates nonparametrically. They established also the asymptotic normality of their local maximum likelihood estimator and investigated a modified EM-type algorithm. Recently Butucea et al. (2017) proposed a Fourier based approach to deal with a new semiparametric topographical mixture model able to capture the characteristics of dichotomously shifted response-type experiments. See also Compiani and Kitamura (2016) for an overview on semiparametric mixtures with a focus on econometric applications.

In this paper we investigate a two-component FMR model, in which one of the conditional component distributions is unknown but assumed symmetric about a location parameter μ, while the other is specified up to some scale parameter σ. The location parameter μ, the scale parameter σ as well as the proportion p are allowed to depend nonparametrically on the covariates. After settling identifiability, we provide local M-estimators for these parameters which converge in the sup-norm at the optimal rates over Hölder-smoothness classes. We also introduce an adaptive version of the estimators based on the Lepski-method, see Lepskii (1992).
Sup-norm bounds show that the local M-estimator properly estimates the functions globally, and allow for a slight additional smoothing of the estimated functions in order to obtain continuous estimates without deteriorating the rates of convergence. Further, uniform rates are the first step for the construction of confidence bands which are a very useful inferential tool, see e.g. Chernozhukov et al. (2014).
Inspired by Butucea et al. (2017), the contrast that we use in the estimation procedure is based on characteristic functions, thus simplifying the approach in Bordes and Vandekerkhove (2010) which requires an additional smoothing when building the contrast. We also develop general useful technical tools based on the Bernstein-inequality in Giné et al. (2000) when the contrast has the form of a U-statistic.
In our analysis we develop general results about rates of convergence in the sup-norm as well as adaptive estimation of local M-estimators which might be of some independent interest, and which can also be applied in various other settings, e.g. to the models in Butucea et al. (2017) or in Huang and Yao (2012).
The paper is organized as follows. In Section 2 we formally introduce the model. Section 3 deals with identifiability of the parameters, for which we provide some general results. Section 4 introduces the estimation methodology and in particular develops the contrast function underlying the M-estimator. In Section 5 we obtain optimal rates of convergence in the sup-norm for our estimators, while

Section 6 deals with adaptivity. In Section 7 we provide results of some numerical experiments, and also analyze the ChipMix data set from Martin-Magniette et al. (2008) which was previously analyzed in Bordes et al. (2013) using linear FMRs. Section 8 presents our general theory for local M-estimators as well as technical tools for contrasts in the form of U-statistics. Finally, Sections 9 - 12 contain the technical proofs.

2. Two-component mixture of location-scale regressions

We consider the following nonparametric regression model

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

for sequences of independent and identically distributed (i.i.d.) random vectors $(X_i)_{i\in\mathbb{N}}$ supported on a compact set $I \subset \mathbb{R}^d$, $d \geq 1$, 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 assumed to be observable while 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 probability density function (pdf), denoted by $\ell : I \to (0,\infty)$, with respect to the Lebesgue measure. The unknown location and scaling functions $\mu : I \to \mathbb{R}$ and $\sigma : I \to (0,\infty)$ partially determine the distributional relationship between the explanatory and response variables along with the unknown mixing function $p : I \to (0,1)$. Finally conditionally on $\{X_i = x\}$, the variables W_i are assumed to have a Bernoulli-distribution with parameter $p(x)$, that is

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

Further we assume that conditionally on $\{X_i = x\}$, the vectors $\varepsilon_{1,i}$ and $\varepsilon_{2,i}$ have zero-symmetric conditional pdfs, denoted respectively f_x and $\bar{f}$, where $\bar{f}$ is assumed to be known and not to depend on x, while f_x is unknown and may depend on x. If we furthermore have the conditional independence relations

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

the random vectors (Y_i, X_i) have the following joint density

$$\begin{aligned} f_{Y,X}(y,x) &:= f_{Y|X}^{\vartheta(\cdot)}(y|x)\ell(x) \qquad (2.1)\\ &= \left[\frac{1-p(x)}{\sigma(x)}\bar{f}\Big(\frac{y}{\sigma(x)}\Big) + p(x)f_x\big(y - \mu(x)\big)\right]\cdot\ell(x)\ , \quad (y,x) \in \mathbb{R}\times I\ , \end{aligned}$$

where the functional parameter

$$\vartheta(x) = (p(x), \sigma(x), \mu(x), f_x)$$

collects all the x-local quantities to be estimated from the data.

3. Identifiability

Regarding the identifiability problem, it is enough to consider model (2.1) without covariate as we aim to estimate the various parameter-functions for each given value of x. Our identification strategy and results will be similar to those in Bordes et al. (2006) and Hohmann and Holzmann (2013). We suppose that both the known pdf $\bar{f}$ as well as the unknown pdf f are zero-symmetric and have finite third-order moments. Hence, we consider mixtures of the following form

$$f_{\text{mix}}(y;\vartheta) = (1-p)\bar{f}(y/\sigma)/\sigma + pf(y-\mu)\,, \quad y \in \mathbb{R}\,, \tag{3.1}$$

where

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

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

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

Note that we may assume that $\bar{f}$ is normalized, that is $\int y^2\,\bar{f}^2(y)\mathrm{d}y = 1$. In the following we provide two sets of identifiability assumptions. The results rely on the symmetry of the component pdfs. Indeed f is symmetric if and only if 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.
Our first assumption imposes a constraint on the true mixing parameter p_* but requires only mild conditions on the component pdfs $\bar{f}$ and f_*.

Assumption 1. The true model parameter $\vartheta_* = (p_*,\sigma_*,\mu_*,f_*)^\top$ and the component pdf $\bar{f}$ satisfy

- **(I1)** $\mu_* \in \mathbb{R}\backslash\{0\}$, $p_* \in (1/2,1)$ and $\sigma_* \in (0,\infty)$,
- **(I2)** $\bar{f} \in \mathcal{E}_3$ and $\varphi_{\bar{f}} > 0$,
- **(I3)** $f_* \in \mathcal{E}_3$ and $\varphi_{f_*} > 0$.

The second assumption does not impose a restriction on the mixing parameter but rather depends on the relationship of both component densities $\bar{f}$ and f_*. That is, the characteristic functions of these densities need to be distinguishable in the tails in one of the following manners.

Condition 1. We consider the two following conditions:

- **(C1)** For large $t \in \mathbb{R}$ it holds that $\varphi_{f_*}(t) \neq 0$ and for all $\sigma > 0$, we have

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

- **(C2)** For large $t \in \mathbb{R}$ it holds that $\varphi_{f_*}(t) \neq 0$, $\varphi_{\bar{f}}(t) \neq 0$ and for all $\sigma > 0$, we have

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

Example 1.

(i) Condition **(C1)** holds when $f_* \sim \text{Laplace}(\mu_1, \sigma_1)$ and $\bar{f} \sim \mathcal{N}(\mu_2, \sigma_2^2)$.
(ii) Condition **(C2)** holds when $\bar{f} \sim t(\nu)$ and $f_* \sim \mathcal{N}(\mu_1, \sigma_1^2)$.
(iii) When both component densities are Gaussian, none of the conditions are satisfied. Identification is still possible under Assumption 1, however.

Admissible unknown component pdfs f_* are aggregated in the class of functions

$$\mathcal{E}_3^{\bar{f}} = \{f \in \mathcal{E}_3 : (\bar{f}, f) \text{ meets one of the conditions } \textbf{(C1)} \text{ or } \textbf{(C2)}\} .$$

The second identifiability assumption is as follows.

Assumption 2. The model parameter $\vartheta_* = (p_*, \sigma_*, \mu_*, f_*)^\top$ and the known component density $\bar{f}$ fulfill

(I′1) $\mu_* \in \mathbb{R}\backslash\{0\}$, $p_* \in (0,1)$, $\sigma_* \in (0, \infty)$,
(I′2) $\bar{f} \in \mathcal{E}_3$,
(I′3) $f_* \in \mathcal{E}_3^{\bar{f}}$.

We can now state the following identifiability theorem, the proof of which is provided in Section 9.

Theorem 3.1 (Identifiability)**.** *If Assumption 1 or 2 holds we have the following identifiability property. If ϑ satisfies $f_{\text{mix}}(y; \vartheta_*) = f_{\text{mix}}(y; \vartheta)$ for almost all $y \in \mathbb{R}$ then $\vartheta = \vartheta_*$.*

Remark 1. It is important to notice that in both identifiability results, the technical conditions are only imposed on the true parameter $\vartheta_* = (p_*, \sigma_*, \mu_*, f_*)^\top$. Identification is then ensured within the whole class of parameters.

4. Estimation Methodology

We first present our estimation methodology in the model (3.1) without covariates. The approach to build a contrast function based on Fourier transformation is inspired by Butucea and Vandekerkhove (2014). In particular, as opposed to Bordes et al. (2006) and Bordes and Vandekerkhove (2010) we do not require an additional smoothing parameter for the indicator to obtain a smooth contrast function. Hence, in this restricted setting our approach yields asymptotically normally distributed estimators at $\sqrt{n}$-rate without additional smoothing.
Specifically, first assume that the observations Y_j have density $f_{\text{mix}}(y; \vartheta_*)$ as in (3.1), where $\bar{f}, f_* \in \mathcal{E}_3$ and

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

The characteristic function of $f_{\text{mix}}(\cdot; \vartheta_*)$ is given by

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

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

$$\Big(\varphi_{f_{\mathrm{mix}}(\cdot;\vartheta_*)}(t) - (1-p_*)\,\varphi_{\bar{f}}(\sigma_* t)\Big)\, e^{-it\mu_*}\,. \tag{4.1}$$

Since $\varphi_{f_{\mathrm{mix}}(\cdot;\vartheta_*)}(t)\, e^{-it\mu_*}$ is the characteristic function of $Y - \mu_*$, we get that the imaginary part of (4.1) satisfies

$$\begin{aligned} 0 &= \Im\Big(\Big(\varphi_{f_{\mathrm{mix}}(\cdot;\vartheta_*)}(t) - (1-p_*)\,\varphi_{\bar{f}}(\sigma_* t)\Big)\, e^{-it\mu_*}\Big) \\ &= \mathbb{E}_{\vartheta_*}\Big[\sin\big((Y-\mu_*)\,t\big)\Big] + (1-p_*)\,\varphi_{\bar{f}}(\sigma_* t)\,\sin(t\mu_*) \end{aligned} \tag{4.2}$$

for all $t \in \mathbb{R}$, where we used that $\varphi_{\bar{f}}(\sigma_* t)$ is real-valued since $\bar{f}$ is symmetric, and where $\mathbb{E}_{\vartheta_*}$ denotes the expectation with respect to the distribution $\mathbb{P}_{\vartheta_*}$ which has density $f_{\mathrm{mix}}(y;\vartheta_*)$ with respect to Lebesgue measure. Hence, setting $H : \mathbb{R}\times\mathbb{R}\times[0,1]\times(0,\infty)\times\mathbb{R} \to [-2,2]$,

$$H(y,t,\theta) = \sin\big((y-\mu)\,t\big) + (1-p)\,\varphi_{\bar{f}}(\sigma\, t)\,\sin(\mu t), \tag{4.3}$$

we can define the contrast function

$$M(\theta;\vartheta_*) := \int_{\mathbb{R}} \mathbb{E}^2_{\vartheta_*}\big[H(Y,t,\theta)\big] q(t)\,\mathrm{d}t \tag{4.4}$$

for some strictly positive density q that is chosen a priori. We have the following identification result for this contrast.

Proposition 4.1 (Contrast property)**.** *Let Assumption 1 or 2 hold. Then the function $M(\cdot;\vartheta_*) : [0,1]\times(0,\infty)\times\mathbb{R} \to [0,4]$ defined in (4.4) is a discrepancy function, that is for $\theta \in [0,1]\times(0,\infty)\times\mathbb{R}$, we have*

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

The proof is provided in Section 9. An estimator $\hat{\theta}_n$ for θ is based on minimizing an empirical version of the contrast given by the U-statistic

$$M_n(\theta) = \frac{1}{n\,(n-1)} \sum_{1\le j\ne k\le n} \int H(Y_j,t,\theta) H(Y_k,t,\theta) q(t)\,\mathrm{d}t\,.$$

An analysis similar to that in Butucea and Vandekerkhove (2014) shows that under appropriate assumptions, the estimator $\hat{\theta}_n$ is asymptotically normally distributed. Let us return to the regression model (2.1). Our general estimation strategy is then analogous to that in Butucea et al. (2017). For the x-local parameter

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

and

$$\vartheta_*(x) = (\theta_*^\top(x), f_x^*)^\top, \quad \text{with} \quad f_x^* \in \mathcal{E}_3,$$

Assumption 1 or 2 is imposed globally. The asymptotic contrast is given by

$$M(\theta, x; \gamma) := \int_{\mathbb{R}} \mathbb{E}_\gamma^2 \big[H(Y, t, \theta) \big| X = x \big] q(t) \, \mathrm{d}t \cdot \ell^2(x) \, , \tag{4.5}$$

where again $\mathbb{E}_\gamma$ denotes the expectation with respect to the distribution $\mathbb{P}_\gamma$, which is the probability measure from the underlying statistical model, i.e.

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

In order to estimate the contrast M, we use a U-statistic type estimator localized at x,

$$\begin{aligned} M_n(\theta, x; h) = & \frac{1}{n\,(n-1)} \sum_{1 \le j \ne k \le n} \Big(\int H(Y_j, t, \theta) H(Y_k, t, \theta) q(t) \, \mathrm{d}t \\ & \cdot K_h(X_j - x) K_h(X_k - x) \Big) \, , \end{aligned} \tag{4.6}$$

where $K : \mathbb{R}^d \to \mathbb{R}$ is a kernel function and $h \in (0, \infty)$ is a bandwidth parameter. The estimator $\hat{\theta}_n : I \to \mathbb{R}^3$ of the parameter function $\theta_*(\cdot)$ is then defined as the pointwise minimizer of (4.6), that is

$$\hat{\theta}_n(x; h) \in \operatorname*{argmin}_{\theta \in \Theta} M_n(\theta, x; h) \, , \tag{4.7}$$

where Θ is a suitable compact subset of $(0, 1) \times (0, \infty) \times \mathbb{R} \backslash \{0\}$ that we specify below.

5. Optimal rate of convergence in the supremum norm

In this section we derive the convergence rate of the estimator $\hat{\theta}_n(x; h)$ for the underlying parameter functions $p_*(\cdot)$, $\mu_*(\cdot)$, $\sigma_*(\cdot)$ over Hölder smoothness classes. We focus on the supremum norm error for the following reasons. First, although the estimator is defined as a pointwise minimizer in (4.7), convergence in the sup-norm shows that it properly estimates the parameter functions $p_*(\cdot)$, $\mu_*(\cdot)$, $\sigma_*(\cdot)$ in a global way. Second, a sup-norm bound allows to slightly smooth the estimated functions in order to obtain continuous estimates, without deteriorating the rates of convergence. Third, uniform rates are the first step for the construction of confidence bands which are a very useful inferential tool, see e.g. Chernozhukov et al. (2014).

Our technical analysis is based on general results for local M-estimators obtained in Section 8.1, and hence is quite different from that in Butucea et al. (2017) who prove pointwise asymptotic normality using undersmoothing. Indeed, our approach could also be applied to their model to obtain similar results as in Theorems 5.1 and 6.1 below.

We investigate estimation over Hölder-smoothness classes of functions. Denote the set of Hölder smooth functions on I with Hölder parameter $\alpha > 0$ and Hölder constant $L > 0$ taking values in some set U by

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

Here $\lfloor\alpha\rfloor = \max\{k \in \mathbb{N}_0 \mid k < \alpha\}$ and we use the standard multi-index notation for multivariate derivatives, i.e. for $k = (k_1, \ldots, k_d)$, we write $\partial^k f = \partial_1^{k_1} \ldots \partial_d^{k_d} f$ and $|k| = k_1 + \ldots + k_d$. Note that if U is bounded we have that

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

We suppose that $p_*(\cdot)$, $\mu_*(\cdot)$, $\sigma_*(\cdot)$ and $\ell(\cdot)$ are Hölder smooth with the same parameters α and $L > 0$. Specifically, for given $U_p \subset (0,1)$, $U_\mu \subset \mathbb{R}\backslash\{0\}$, $U_\sigma \subset (0,\infty)$, $U_\ell \subset (0,\infty)$ we consider the set of parameters

$$\begin{aligned}\Gamma(\alpha) = \big\{ & \gamma = (\theta_*^\top(\cdot), \ell(\cdot))^\top \mid \ell \in \mathcal{H}(\alpha, L, U_\ell),\ \theta_* = (p_*, \mu_*, \sigma_*)^\top \text{with} \\ & p_* \in \mathcal{H}(\alpha, L, U_p),\ \mu_* \in \mathcal{H}(\alpha, L, U_\mu),\ \sigma_* \in \mathcal{H}(\alpha, L, U_\sigma) \big\}\,.\end{aligned} \tag{5.1}$$

For convenience the sets U_p, U_σ, U_μ, U_ℓ in the definition of $\Gamma(\alpha)$ in (5.1) are assumed to be compact rectangular sets. We shall take Θ in the definition (4.7) of the estimator $\hat{\theta}_n(x; h)$ to be

$$\Theta = U_p \times U_\sigma \times U_\mu. \tag{5.2}$$

Note that we excluded the conditional density f_x^* of ε_1 given $X = x$ from the parameter set. Indeed we of course do not assume that this is known, but in their present form the rates are not uniform with respect to this parameter. Extensions are possible but would result in still higher technical complexity.

Assumption 3.

- **(M1)** The identification Assumption 1 or 2 is fulfilled for all $x \in I$.
- **(M2)** For each y we have that $f_\cdot^*(y) \in \mathcal{H}(\alpha, L(y), U)$ for some integrable and bounded function $L(\cdot)$ and some compact set $U \subset [0,\infty)$. In addition, for $x \in I$ the characteristic function $\varphi_{f_x^*}$ of the density f_x^* is strictly positive.
- **(M3)** The known component density $\bar{f}$ fulfils **(I2)** and the functions $y \mapsto y\bar{f}(y)$, $\bar{f}$ and $\partial^2\varphi_{\bar{f}}$ are bounded and we have that $\lim_{|t|\to\infty} t\partial\varphi_{\bar{f}}(t) = 0$.

- **(K1)** The kernel $K : \mathbb{R}^d \to \mathbb{R}$ is Lipschitz continuous with Lipschitz constant $L_K > 0$ and has support $[-1,1]^d$.
- **(K2)** The kernel K is of order α, i.e. for all $|k| \le \lfloor\alpha\rfloor$ it holds that

$$\int z^k K(z)\,\mathrm{d}z = \int z_1^{k_1} \cdot \ldots \cdot z_d^{k_d} K(z)\,\mathrm{d}z = 0\,.$$

(K3) The probability density q has a finite third absolute moment and is bounded.

Theorem 5.1 (Main result: Rate of convergence)**.** *Under Assumption 3, given a compact rectangle $J \subset \mathrm{int}(I)$, if we let $h_n \sim \left(\frac{\log n}{n}\right)^{\frac{1}{2\alpha+d}}$, we have that*

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

Thus, the estimator $\hat{\theta}_n(\cdot;h_n)$ has the convergence rate $\left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+d}}$ in the sup-norm for convergence in probability over the parameter set $\Gamma(\alpha)$. A classic result from Stone (1982) states that this rate is optimal for nonparametric regression in d dimensions over Hölder smoothness classes. The proof of Theorem 5.1 which relies on the theory presented in Section 8.1 is given in Section 10.

6. Adaptive estimation

In Theorem 5.1, the choice of the bandwidth $h_n \sim \left(\frac{\log n}{n}\right)^{\frac{1}{2\alpha+d}}$ requires a-priori knowledge of the smoothness parameter α. In this section we shall make the estimator $\hat{\theta}_n(x;h)$ in (4.7) adaptive w.r.t. this parameter by using the Lepski-method, see Lepskii (1992), Lepski et al. (1997) and Golubev et al. (2000).
We shall use an indirect approach and choose an adaptive bandwidth based on the gradients of the contrast functions in (4.5) and (4.6),

$$S_n(\theta,x;h) = \partial_\theta M_n(\theta,x;h) \,, \qquad S(\theta,x;\gamma) = \partial_\theta M(\theta,x;\gamma), \tag{6.1}$$

where $\partial_\theta = (\partial_p, \partial_\sigma, \partial_\mu)^\top$. We let

$$h(\alpha) = h_n(\alpha) = (\log n/n)^{1/(2\alpha+d)} \quad \text{and} \quad r(\alpha) = r_n(\alpha) = h(\alpha)^\alpha$$

which we consider over a grid of smoothness parameters

$$\alpha_k = a + k\frac{b-a}{N}, \qquad k = 0,\dots,N \,,$$

where $N = \lceil \log n \rceil = \min\{k \in \mathbb{N} \mid k > \log n\}$, and set

$$h_k = h(\alpha_k), \qquad r_k = r(\alpha_k).$$

For a sufficiently large constant $C_{\mathrm{Lep}} < \infty$ we consider the Lepski choice

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

which leads to the estimator

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

In order to make use of the highest possible smoothness order b, we need the following assumption.

($\widetilde{\mathbf{K2}}$) The kernel K is of order $\lfloor b \rfloor$.

Theorem 6.1 (Main result: Adaptive rate of convergence)**.** *Let $0 < a < b < \infty$ and let K be a kernel fulfilling Assumptions* **(K1)** *and* **($\widetilde{\mathbf{K2}}$)**. *Then, under Assumptions* **(M1)** - **(M3)** *and* **(K3)**, *for any compact rectangular set $J \subset \mathrm{int}(I)$ and for sufficiently large $C_{\mathrm{Lep}} > 0$ we have that*

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

The proof of this theorem, which is given in Section 10, is again based on a general adaptivity result for local M-estimators obtained Section 8.1.

7. Simulations and real data illustration

7.1. Simulations

We propose in this section to investigate the finite sample size properties, in the supremum norm sense, of the functional estimator $\hat{\theta}_n(x; h_n) = (\hat{p}_n(x), \hat{\sigma}_n(x), \hat{\mu}_n(x))$ over two models **(M1)** and **(M2)** described below in dimension $d = 1$. Commonly to both models we choose

$$p(x) = 0.75 - 0.15 \sin\left(\frac{x}{4}\right), \ \mu(x) = 1.5 + \frac{1}{2}\sin\left(\frac{x}{2}\right),$$
$$\sigma(x) = \frac{1}{2} + \frac{1}{4}\sin\left(\frac{x}{4}\right).$$

and

$$(\mathbf{M1} : \mathbf{Gaussian}) \quad X \sim \mathcal{N}(3,7),\ \varepsilon_{1,i} | \{X_i = x\} \sim \mathcal{N}\left(0, \frac{1}{4}\right),\ \varepsilon_{2,i} \sim \mathcal{N}(0,1),$$

$$(\mathbf{M2} : \mathbf{Laplace}) \quad X \sim \mathcal{N}(3,7),\ \varepsilon_{1,i} | \{X_i = x\} \sim \mathrm{Laplace}(0, \frac{1}{2\sqrt{2}}),\ \varepsilon_{2,i} \sim \mathcal{N}(0,1).$$

Notice that we set the variance of $\varepsilon_{1,i} | \{X_i = x\}$ equal to $1/4$ in both models **(M1)** and **(M2)** for fair comparison. Identifiability also of model **(M1)** is guaranteed by Theorem 3.1 since Assumption 1 is satisfied.
The density q in the empirical contrast $M_n(\vartheta, x; h)$ in (4.6) is a $\mathcal{N}(0,1)$ distribution, the kernel $K(\cdot) = 1/2(1 - |x|)\mathbb{I}_{-1\leq x\leq 1}$ (triangular kernel) and

$$h_n = \kappa \left[1.06 \times \left(\frac{\hat{\sigma}_X + \hat{\sigma}_Y}{2} \right) n^{-1/5} \right]. \tag{7.1}$$

where κ is a smoothness/scaling parameter the influence of which is to be tested. The general form within brackets is a sort of rule of thumb. Thus we refrain from implementing the Lepski search and instead manually investigate the influence of the bandwidth over a suitable grid of values. The initialization is done at:

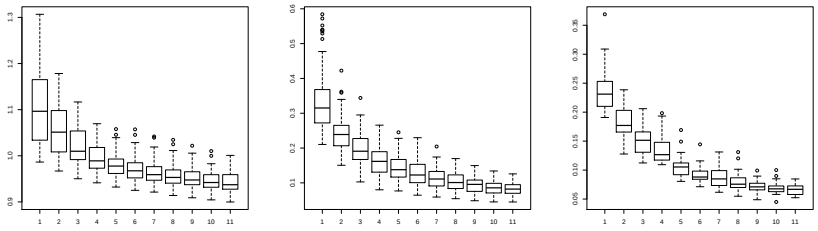
$p(x)_{\texttt{initial}} = p(x) + \text{unif}(-0.1, 0.1)$, $\mu_{\texttt{initial}} = \mu(x) + \text{unif}(-0.25, 0.25)$ and $\sigma_{\texttt{initial}} = \sigma(x) + \text{unif}(-0.1, 0.1)$ for model **(M1)**.
We compute our estimator $\theta_n(\cdot, h_n)$ over a testing grid

$$\mathcal{G} = \{x_k = -10 + k;\ 1 \leq k \leq K = 25\},$$

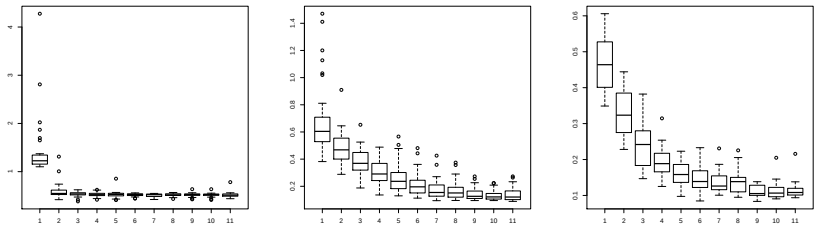
which is basically the interval $[-5, 10]$ divided in cells of size 0.05. In Figure 1, respectively Figure 4, we present the behavior of the supremum error distribution over the location, scaling and proportion functions in model **(M1)**, respectively model **(M2)**, for $n = 4000$, $10,000$ and $20,000$ and index values of κ defined in (7.1). In Figure 2, resp. 3, we display one single run performance for $n = 4000$, resp. $n = 10,000$, under $\kappa = 4$ and 10, to illustrate the influence of the sample size n and the scaling parameter κ on our method.
Comments on Figures 1–4. We remark first that, despite the fact that the variance of the second component is common to models **(M1)** and **(M2)**, the performances in terms of bias and variance of the supremum norm are dramatically better for model **(M2)**. While this seems to be somewhat in contrast to our theory since the Gaussian density is much smoother than the Laplace density, we suspect that the effect is due to better separability of the two densities (Gaussian and Laplace) in model **(M2)** as compared to model **(M1)** (both Gaussian). Further, Figures 2 and 3 show the positive impact of the sample size on the largest estimation deviation over the different parameter functions. We clearly see that the estimated curves better "hold onto" the target curves when the sample size increases from 4,000 to 10,000. Let us finally notice that the most difficult parameter to control is the scaling as illustrated in Figures 2 (b) and 3 (b) by the quite large amplitude of the oscillations along the graphs.

7.2. Application to NimbleGen high density array

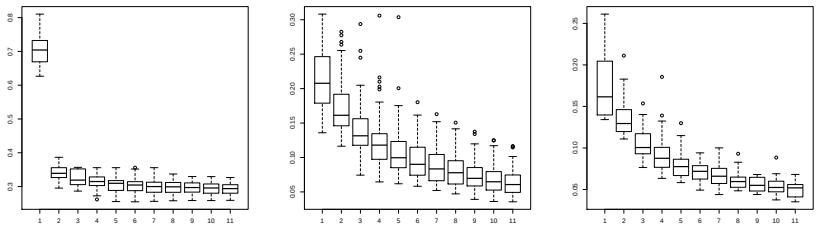
We consider the NimbleGen high density array dataset analyzed by Martin-Magniette et al. Martin-Magniette et al. (2008) and Bordes et al Bordes et al. (2013). The aim of these authors was to fit a simpler linear model than (2.1), where basically $p(x) = p \in (0, 1)$ is fixed, the location function $\mu_x = \alpha + \beta x$ where α and β are respectively the intercept and slope of the second component linear regression function, and the scaling function $\sigma(x) = \sigma$ is known. Originally the dataset, produced by a two color ChIP-chip experiment, consists of $n = 176,343$ observations $(x_i, \tilde{y}_i)$. A parametric mixture of linear regressions with two unknown components was fitted first to the data by Martin-Magniette et al. Martin-Magniette et al. (2008) under the assumption of normal errors using an EM approach. More details can be found in Vandekerkhove Vandekerkhove (Vandekerkhove). The latter author suggested to consider that the intercept and the slope of the first component axis were precisely estimated by the values 1.47 and 0.82, respectively, obtained by Martin-Magniette et al. Martin-Magniette et al. (2008), and applied the transformation $y_i = \tilde{y}_i - (1.47 + 0.82x_i)$ to obtain a dataset (x_i, y_i) that fits into the setting considered in this work (centered first component). The transformed dataset scatter plot is displayed in Figure 5 along



(a) Location, $n = 4000,\ 10,000$ and $20,000$.



(b) Scaling, $n = 4000,\ 10,000$ and $20,000$.

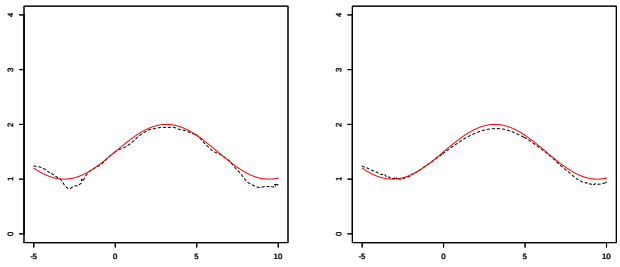


(c) Proportion, $n = 4000,\ 10,000$ and $20,000$.

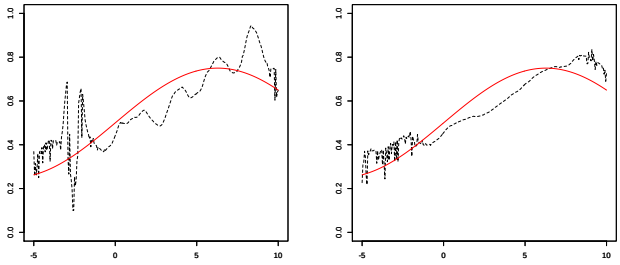
]

Fig 1: Under model **(M1)**, behavior of the supremum error distribution over the location, scaling and proportion, functions (rows) for $n = 4000,\ 10,000$ and $20,000$ (columns). The index under each boxplot corresponds to the value of κ, involved in (7.1), under which the supremum empirical distribution is obtained.

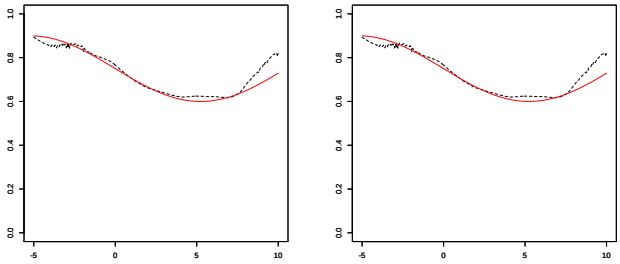
with the linear regression functions obtained by Martin et al. Martin-Magniette et al. (2008) (blue dashed line) and Bordes et al. Bordes et al. (2013) (blue solid line) and the nonlinear regression function fitted by our method (red solid line). Let us also remind that the estimated value of p found by these authors is very similar and about 0.35. In Figure 6 we display the graph of $(x, y) \mapsto 1/\hat{\sigma}_n(x) f(y/\hat{\sigma}_n(x))$ over a (x, y)-grid to illustrate the influence of the scaling on the first component shape population. In Figure 7 we display successively



(a) Location function estimation (dashed line) and true location function (solid red line), $n = 4000$. Left: $\kappa = 4$. Right: $\kappa = 10$.



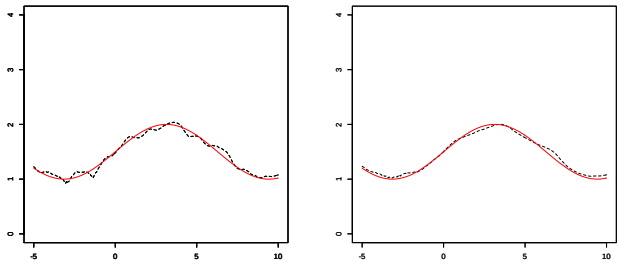
(b) Scaling function estimation (dashed line) and true location function (solid red line), $n = 4000$. Left: $\kappa = 4$. Right: $\kappa = 10$.



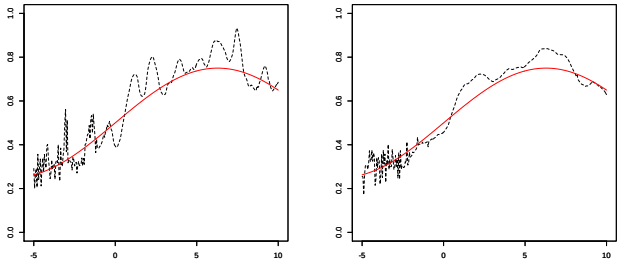
(c) Proportion function estimation (dashed line) and true location function (solid red line), $n = 4000$. Left: $\kappa = 4$. Right: $\kappa = 10$.

Fig 2: One single run fitting example under model **(M1)** : left column, resp. right column, is obtained for $n = 4000$ and $\kappa = 4$, resp. $\kappa = 10$.

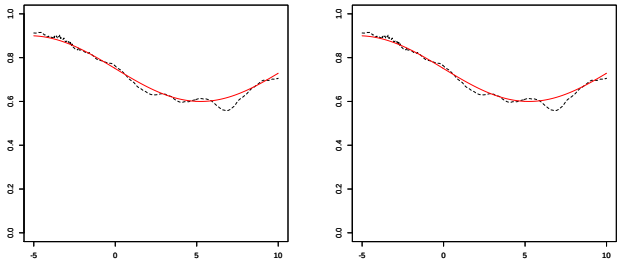
the results obtained on the location, scaling and proportion functions over 10 model fitting attempts, sourcing every time different $10,000$-size samples from



(a) Location function estimation (dashed line) and true location function (solid red line), $n = 10,000$. Left: $\kappa = 4$. Right: $\kappa = 10$.



(b) Scaling function estimation (dashed line) and true location function (solid red line), $n = 10,000$. Left: $\kappa = 4$. Right: $\kappa = 10$.

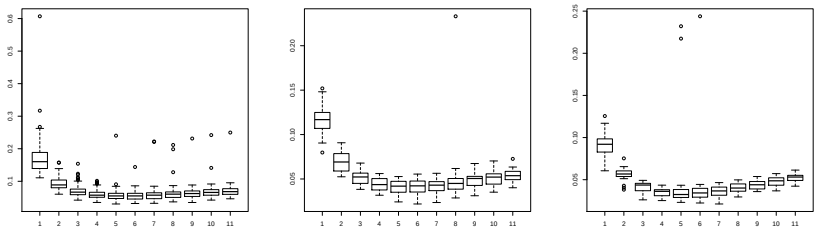


(c) Proportion function estimation (dashed line) and true location function (solid red line), $n = 10,000$. Left: $\kappa = 4$. Right: $\kappa = 10$.

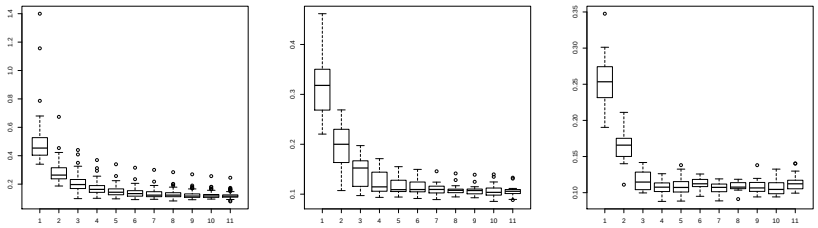
Fig 3: One single run fitting example under model **(M1)** : left column, resp. right column, is obtained for $n = 10,000$ and $\kappa = 4$, resp. $\kappa = 10$.

the transformed NimbleGen dataset.

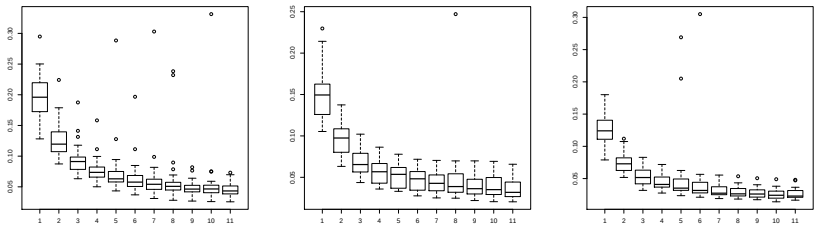
Comments on Figure 5–7. We can observe on Figure 5 that the estimated loca-



(a) Location, $n = 4000,\ 10,000$ and $20,000$.



(b) Scaling, $n = 4000,\ 10,000$ and $20,000$.



(c) Proportion, $n = 4000,\ 10,000$ and $20,000$.

Fig 4: Under model **(M2)**, behavior of the supremum error distribution over the proportion, scaling and location functions (rows) for $n = 4000,\ 10,000$ and $20,000$ (columns). The index under each boxplot corresponds to the value of κ, involved in (7.1), under which the supremum empirical distribution is obtained.

tion function obtained by our method is clearly below the regression lines proposed by Martin-Magniette et al. Martin-Magniette et al. (2008) (blue dashed line) and Bordes et al. Bordes et al. (2013) (blue solid line). The consequence of this is that our method implicitly considers that the unknown component has a more spread out distribution, due to the symmetry assumption, than the one obtained by the previous authors. This remark also implies that there is a stronger overlap between the first and the second component which explains

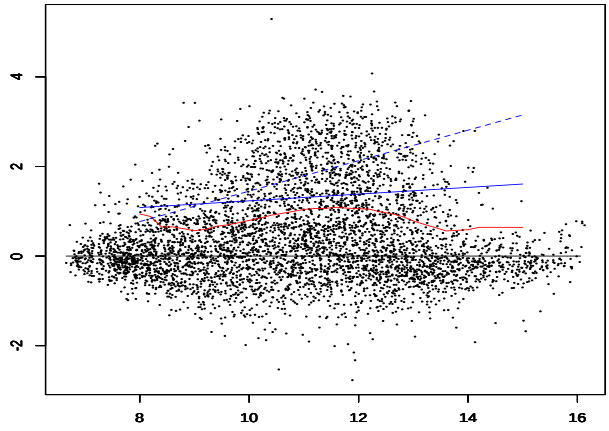


Fig 5: The transformed NimbleGen dataset along with the linear regression functions obtained by Martin-Magniette et al. Martin-Magniette et al. (2008) (blue dashed line), Bordes et al. Bordes et al. (2013) (blue solid line) and the nonlinear regression function fitted by our method (red solid line).

that the clearly visible less dense area lying over the interval $[9, 13]$ is the result of an overlap of a dominant second component (the upper part of the scatter plot keeps being constantly dense) and a weak first component as it is validated by the pattern of the proportion parameter estimates displayed in Figure 7 (c). Further we can observe in Figure 5 that the first component shrinks slightly over the interval $[10, 13]$ which is also detected by our method as it is demonstrated on Figure 7 (b).

8. Local M-estimators and U-processes

8.1. General estimation theory for local M-estimators

In this section we develop rates of convergence and adpative estimation for general local M-estimators. The proofs of the results in this section are given in Section 11.

Let $\Gamma(\alpha)$, $\alpha \in [a, b]$ be sets which index statistical models $(\mathbb{P}_\gamma)_{\gamma\in\Gamma(\alpha)}$ on some measurable space. Let $I \subset \mathbb{R}^d$ be a compact rectangle, and let $\Theta \subseteq \mathbb{R}^m$.

Suppose that the deterministic contrast function $M(\cdot, \cdot; \gamma) : \Theta \times I \to \mathbb{R}$ is uniquely minimized in its first argument by $\theta_*(x; \gamma)$, i.e.

$$\theta_*(x; \gamma) = \underset{\theta\in\Theta}{\operatorname{argmin}}\, M(\theta, x; \gamma) \,. \tag{8.1}$$

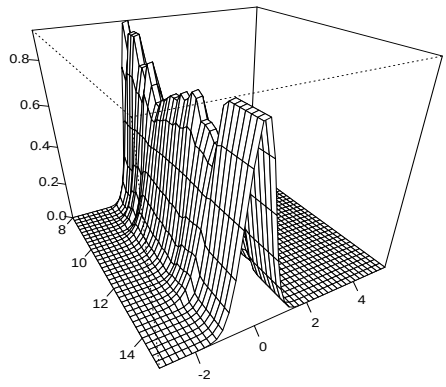


Fig 6: The graph associated to the NimbleGen dataset of the mapping $(x, y) \mapsto 1/\hat{\sigma}_n(x) f(y/\hat{\sigma}_n(x))$ over a (x, y)-grid.

The function $M(\theta, x; \gamma)$ is assumed to be a limiting version of a sequence of random contrast functions $M_n(\theta, x; \alpha)$ under $\mathbb{P}_\gamma$. In our specific model, the parameter α corresponds to the Hölder-degree of smoothness in the previous sections, where $M_n(\theta, x; \alpha) = M_n(\theta, x; h_n(\alpha))$ is given in (4.6) with $h_n(\alpha) = \left(\frac{\log n}{n}\right)^{\frac{1}{2\alpha+d}}$, and $M(\theta, x; \gamma)$ in (4.5).
We suppose that $M_n(\cdot, x; \alpha)$ are minimized by some $\hat{\theta}_n(x; \alpha)$, i.e.

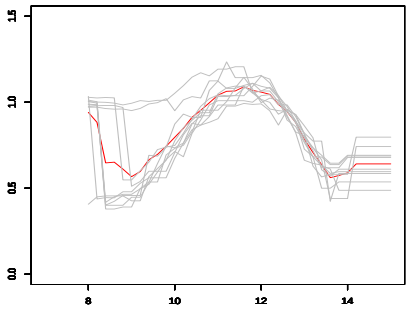
$$\hat{\theta}_n(x; \alpha) \in \operatorname*{argmin}_{\theta \in \Theta} M_n(\theta, x; \alpha) \, . \tag{8.2}$$

Consider the gradients of the contrast functions

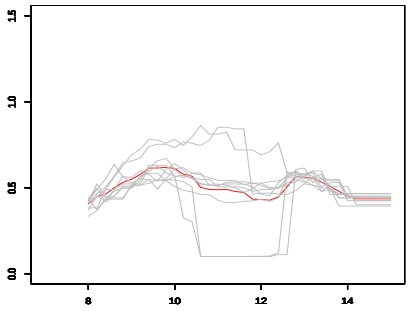
$$S_n(\cdot, \cdot; \alpha) := \partial_\theta M_n(\cdot, \cdot; \alpha) \, , \quad S(\cdot, \cdot; \gamma) := \partial_\theta M(\cdot, \cdot; \gamma) \, , \quad \gamma \in \Gamma(\alpha) \, ,$$

where $\partial_\theta = (\partial_{\theta_1}, \ldots, \partial_{\theta_m})^\top$. We formulate a result on rates of convergence in sup-norm when the nuisance parameter α is known a priori, and subsequently formulate a Lepski-type method to obtain estimates which are adaptive with respect to α. We work with the following high-level assumptions.

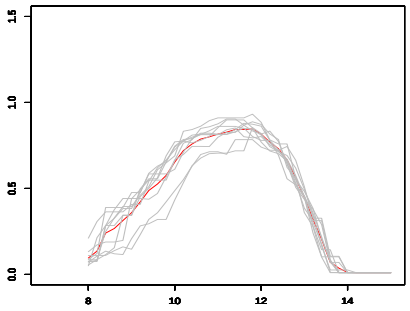
Assumption 4. Assume that Θ is compact and convex with $\Theta = \overline{\operatorname{int}(\Theta)}$. For convenience to avoid boundary issues, assume that the contrast functions are defined on an open and convex set $\Xi \supset \Theta$. Given $0 < a < b < \infty$ and $\alpha \in [a, b]$ let $(\Gamma(\alpha), \|\cdot\|_\alpha)$ be subsets of normed spaces. Further let $r(\alpha) = r(\alpha; n)$ be given rates of convergence, which tend to ∞ in n for given α, and increase in α for given n.



(a) Plot of 10 location function estimation attempts (grey solid lines) along with their average function (red solid line).



(b) Plot of 10 scaling function estimation attempts (grey solid lines) along with their average function (red solid line).



(c) Plot of 10 proportion function estimation attempts (grey solid lines) along with their average function (red solid line).

Fig 7: Plots of 10 estimation attempts based on samples of size $n = 10,000$ from the transformed NimbleGen dataset along with their average function.

(A1) Let $(\Gamma(\alpha), \|\cdot\|_\alpha)$ be 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

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

(A2) The map $(\theta, x; \gamma) \mapsto M(\theta, x; \gamma)$ is continuous. Further for every $x \in I$, $\alpha \in [a, b]$, $\gamma \in \Gamma(\alpha)$, the contrast $M(\cdot, x; \gamma)$ attains a unique minimum at $\theta_*(x; \gamma)$, and the map $(x; \gamma) \mapsto \theta_*(x; \gamma)$ is continuous.

(A3) For all $x \in I$, $\alpha \in [a, b]$, $\gamma \in \Gamma(a)$, the function $M(\cdot, x; \gamma)$ is twice continuously differentiable in its first argument and the Hessian matrix

$$V_x\big(\theta_*(x;\gamma);\gamma\big) := \partial_\theta \, \partial_\theta^\top \, M\big(\theta_*(x;\gamma), x; \gamma\big)$$

is positive definite. In particular the eigenvalues $\lambda^1_{x,\gamma;\alpha} \geq \cdots \geq \lambda^m_{x,\gamma;\alpha}$ of the matrices $V_x\big(\theta_*(x;\gamma);\gamma\big)$ are positive. Furthermore, the map

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

is continuous.

(A4) The Hessian matrices $V_x(\cdot;\gamma)$ are uniformly Lipschitz continuous in θ, i.e. for all $\theta, \theta' \in \Xi$, we have

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

where the Lipschitz constant $L_{\text{Hess}} < \infty$ depends only on Ξ, I, a, b and $\Gamma(\alpha)$.

(A5) The empirical contrast is continuously differentiable in its first argument and for the gradients

$$S_n(\theta, x; \alpha) := \partial_\theta M_n(\theta, x; \alpha) \, , \quad S(\theta, x; \gamma) := \partial_\theta M(\theta, x; \gamma) \tag{8.3}$$

it holds that for some $C^{**} < \infty$,

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

(A6) The empirical contrast M_n is uniformly consistent for M, i.e. for $\varepsilon > 0$ it holds that

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

Theorem 8.1 (General rate of convergence: twice differentiable contrast)**.** *Under Assumption 4,* **(A1)** - **(A6)**, *for any $\alpha \in [a, b]$, every sequence $\hat\theta_n(x; \alpha)$ of minimizers in* (8.2) *satisfies*

$$\lim_{\delta\to\infty} \limsup_{n\to\infty} \sup_{\gamma\in\Gamma(\alpha)} \mathbb{P}_\gamma \Big(r(\alpha)^{-1} \sup_{x\in I} \big\|\hat\theta_n(x;\alpha) - \theta_*(x;\gamma)\big\| \geq \delta \Big) = 0 \, .$$

The result shows that under the conditions of the theorem, the local M-estimator $\hat{\theta}_n(x;\alpha)$ inherits its rate of convergence from that of the gradients as stated in **(A5)**. In our setting, this rate will be the sup-norm rate in d dimensions over α-Hölder classes, that is $r(\alpha) = \big(\log n/n\big)^{\frac{\alpha}{2\alpha+d}}$.
Let us turn to adaptive estimation with respect to α. Our approach will be to use the Lepski method for the gradients S_n in (8.3) and hence to obtain a data driven nuisance parameter $\hat{\alpha}_n \in [a,b]$ so that

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

and then to use the estimator $\hat{\theta}_n(x;\hat{\alpha}_n)$. As in Section 6 we let $\alpha_k = a + k\,(b-a)/N$, $k = 0,\dots,N = \lceil \log n\rceil$ and $r_k = r(\alpha_k)$. For the choice

$$\hat{k}_n = \hat{k} = \max\Big\{0 \le k \le N \mid \sup_{x\in I,\theta\in\Theta} \big\|S_n(\theta,x;\alpha_k) - S_n(\theta,x;\alpha_l)\big\| \le C_{\mathrm{Lep}} r_l \quad \forall\, 0 \le l \le k\Big\}\,, \tag{8.4}$$

where the Lepski constant $C_{\mathrm{Lep}} < \infty$ has to be chosen large enough, we let $\hat{\alpha}_n = \alpha_{\hat{k}}$ and

$$\widehat{\theta}^{\mathrm{ad}}_n(x) = \hat{\theta}_n(x;\hat{\alpha}_n) = \operatorname*{argmin}_{\theta\in\Theta} M_n(\theta,x;\hat{\alpha}_n)\,. \tag{8.5}$$

The following high-level assumption allows to bound the probability of stopping early in the selection rule (8.4).

(A7) 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

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

C^{**} is specified in **(A5)** and $0 \le k_n(\alpha) \le N-1$ is chosen so that $\alpha_{k_n(\alpha)} \le \alpha \le \alpha_{k_n(\alpha)+1}$.

Theorem 8.2 (General rate of convergence: Adaptivity)**.** *Under Assumption 4,* **(A1)** - **(A6)** *and* **(A7)** *for sufficiently large choice of* C_{Lep}*, the estimator* $\widehat{\theta}^{ad}_n(\cdot)$ *defined in* (8.5) *satisfies*

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

8.2. Uniform bounds for U-processes

In this section we provide tools which allow us to deal with the stochastic components in the high-level assumptions **(A5)**, **(A6)** and **(A7)** in case the contrast function is a local U-statistic such as

$$M_n(\theta, x; h) := \frac{1}{n(n-1)} \sum_{1 \le j \ne k \le n} \tau(Y_j, Y_k, \theta) K_h(X_j - x) K_h(X_k - x) \,. \tag{8.6}$$

Here τ is a smooth function that is symmetric in its first two arguments, K is a kernel function and $h > 0$ is a bandwidth parameter. Proofs for the results in this section are given in Section 12. The first result will be used to take care of **(A6)** as well as of **(A5)** when applied to the coordinates of the gradient w.r.t. θ.

Theorem 8.3 (Uniform stochastic error for U-statistics)**.** *Consider the local U-statistics $M_n(\theta, x; h)$ as in (8.6), where the sequence $(Z_n)_n = \big((Y_n, X_n^\top)^\top\big)_n$ of i.i.d. random vectors have Lebesgue densities*

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

The support $I \subset \mathbb{R}^d$ of ℓ_γ is supposed to be a compact cuboid, and $\sup_{\gamma\in\Gamma} \|\ell_\gamma\|_\infty < \infty$. Further, $K : \mathbb{R}^d \to \mathbb{R}$ is a Lipschitz continuous and bounded L^2-kernel; for some non-empty set A, $(h_n(\alpha))_{n\in\mathbb{N}}$, $\alpha \in A$ is sequences of bandwidth parameters so that

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

and $\tau : \mathbb{R} \times \mathbb{R} \times \Theta \to [0, \infty)$ is a bounded function and $\Theta \subset \mathbb{R}^m$ is a compact and convex set with $\Theta = \overline{\mathrm{int}(\Theta)}$. The function τ is symmetric in its first two arguments and satisfies

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

for some constant $L_\tau < \infty$. Then we have for any $\rho \in [1, \infty)$ and any compact set $J \subset \mathrm{int}(I)$ that

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

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 2. If τ (and hence M_n) take values in $\mathbb{R}^k$ (e.g. the gradient of a U-statistic) it will be enough to check that every coordinate function fulfills the assumptions of Theorem 8.3.

The next result, which takes care of **(A7)**, is directly formulated for the gradient.

Lemma 8.4. *Let M_n be a U-statistic as in (8.6) that is differentiable in θ. Let the assumptions of Theorem 8.3 hold for the coordinates of the gradient*

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

Then for positive constants $\tilde{c}_1, \tilde{c}_2 > 0$, there is an increasing linear function u (depending on $\tilde{c}_1, \tilde{c}_2$) such that for sufficiently large values of C_{Lep} we have that

$$\limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \sup_{0\le l\le j\le k_n(\alpha)} n^{u(C_{\mathrm{Lep}})} \tilde{p}_{lj} < \infty \; , \quad \textit{where}$$

$$\tilde{p}_{lj} = \mathbb{P}\Big(\sup_{x\in J, \theta\in\Theta} \| S_n(\theta, x; h_j) - \mathbb{E}_\gamma\big[S_n(\theta, x; h_j)\big] \| > \big(\tilde{c}_1 C_{\mathrm{Lep}} - \tilde{c}_2\big) r_l \Big) \; .$$

Acknowledgements

The authors would like to thank Cristina Butucea for helpful discussion in the early stage of this project. H.H. gratefully acknowledges financial support from the DFG, grant HO 3260/5-1.

References

Bordes, L., C. Delmas, and P. Vandekerkhove (2006). Semiparametric estimation of a two-component mixture model where one component is known. *Scandinavian Journal of Statistics 33*(4), 733–752.

Bordes, L., I. Kojadinovic, and P. Vandekerkhove (2013). Semiparametric estimation of a mixture of two linear regressions in which one component is known. *Preprint*.

Bordes, L., I. Kojadinovic, P. Vandekerkhove, et al. (2013). Semiparametric estimation of a two-component mixture of linear regressions in which one component is known. *Electronic Journal of Statistics 7*, 2603–2644.

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.

Butucea, C., R. N. Tzoumpe, P. Vandekerkhove, et al. (2017). Semiparametric topographical mixture models with symmetric errors. *Bernoulli 23*(2), 825–862.

Butucea, C. and P. Vandekerkhove (2014). Semiparametric mixtures of symmetric distributions. *Scandinavian Journal of Statistics 41*(1), 227–239.

Chernozhukov, V., D. Chetverikov, K. Kato, et al. (2014). Anti-concentration and honest, adaptive confidence bands. *The Annals of Statistics 42*(5), 1787–1818.

Compiani, G. and Y. Kitamura (2016). Using mixtures in econometric models: a brief review and some new results.

De Veaux, R. D. (1989). Mixtures of linear regressions. *Computational Statistics & Data Analysis 8*(3), 227–245.

Driver, B. K. (2003, June). Analysis tools with applications. *Lecture Notes*.

Giné, E., R. Latała, and J. Zinn (2000). Exponential and moment inequalities for U-statistics. In *High Dimensional Probability II*, Volume 47, pp. 13–38. Springer.

Golubev, F., O. Lepski, and B. Levit (2000). On adaptive estimation using the sup-norm losses.

Hohmann, D. and H. Holzmann (2013). Semiparametric location mixtures with distinct components. *Statistics 47*(2), 348–362.

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.

Hunter, D. R. and D. S. Young (2012). Semiparametric mixtures of regressions. *Journal of Nonparametric Statistics 24*(1), 19–38.

Lepski, O. V., V. G. Spokoiny, et al. (1997). Optimal pointwise adaptive methods in nonparametric estimation. *The Annals of Statistics 25*(6), 2512–2546.

Lepskii, O. (1992). Asymptotically minimax adaptive estimation. i: Upper bounds. optimally adaptive estimates. *Theory of Probability & Its Applications 36*(4), 682–697.

Martin-Magniette, M.-L., T. Mary-Huard, C. Bérard, and S. Robin (2008). Chipmix: mixture model of regressions for two-color chip–chip analysis. *Bioinformatics 24*(16), i181–i186.

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.

Städler, N., P. Bühlmann, and S. Van De Geer (2010). ℓ-1-penalization for mixture regression models. *Test 19*(2), 209–256.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, 1040–1053.

van der Vaart, A. and J. A. Wellner (1996). *Weak convergence and empirical processes. With applications to statistics*. New York, NY: Springer.

Vandekerkhove, P. Journal of nonparametric statistics. *25*, 181–208.

Zhu, H.-T. and H. Zhang (2004). Hypothesis testing in mixture regression models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66*(1), 3–16.

9. Proofs for Sections 3 and 4

The following simple lemma states that from $\mu = \mu_*$ it follows that $\vartheta = \vartheta_*$ under the assumptions of Theorem 3.1.

Lemma 9.1. *Let $\vartheta_i = (p_i, \sigma_i, \mu_i, f_i)^\top$, $i = 1, 2$ be two parameter vectors for the model. If $p_1 \in (0, 1)$, $\mu_1 = \mu_2 \neq 0$, and $f_{\text{mix}}(y; \vartheta_1) = f_{\text{mix}}(y; \vartheta_2)$ 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.*

As mentioned above we assume that $\int y^2 \bar{f}(y)\,\mathrm{d}y = 1$.

Proof. Denote by ξ_i the second order moment of f_i, $i = 1, 2$. Using the symmetry of $\bar{f}, f_1$ and f_2 leads to the moment equations

$$p_1\mu_1 = p_2\mu_2 \,, \tag{9.1}$$
$$(1-p_1)\sigma_1^2 + p_1(\xi_1 + \mu_1^2) = (1-p_2)\sigma_2^2 + p_2(\xi_2 + \mu_2^2) \,, \tag{9.2}$$
$$p_1(3\xi_1\mu_1 + \mu_1^3) = p_2(3\xi_2\mu_2 + \mu_2^3) \,. \tag{9.3}$$

Under the assumptions in Lemma 9.1, (9.1)-(9.3) imply that

$$p_1\mu_1 = p_2\mu_1 \,, \tag{9.4}$$
$$(1-p_1)\sigma_1^2 + p_1(\xi_1 + \mu_1^2) = (1-p_2)\sigma_2^2 + p_2(\xi_2 + \mu_1^2) \,, \tag{9.5}$$
$$p_1(3\xi_1\mu_1 + \mu_1^3) = p_2(3\xi_2\mu_1 + \mu_1^3) \,. \tag{9.6}$$

As $\mu_1 \neq 0$, (9.4) gives $p_1 = p_2$. Then, (9.6) and $p_1\mu_1 \neq 0$ lead to $\xi_1 = \xi_2$, yielding $\sigma_1 = \sigma_2$ by (9.5) as $p_1 \neq 1$. Finally, through

$$\begin{aligned} f_1(y) &= \frac{1}{p_1}\, f_{\mathrm{mix}}(y + \mu_1; \vartheta_1) - \frac{(1-p_1)\bar{f}((y+\mu_1)/\sigma_1)}{\sigma_1 p_1} \\ &\overset{a.s.}{=} \frac{1}{p_2}\, f_{\mathrm{mix}}(y + \mu_2; \vartheta_2) - \frac{(1-p_2)\bar{f}((y+\mu_2)/\sigma_2)}{\sigma_2 p_2} = f_2(y) \end{aligned}$$

we obtain $f_1 = f_2$ almost surely, thus $\vartheta_1 = \vartheta_2$. □

Proof of Theorem 3.1. Assume that

$$f_{\mathrm{mix}}(y; \vartheta_*) = f_{\mathrm{mix}}(y; \vartheta) \quad \text{for almost all} \quad y \in \mathbb{R} \tag{9.7}$$

for some $\vartheta = (p, \sigma, \mu, f)^\top$. Taking the Fourier transform in (9.7), using that the Fourier transforms of $\bar{f}$, f_*, f are real-valued and considering real and imaginary part separately gives for all $t \in \mathbb{R}$ that

$$\begin{aligned} (1-p_*)\varphi_{\bar{f}}(\sigma_* t) - (1-p)\varphi_{\bar{f}}(\sigma t) + p_* \cos(\mu_* t)\varphi_{f_*}(t) &= p\cos(\mu t)\varphi_f(t) \,, \\ p_* \sin(\mu_* t)\varphi_{f_*}(t) &= p\sin(\mu t)\varphi_f(t) \,. \end{aligned} \tag{9.8}$$

Multiplying these equations by $\sin(\mu t)$ and $\cos(\mu t)$, respectively, and using the trigonometric identities yields

$$\big[(1-p_*)\varphi_{\bar{f}}(\sigma_* t) - (1-p)\varphi_{\bar{f}}(\sigma t)\big]\sin(\mu t) = p_*\varphi_{f_*}(t)\sin\big((\mu_* - \mu)t\big) \,, \qquad t \in \mathbb{R} \,. \tag{9.9}$$

Now, as the first moments of $f_{\mathrm{mix}}(\cdot; \vartheta)$ and $f_{\mathrm{mix}}(\cdot; \vartheta_*)$ have to coincide, we have

$$p\mu = p_*\mu_* \,, \tag{9.10}$$

which directly implies $p, \mu \neq 0$.

Proof under Assumption 1. According to (9.10), we conclude that $t = \frac{\pi}{\mu}$ is a zero of the left-hand side of (9.9), giving $\sin\left(\frac{\mu_* - \mu}{\mu}\pi\right) = 0$ 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 (9.10), 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 9.1.

Proof under Assumption 2. Suppose that Condition **(C1)** holds. Assume t to be so large that $\varphi_{f_*}(t) \neq 0$ holds. Dividing (9.9) by $\varphi_{f_*}(t)$ and taking limits in t gives

$$\lim_{t\to\infty} p_* \sin\big((\mu_* - \mu)t\big) = 0$$

according to Condition **(C1)**. As $p_* > 0$ and sin is periodic, it follows $\mu_* = \mu$ and since $\mu_* \neq 0$, we obtain $\vartheta_* = \vartheta$ by Lemma 9.1.

Since for $\mu_* = \mu \neq 0$ identification follows directly by Lemma 9.1, we assume $\mu_* \neq \mu$ and derive a contradiction to show identification under the other conditions.

Now suppose that Condition **(C2)** holds. We need to consider three cases.

Case 1: $\sigma = \sigma_*$. If we divide (9.9) by $\varphi_{\bar{f}}(\sigma_* t)$ and let $t \to \infty$, the right-hand side tends to 0 and hence

$$\lim_{t\to\infty} \big((1 - p_*) - (1 - p)\big) \sin(\mu t) = 0 \,.$$

As $\mu \neq 0$, this is only possible if $p = p_*$, in which case (9.10) implies $\mu = \mu_*$, a contradiction.

Case 2: $\sigma < \sigma_*$. If we divide (9.9) 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 (9.9) reduces to

$$(1 - p_*)\varphi_{\bar{f}}(\sigma_* t) \sin(\mu t) = p_* \varphi_{f_*}(t) \sin\big((\mu_* - \mu)t\big) \,, \quad t \in \mathbb{R} \,.$$

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 (9.9) 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.

□

Proof of Proposition 4.1. Since $q > 0$, by continuity it suffices to prove the equivalence

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

By (4.2) we have that $\mathbb{E}_{\vartheta_*}\big[H(Y,t,\theta_*)\big] = 0$ for all $t \in \mathbb{R}$.
For the converse, suppose now that $\theta \in [0,1] \times (0,\infty) \times \mathbb{R}$ is such that for all $t \in \mathbb{R}$,

$$\mathbb{E}_{\vartheta_*}\big[H(Y,t,\theta)\big] = \mathbb{E}_{\vartheta_*}\big[\sin\big((Y-\mu)\,t\big)\big] + (1-p)\,\varphi_{\bar{f}}(\sigma\,t)\,\sin(t\mu) = 0\,.$$

Since

$$\mathbb{E}_{\vartheta_*}\big[\sin\big((Y-\mu)t\big)\big] = \Im\bigg(\int e^{it(y-\mu)} f_{\mathrm{mix}}(y;\vartheta_*)\,\mathrm{d}y\bigg) = \Im\big(\varphi_{f_{\mathrm{mix}}(\cdot+\mu;\vartheta_*)}(t)\big)$$

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 $t \in \mathbb{R}$,

$$\mathbb{E}_{\vartheta_*}\big[H(Y,t,\theta)\big] = \Im\left(\varphi_{f_{\mathrm{mix}}(\cdot+\mu;\vartheta_*)-\frac{1-p}{\sigma}\bar{f}\left(\frac{\cdot+\mu}{\sigma}\right)}(t)\right) = 0\,.$$

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 (9.8), 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 1 is fulfilled we can repeat the proof of Theorem 3.1 starting after (9.9). Note that we cannot use Theorem 3.1 to confirm the result because $\tau(\cdot;\theta|\vartheta_*)$ does not have to be a density. The same method works under Assumption 2. Finally the contrast property for M is straightforward since q is a strictly positive weight function over $\mathbb{R}$. □

10. Proofs of Theorems 5.1 and 6.1

10.1. Outline

In this section we provide the proofs of Theorems 5.1 and 6.1. The strategy is to check the assumptions **(A1)** - **(A6)** as well as **(A7)** in Section 8.1 for our particular model, and then to apply Theorems 8.1 and 8.2.
Assumption **(A1)** is satisfied for Hölder classes if the diameter of the domain I is ≤ 1. The general case can be deduced by rescaling. Hölder spaces are indeed compactly nested, see (Driver, 2003, Theorem 5.14).
The continuity in Assumption **(A2)** follows from the form (4.5) of the contrast together with (10.7) for the conditional expectation, as well as the continuity assumptions in our model. Uniqueness of the minimizer (in fact the zero) follows from Proposition 4.1 applied to the conditional density of Y given X, as well as the assumed positivity of q in (4.4).
The remainder of the section is organized as follows. The main lemmas which take care of **(A3)** - **(A7)** are stated in Section 10.2. Technical lemmas concerning derivatives of the contrast in our model are presented in Section 10.3. The proofs of the main Lemmas 10.1 and 10.2 are then given in Section 10.4.

10.2. Main Lemmas

We choose some open rectangle Ξ with closure $\overline{\Xi}$ contained in $(0,1) \times (0,\infty) \times \mathbb{R}\backslash\{0\}$ which contains the parameterset Θ in (5.2), $\Theta \subset \Xi$.
To prove **(A3)** and **(A4)**, let $V_x(\theta;\gamma)$ denote the Hessian matrix of $M(\cdot,x;\gamma)$ in (4.5) evaluated at θ.

Lemma 10.1. *Let $0 < a \leq b < \infty$. Under Assumptions* **(M2)**, **(M3)**, **(K3)**, *Conditions* **(A3)** *and* **(A4)** *hold for any compact rectangle $J \subset \mathrm{int}(I)$. 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 $\theta, \theta' \in \overline{\Xi}$, we have

$$\sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \sup_{x\in J} \|V_x(\theta;\gamma) - V_x(\theta';\gamma)\| \leq C\|\theta - \theta'\|_1 \ ,$$

where C depends only on $\overline{\Xi}$, I and q.

The following lemma then takes care of **(A5)**, **(A6)** and **(A7)**. In its statement, for the uniform rate for **(A5)** we discuss separately the bias and variance components for the gradients $S_n\big(\theta, x; h\big)$ in (6.1).

Lemma 10.2. *Let $0 < a \leq b < \infty$. Under Assumptions* **(M2)**, **(M3)**, **(K3)**, *for some kernel K fulfilling Assumptions* **(K1)** *and* **($\widetilde{\text{K2}}$)** *and sequences of bandwidth parameters $h_n(\alpha)$, $\alpha \in [a,b]$ so that*

$$\sup_{\alpha\in[a,b]} h_n(\alpha) \to 0, \qquad \sup_{\alpha\in[a,b]} \frac{\log n}{n h_n(\alpha)^d} \to 0 \ ,$$

Conditions **(A5)**, **(A6)** *and* **(A7)** *hold for any compact cuboid* $J \subset \mathrm{int}(I)$. *To be specific on* **(A5)**, *we have for any compact retangle* $J \subset \mathrm{int}(I)$

$$\limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \sup_{x\in J,\theta\in\Theta} h_n(\alpha)^{-\alpha} \big\| \mathbb{E}_\gamma\big[S_n\big(\theta,x;h_n(\alpha)\big)\big] - S(\theta,x;\gamma)\big\| \le C_* \,, \tag{10.1}$$

$$\limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(a)} \left(\frac{\log n}{n h_n(\alpha)^d}\right)^{-1} \mathbb{E}_\gamma\Bigg[\sup_{x\in J,\theta\in\Theta} \big\| S_n\big(\theta,x;h_n(\alpha)\big) - \mathbb{E}_\gamma\big[S_n\big(\theta,x;h_n(\alpha)\big)\big]\big\|^2 \Bigg] \le C_{\mathrm{STOCH}} \,. \tag{10.2}$$

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*

$$\limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \left(\frac{\log n}{n}\right)^{-\frac{2\alpha}{2\alpha+d}} \mathbb{E}_\gamma\Bigg[\sup_{x\in J,\theta\in\Theta} \big\| S_n\big(\theta,x;h_n(\alpha)\big) - S(\theta,x;\gamma)\big\|^2 \Bigg] \le C \,.$$

The proofs of Lemmas 10.1 and 10.2 are given in Section 10.4.

10.3. Derivatives associated with the contrast function

The following lemma lists the derivatives of the function $H(y,t,\theta)$ in (4.3), as well as some useful bounds.

Lemma 10.3. *The derivatives of the function* $H(y,t,\theta)$ *in* (4.3) *are given by*

$$\begin{aligned}
\partial_p H(y,t,\theta) &= -\,\varphi_{\bar f}(\sigma t)\sin(\mu t) \,,\\
\partial_\sigma H(y,t,\theta) &= t(1-p)\partial\varphi_{\bar f}(\sigma t)\sin(\mu t) \,,\\
\partial_\mu H(y,t,\theta) &= -\,t\cos\big((y-\mu)t\big) + t(1-p)\varphi_{\bar f}(\sigma t)\cos(\mu t) \,.
\end{aligned}$$

Under Assumption **(M3)**, *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) $\sup_{y,t\in\mathbb{R}} \sup_{\theta\in\Xi} |H(y,t,\theta)| \le C$,
(ii) $\sup_{y\in\mathbb{R}} \sup_{\theta\in\Xi} \big\|\partial_\theta H(y,t,\theta)\big\| \le C\big(1+|t|\big)$,
(iii) $\sup_{y\in\mathbb{R}} \sup_{\theta\in\Xi} \big\|\partial_\theta\,\partial_\theta^\top H(y,t,\theta)\big\| \le C\big(1+t^2\big)$,
(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) $\sup_{y\in\mathbb{R}} \big\|\partial_\theta H(y,t,\theta) - \partial_\theta H(y,t,\tilde\theta)\big\| \le C\big(1+t^2\big)\|\theta-\tilde\theta\|$,
(vi) $\sup_{y\in\mathbb{R}} \big\|\partial_\theta\,\partial_\theta^\top H(y,t,\theta) - \partial_\theta\,\partial_\theta^\top H(y,t,\tilde\theta)\big\| \le C\big(1+|t|^3\big)\|\theta-\tilde\theta\|$.

Proof of Lemma 10.3. The derivatives of $H(y,t,\theta)$ are obtained by straightforward calculation. Properties (i)-(iii) are immediate 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

$$\begin{aligned}\left|\partial^k\varphi_{\bar f}(\sigma t)-\partial^k\varphi_{\bar f}(\sigma' t)\right| &\le \int\left|\exp(i\sigma ty)-\exp(i\sigma' ty)\right|\cdot\left|i^k y^k\bar f(y)\right|\mathrm{d}y\\ &\le |t|\,|\sigma-\sigma'|\int|y|^{k+1}\bar f(y)\,\mathrm{d}y\,,\quad k=0,1,2\,. \end{aligned}\tag{10.3}$$

□

Let us now turn to the derivatives of the asymptotic contrast $M(\theta,x;\gamma)$ in (4.5) and its Hessian $V_x(\theta;\gamma)$.

Lemma 10.4. *We have under our assumptions that*

$$\begin{aligned}\partial_\theta M(\theta,x;\gamma) =& 2\int\mathbb{E}_\gamma\Big[H(Y,t,\theta)\Big|X=x\Big]\cdot\mathbb{E}_\gamma\Big[\partial_\theta H(Y,t,\theta)\Big|X=x\Big]q(t)\,\mathrm{d}t\cdot\ell^2(x)\,,\\ V_x(\theta;\gamma) =& 2\int\Big(\mathbb{E}_\gamma\Big[\partial_\theta H(Y,t,\theta)\Big|X=x\Big]^\top\mathbb{E}_\gamma\Big[\partial_\theta H(Y,t,\theta)\Big|X=x\Big]\\ &+\mathbb{E}_\gamma\Big[H(Y,t,\theta)\Big|X=x\Big]\cdot\mathbb{E}_\gamma\Big[\partial^2_{\theta^2}H(Y,t,\theta)\Big|X=x\Big]\Big)q(t)\,\mathrm{d}t\cdot\ell^2(x)\,,\end{aligned}\tag{10.4}$$

where

$$\mathbb{E}_\gamma\Big[\partial_p H(Y,t,\theta)\Big|X=x\Big] = -\varphi_{\bar f}(\sigma t)\sin(\mu t)\,,\tag{10.5}$$

$$\mathbb{E}_\gamma\Big[\partial_\sigma H(Y,t,\theta)\Big|X=x\Big] = t(1-p)\partial\varphi_{\bar f}(\sigma t)\sin(\mu t)\,,$$

$$\begin{aligned}\mathbb{E}_\gamma\Big[\partial_\mu H(Y,t,\theta)\Big|X=x\Big] =& \,t\cos(\mu t)\Big((1-p)\varphi_{\bar f}(\sigma t)-\big(1-p_*(x)\big)\varphi_{\bar f}\big(\sigma_*(x)t\big)\Big)\\ &-p_*(x)t\varphi_{f_x^*}(t)\cos\big((\mu_*(x)-\mu)t\big)\,,\end{aligned}\tag{10.6}$$

$$\mathbb{E}_\gamma\Big[\partial^2_{p^2}H(Y,t,\theta)\Big|X=x\Big] = 0\,,$$

$$\mathbb{E}_\gamma\Big[\partial_p\partial_\sigma H(Y,t,\theta)\Big|X=x\Big] = -t\partial\varphi_{\bar f}(\sigma t)\sin(\mu t)\,,$$

$$\mathbb{E}_\gamma\Big[\partial_p\partial_\mu H(Y,t,\theta)\Big|X=x\Big] = -t\varphi_{\bar f}(\sigma t)\cos(\mu t)\,,$$

$$\mathbb{E}_\gamma\Big[\partial^2_{\sigma^2}H(Y,t,\theta)\Big|X=x\Big] = t^2(1-p)\partial^2\varphi_{\bar f}(\sigma t)\sin(\mu t)\,,$$

$$\mathbb{E}_\gamma\Big[\partial_\sigma\partial_\mu H(Y,t,\theta)\Big|X=x\Big] = t^2(1-p)\partial\varphi_{\bar f}(\sigma t)\cos(\mu t)\,,$$

$$\mathbb{E}_\gamma\Big[\partial^2_{\mu^2}H(Y,t,\theta)\Big|X=x\Big] = -t^2\mathbb{E}_\gamma\Big[\sin\big((Y-\mu)t\big)\Big|X=x\Big]-t^2(1-p)\varphi_{\bar f}(\sigma t)\sin(\mu t)\,.$$

Proof of Lemma 10.4. We have that

$$\begin{aligned}
&\mathbb{E}_\gamma\Big[\sin\big((Y-\mu)t\big)\Big|X=x\Big]\\
&=\int \Im\Big(\exp\big(i(y-\mu)t\big)\Big)\Big(\frac{1-p_*(x)}{\sigma_*(x)}\bar{f}\Big(\frac{y}{\sigma_*(x)}\Big)+p_*(x)f_x^*\big(y-\mu_*(x)\big)\Big)\,\mathrm{d}y\\
&=\big(1-p_*(x)\big)\sin(-\mu t)\varphi_{\bar{f}}\big(\sigma_*(x)t\big)+p_*(x)\sin\big((\mu_*(x)-\mu)t\big).\varphi_{f_x^*}(t)\ .
\end{aligned}$$

Therefore, from the definition of $H(y,t,\theta)$ in (4.3) we deduce that

$$\begin{aligned}
\mathbb{E}_\gamma\Big[H(Y,t,\theta)\Big|X=x\Big]&=\sin(\mu t)\Big((1-p)\varphi_{\bar{f}}(\sigma t)-\big(1-p_*(x)\big)\varphi_{\bar{f}}\big(\sigma_*(x)t\big)\Big)\\
&\quad+p_*(x)\sin\big((\mu_*(x)-\mu_*)t\big)\varphi_{f_x^*}(t).
\end{aligned}\tag{10.7}$$

Taking derivatives under the integral gives (10.4). The derivatives (10.5) - (10.6) are obtained by straightforward computation. □

10.4. Proofs of Lemmas 10.1 and 10.2

Proof of Lemma 10.1. (i) Let us start by showing that for each given $x\in J$, 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 t, (10.4) reduces to

$$\begin{aligned}
&V_x(\theta_*(x);\gamma)\\
&=2\int \mathbb{E}_\gamma\Big[\partial_\theta H\big(Y,t,\theta_*(x)\big)\Big|X=x\Big]^\top\mathbb{E}_\gamma\Big[\partial_\theta H\big(Y,t,\theta_*(x)\big)\Big|X=x\Big]q(t)\,\mathrm{d}t\cdot\ell^2(x)\ .
\end{aligned}$$

When inserting the true parameter $\theta_*(x)$, the derivatives (10.5) - (10.6) 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 $v^\top=(v_1,v_2,v_3)\in\mathbb{R}^3$ so that

$$0=v^\top V_x(\theta_*(x);\gamma)v=2\int\Big(\mathbb{E}_\gamma\Big[\partial_\theta H\big(Y,t,\theta_*(x)\big)\Big|X=x\Big]v\Big)^2q(t)\,\mathrm{d}t\cdot\ell^2(x)\ .$$

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

$$\begin{aligned}0 &= v_1\mathbb{E}_\gamma\Big[\partial_p H\big(Y,t,\theta_*(x)\big)\Big|X=x\Big] + v_2\mathbb{E}_\gamma\Big[\partial_\sigma H\big(Y,t,\theta_*(x)\big)\Big|X=x\Big]\\ &\quad + v_3\mathbb{E}_\gamma\Big[\partial_\mu H\big(Y,t,\theta_*(x)\big)\Big|X=x\Big]\\ &= -v_1\varphi_{\bar f}\big(\sigma_*(x)t\big)\sin\big(\mu_*(x)t\big)\\ &\quad + v_2 t\big(1-p_*(x)\big)\partial\varphi_{\bar f}\big(\sigma_*(x)t\big)\sin\big(\mu_*(x)t\big) - v_3 t p_*(x)\varphi_{f_x^*}(t) \qquad (10.8)\\ &=: g(t)\end{aligned}$$

for all $t\in\mathbb{R}$. It remains to show that $v=0$.

First note that the first and second summand in (10.8) 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 (10.8), so that the matrix $V_x(\theta_*(x);\gamma)$ is indeed positive definite.

(ii) This is immediate from (10.4), the Lipschitz continuity of the derivatives in Lemma 10.4, and the fact that q has finite moments of order up to 3.

□

Before turning to the proof of Lemma 10.2 we show two lemmas which are required to deal with the bias in (10.1). The first lemma gives a well-known bound on the bias when using higher-order kernels for functions from Hölder classes.

Lemma 10.5. *Let $0<a\le b<\infty$, $K:\mathbb{R}^d\to\mathbb{R}$ 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 $J\subset\operatorname{int}(I)$, there is some constant $0<C_{\mathrm{Hol}}<\infty$ depending only on $[a,b]$, L, U and K so that*

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

Proof. Fix any $\ell \in \mathcal{H}(\alpha, L, U)$, $x \in J$, $\alpha \in [a, b]$ 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

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

because according to the multinomial theorem, we have

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

□

Lemma 10.6. *If the kernel K fulfills Assumptions* **(K1)** *and* **($\widetilde{\text{K2}}$)**, *then under Assumptions* **(M2)**, **(M3)**, *for any compact rectangle $J \subset \mathrm{int}(I)$ there exists a $C > 0$ such that*

$$\begin{aligned}
&\sup_* h^{-\alpha} \sup_{x \in J, \theta \in \Theta} \Big| \mathbb{E}_\gamma\big[H(Y, t, \theta) \big| X = x\big] - \Big(\mathbb{E}_\gamma\big[H(Y, t, \theta) \big| X = \cdot\big] * K_h\Big)(x) \Big| \leq C\big(1 + |t|\big)\,, \\
&\sup_* h^{-\alpha} \sup_{x \in J, \theta \in \Theta} \Big| \mathbb{E}_\gamma\big[\partial_\theta H(Y, t, \theta) \big| X = x\big] \\
&\qquad\qquad - \Big(\mathbb{E}_\gamma\big[\partial_\theta H(Y, t, \theta) \big| X = \cdot\big] * K_h\Big)(x) \Big| \leq C\big(1 + t^2\big)\,,
\end{aligned}$$

where the suprema are taken over $\alpha \in [a, b]$, $\gamma \in \Gamma(\alpha)$, $h \in (0, \infty)$.

Proof of Lemma 10.6. Consider the first statement. For $\alpha \in [a, b]$, $\gamma \in \Gamma(\alpha)$,

$h \in (0,\infty)$ and $x \in J$ we estimate

$$\begin{aligned}
&\Big|\mathbb{E}_\gamma\big[H(Y,t,\theta)\big|X=x\big] - \Big(\mathbb{E}_\gamma\big[H(Y,t,\theta)\big|X=\cdot\big] * K_h\Big)(x)\Big| \\
&= \left|\int \Big(\mathbb{E}_\gamma\big[H(Y,t,\theta)\big|X=x\big] - \mathbb{E}_\gamma\big[H(Y,t,\theta)\big|X=z+x\big]\Big)K_h(z)\,\mathrm{d}z\right| \\
&= \left|\int \Big(\mathbb{E}_\gamma\big[\sin\big((Y-\mu)t\big)\big|X=x\big] - \mathbb{E}_\gamma\big[\sin\big((Y-\mu)t\big)\big|X=z+x\big]\Big)K_h(z)\,\mathrm{d}z\right| \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(by (10.7))} \\
&= \Bigg|\int \Big(\sin(-\mu t)\big(1-p_*(x)\big)\varphi_{\bar f}\big(\sigma_*(x)t\big) - \sin(-\mu t)\big(1-p_*(z+x)\big)\varphi_{\bar f}\big(\sigma_*(z+x)t\big) \\
&\quad + \sin\big((\mu_*(x)-\mu)t\big)p_*(x)\varphi_{f^*_x}(t) - \sin\big((\mu_*(z+x)-\mu)t\big)p_*(z+x)\varphi_{f^*_{z+x}}(t)\Big)K_h(z)\,\mathrm{d}z\Bigg| \\
&\le \bar C\,\big(1+|t|\big)\left|\int\Big(\big(p_*(x)-p_*(z+x)\big)\right. \\
&\qquad\qquad \left. + \big(\sigma_*(x)-\sigma_*(z+x)\big) + \big(\mu_*(x)-\mu_*(z+x)\big)\Big)K_h(z)\,\mathrm{d}z\right| \qquad (10.9) \\
&\quad + \bar C\left|\int\Big(\varphi_{f^*_x}(t) - \varphi_{f^*_{z+x}}(t)\Big)K_h(z)\,\mathrm{d}z\right|, \qquad (10.10)
\end{aligned}$$

where the inequality follows from the boundednes of characteristic functions by 1, by boundedness and Lipschitz continuity of sin, cos, by compactness of $\overline{\Xi}$ and by (10.3) for $k=0$.
The term (10.9) is treated directly by Lemma 10.5, which gives a standard bias estimate for Hölder functions using higher-order kernels. The term (10.10) 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 10.5 is applicable again. The second estimate follows by similar calculations. □

Proof of Lemma 10.2. First let us prove (10.2). We shall show that the assumptions of Theorem 8.3 are fulfilled. The gradient of the empirical contrast M_n is given by

$$S_n(\theta,x;h) = \frac{2}{n(n-1)}\sum_{1\le j\ne k\le n}\int H(Y_j,t,\theta)\partial_\theta H(Y_k,t,\theta)q(t)\,\mathrm{d}t \\ K_h(X_j-x)K_h(X_k-x)\,.$$

According to Lemma 10.3 (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 8.3, from which we obtain (10.2).

Second, let us prove (10.1). We will show that for all θ, x, α, γ, h, we have

$$\begin{aligned}
&\big\|\mathbb{E}_\gamma\big[S_n(\theta,x;h)\big] - S(\theta,x;\gamma)\big\| \\
&\le 2\int \Big\|\Big(\big(\mathbb{E}_\gamma\big[H(Y,t,\theta)\big|X=\cdot\big]\ell\big) * K_h\Big)(x)\cdot\Big(\Big(\mathbb{E}_\gamma\big[\partial_\theta H(Y,t,\theta)\big|X=\cdot\big]\ell\Big) * K_h\Big)(x) \\
&\qquad - \ell^2(x)\mathbb{E}_\gamma\big[H(Y,t,\theta)\big|X=x\big]\cdot\mathbb{E}_\gamma\big[\partial_\theta H(Y,t,\theta)\big|X=x\big]\Big\| q(t)\,\mathrm{d}t \qquad (10.11)\\
&\lesssim h^\alpha.
\end{aligned}$$

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 (10.11). 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 10.3 (i) and (ii), it is enough to examine occurring differences. We estimate

$$\begin{aligned}
&\Big|\Big(\big(\mathbb{E}_\gamma\big[H(Y,t,\theta)\big|X=\cdot\big]\ell\big) * K_h\Big)(x) - \ell(x)\mathbb{E}_\gamma\big[H(Y,t,\theta)\big|X=x\big]\Big| \qquad (10.12)\\
&\le \Big|\Big(\big(\mathbb{E}_\gamma\big[H(Y,t,\theta)\big|X=\cdot\big]\ell\big) * K_h\Big)(x) - \ell(x)\Big(\big(\mathbb{E}_\gamma\big[H(Y,t,\theta)\big|X=\cdot\big]\big) * K_h\Big)(x)\Big| \\
&\quad + \Big|\ell(x)\Big(\big(\mathbb{E}_\gamma\big[H(Y,t,\theta)\big|X=\cdot\big]\big) * K_h\Big)(x) - \ell(x)\mathbb{E}_\gamma\big[H(Y,t,\theta)\big|X=x\big]\Big|,
\end{aligned}$$

where the first summand is treated by Lemma 10.3 (i) and the fact that ℓ is Hölder-α-smooth, so that by Lemma 10.5,

$$\big|\big(\ell * K_h\big)(x) - \ell(x)\big| \lesssim h^\alpha\,.$$

The second summand is dealt with by Lemma 10.6 so that

$$(10.12) \lesssim h^\alpha\big(1+|t|\big)\,.$$

Analogously, we derive that

$$\begin{aligned}
&\Big\|\Big(\big(\mathbb{E}_\gamma\big[\partial_\theta H(Y,t,\theta)\big|X=\cdot\big]\ell\big) * K_h\Big)(x) - \ell(x)\mathbb{E}_\gamma\big[\partial_\theta H(Y,t,\theta)\big|X=x\big]\Big\| \\
&\le \Big\|\Big(\big(\mathbb{E}_\gamma\big[\partial_\theta H(Y,t,\theta)\big|X=\cdot\big]\ell\big) * K_h\Big)(x) - \ell(x)\Big(\big(\mathbb{E}_\gamma\big[\partial_\theta H(Y,t,\theta)\big|X=\cdot\big]\big) * K_h\Big)(x)\Big\| \\
&\quad + \Big\|\ell(x)\Big(\big(\mathbb{E}_\gamma\big[\partial_\theta H(Y,t,\theta)\big|X=\cdot\big]\big) * K_h\Big)(x) - \ell(x)\mathbb{E}_\gamma\big[\partial_\theta H(Y,t,\theta)\big|X=x\big]\Big\| \\
&\lesssim h^\alpha\big(1+t^2\big).
\end{aligned}$$

Since q has finite third moments, (10.1) follows. Together, (10.1) and (10.2) imply **(A5)**. Lemma 8.4 directly gives **(A7)**. Finally, **(A6)** is obtained similarly but simpler than **(A5)**. This concludes the proof of the lemma.

□

11. Proofs for Section 8.1

This section provides the proofs for Theorems 8.1 and 8.2. It is organized as follows. In Section 11.1, in Theorems 11.1 and 11.2 we extend results in van der Vaart and Wellner (1996) for consistency and rates of convergence of M-estimators, specifically (van der Vaart and Wellner, 1996, Theorem 3.2.3) and (van der Vaart and Wellner, 1996, Corollary 3.2.5) by making them uniform over the probability model as well as introducing a covariate parameter x. The proof of Theorem 8.1 in Section 11.2 then requires to check the assumptions of Theorem 11.2. For the adaptive result, Theorem 8.2, we show in Lemma 11.3 that the Lepski-choice adaptively estimates the gradient of the contrast. Then Theorem 11.2 can again be used to obtain the adaptive rate of convergence for $\widehat{\theta}_n^{\mathrm{ad}}(\cdot)$. Finally, the proofs of Theorems 11.1 and 11.2 are provided in Section 11.3.

11.1. Consistency and rates of uniform convergence for M-estimators

We start with the following general results on consistency and uniform rates of convergence, the proofs of which are provided in Section 11.3. We fix a parameter value α, and drop it in the notation, and also write $\Gamma = \Gamma(\alpha)$. For brevity, we shall also often write $\theta_* = \theta_*(x,\gamma)$ for the minimizer in (8.1).

Theorem 11.1 (Uniform consistency)**.** *Let Θ be a normed space with norm $\|\cdot\|$ and assume that*

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

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 $M(\theta,x;\gamma) - M(\theta_*,x;\gamma) < \eta$ we have $\|\theta - \theta_*\| < \varepsilon$.*

Then the estimator $\hat{\theta}_n(\cdot)$ is uniformly consistent, i.e. for all $\varepsilon > 0$, we have

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

The following theorem is a generalization of (van der Vaart and Wellner, 1996, Theorem 3.2.5) that gives conditions for uniform convergence rates uniformly over the model parameters γ for possibly unidentifiable models.

Theorem 11.2 (Rate of convergence: General result in sup-norm)**.** *Let the following assumptions be satisfied.*

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

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

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

where the third infimum is taken over $\{\theta \in \Theta : \|\theta - \theta_*\| = \varepsilon\}$*, the fourth supremum is taken over* $\{\theta \in \Theta : \|\theta - \theta_*\| \leq \varepsilon\}$ *and* θ_* *is the minimizer of* $M(\cdot, x; \gamma)$*. Furthermore,* $W_n(\theta, x; \gamma) := M_n(\theta, x) - M(\theta, x; \gamma)$*,* $\phi_n : (0, \infty) \to (0, \infty)$ *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* $\sup_{\gamma\in\Gamma} \mathbb{P}_\gamma\big(\sup_{x\in I} \big\|\hat{\theta}_n(x) - \theta_*\big\| \geq \delta\big) = o(1)$*.*

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*

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

11.2. Proofs of Theorems 8.1 and 8.2

Proof of Theorem 8.1. We need to check the assumptions of Theorem 11.2 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 $\varepsilon < \bar{\varepsilon}$. Then for any $\gamma \in \Gamma$, $x \in I$,

$$\big\{\theta \in \Xi : \|\theta - \theta_*(x; \gamma)\| \leq \varepsilon\big\} \subset \Xi \,.$$

Let us prove the first point of (i). For any $\gamma \in \Gamma$, $x \in I$, 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

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

according to **(A3)** and **(A4)**, where $\lambda^m_{x,\gamma}$ 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 **(A3)**, so that we can deduce

$$\inf_{\gamma\in\Gamma}\inf_{x\in I}\lambda^m_{x,\gamma}>0$$

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

Let us prove the second part of (i).
By applying the fundamental theorem of calculus on the path $[\theta,\theta_*]$ using the functions

$$t\mapsto W_n\big(\theta+\tfrac{t}{\|\theta-\theta_*\|}\cdot(\theta_*-\theta),x;\gamma\big),\quad t\in[0,\|\theta-\theta_*\|]$$

gives for any n, γ that

$$
\begin{aligned}
&\sup_{x\in I}\sup_{\substack{\theta\in\Theta:\\ \|\theta-\theta_*(x;\gamma)\|\leq\varepsilon}}\big|W_n(\theta,x;\gamma)-W_n\big(\theta_*(x;\gamma),x;\gamma\big)\big|\\
&\leq\sup_{x\in I}\sup_{\substack{\theta\in\Theta:\\ \|\theta-\theta_*(x;\gamma)\|\leq\varepsilon}}\int_0^{\|\theta-\theta_*\|}\left|\tfrac{\partial}{\partial t}W_n\big(\theta+\tfrac{t}{\|\theta-\theta_*\|}\cdot(\theta_*-\theta),x;\gamma\big)\Big|_{t=s}\right|\mathrm{d}s\\
&\leq\varepsilon\sup_{x\in I}\sup_{\substack{\theta\in\Theta:\\ \|\theta-\theta_*(x;\gamma)\|\leq\varepsilon}}\sup_{\vartheta\in[\theta,\theta_*]}\left|\partial_{\frac{\theta_*-\theta}{\|\theta-\theta_*\|}}W_n(\vartheta,x;\gamma)\right|\\
&\leq\varepsilon\sup_{x\in I}\sup_{\theta\in\Theta}\big\|S_n(\theta,x)-S(\theta,x;\gamma)\big\|_1\,,
\end{aligned}
$$

where we bounded the directional derivatives by the gradient. Hence, the second part of (i) is given directly by **(A5)**.

We will prove (ii), i.e. the uniform consistency of $\hat{\theta}_n(\cdot)$, by using Theorem 11.1. As uniform consistency of the contrast M_n is given by **(A6)**, only $(*)$ in the assumptions of Theorem 11.1 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 $n \in \mathbb{N}$

$$M(\theta_n, x_n; \gamma_n) - M\big(\theta_*(x_n; \gamma_n), x_n; \gamma_n\big) < \eta_n\,, \quad \big\|\theta_n - \theta_*(x_n; \gamma_n)\big\| \geq \varepsilon\,. \tag{11.1}$$

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 **(A2)**,we have $M(\theta', x'; \gamma') = M(\theta_*(x'; \gamma'), x'; \gamma')$. Now, according to the right-hand side of (11.1), we have

$$\big\|\theta' - \theta_*(x'; \gamma')\big\| \geq \liminf_{k\to\infty} \big\|\theta_{n_k} - \theta_*(x_{n_k}; \gamma_{n_k})\big\| \geq \varepsilon\,,$$

a contradiction as $\theta \mapsto M(\theta, x'; \gamma')$ is only minimized at $\theta_*(x'; \gamma')$.

□

The main ingredient in the proof of Theorem 8.2 is the following lemma.

Lemma 11.3. *Under the assumptions of Theorem 8.1 we have that*

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

Proof of Theorem 8.2. Using Lemma 11.3, the proof works analogously to the one of Theorem 8.1. We shall apply Theorem 11.2 for $M_n(\cdot,\cdot) := M_n(\cdot,\cdot;\hat{\alpha}_n)$, $S_n(\cdot,\cdot) := S_n(\cdot,\cdot;\hat{\alpha}_n)$ and $\Gamma = \big\{(\alpha,\gamma) : \alpha \in [a,b], \gamma \in \Gamma(\alpha)\big\}$, which is compact with respect to $\max\{|\cdot|, \|\cdot\|_a\}$, cf. Lemma 11.4 below. Further set $r_{n,\gamma} = t_{n,\gamma} = r(\alpha)^{-1}$, $\phi_n = \mathrm{id}$, $\eta = \varepsilon^*$.
In order to prove uniform consistency of the estimator $\widehat{\theta}_n^{\mathrm{ad}}(\cdot)$, we first use **(A6)**, yielding

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

and then proceed analogously to the proof of Theorem 8.1

□

Lemma 11.4. *Under Assumption* **(A1)**, *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 11.4. 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. □

Proof of Lemma 11.3. 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 $\alpha \in [a,b]$, $\gamma \in \Gamma(\alpha)$

$$\begin{aligned}
&\mathbb{E}_\gamma\Big[\sup_{x\in I,\theta\in\Theta} \big\|S_n(\theta,x;\beta_{\hat{k}}) - S(\theta,x;\gamma)\big\|\Big] \\
\leq &\mathbb{E}_\gamma\Big[\sup_{x\in I,\theta\in\Theta} \big\|S_n(\theta,x;\beta_{\hat{k}}) - S(\theta,x;\gamma)\big\| \mathbb{1}_{\hat{k}\leq k_n(\alpha)-1}\Big] \qquad (11.2)\\
&+ \mathbb{E}_\gamma\Big[\sup_{x\in I,\theta\in\Theta} \big\|S_n(\theta,x;\beta_{\hat{k}}) - S(\theta,x;\gamma)\big\| \mathbb{1}_{\hat{k}\geq k_n(\alpha)}\Big] \,. \qquad (11.3)
\end{aligned}$$

The term (11.3) can be handled by a zero-addition of the term $S_n(\theta,x;\beta_{k_n(\alpha)})$ within the supremum, i.e.

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

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,

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

where we used that the net over $[a,b]$ grows logarithmically. We get the desired bound on (11.3), i.e.

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

Next let us examine (11.2). By using Cauchy-Schwarz' inequality, we get for all $\alpha\in[a,b]$, $\gamma\in\Gamma(\alpha)$ that

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

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

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

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

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

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

yielding

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

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

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

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

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

In summary we obtain

$$\begin{aligned}
&\limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} r(\alpha)^{-1}\mathbb{E}_\gamma\Big[\sup_{x\in I,\theta\in\Theta}\big\|S_n(\theta,x;\hat\alpha_n)-S(\theta,x;\gamma)\big\|\Big]\\
&\le (C_{\mathrm{Lep}}+C^*)\exp(d(b-a))\\
&\quad + C^*\limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)}\\
&\sup_{a\le\beta\le\alpha} r(\alpha)^{-1}\Big(\mathbb{E}_\gamma\Big[\Big(\sup_{x\in I,\theta\in\Theta}\big\|S_n(\theta,x;\beta)-S(\theta,x;\gamma)\big\|\Big)^2\Big]\Big)^{\frac12}\cdot\log(n)^{\frac32}\sup_{0\le l\le j\le k_n(\alpha)} p_{lj}^{\frac12}\,,
\end{aligned}$$

where

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

It remains to show that

$$0=\limsup_{n\to\infty}\Bigg\{\sup_{\alpha\in[a,b]}\sup_{\gamma\in\Gamma(\alpha)}\sup_{a\le\beta\le\alpha} r(\alpha)^{-1}\Big(\mathbb{E}_\gamma\Big[\Big(\sup_{x\in I,\theta\in\Theta}\big\|S_n(\theta,x;\beta)-S(\theta,x;\gamma)\big\|\Big)^2\Big]\Big)^{\frac12} \tag{11.4}$$

$$\cdot\log(n)^{\frac32}\sup_{0\le l\le j\le k_n(\alpha)} p_{lj}^{\frac12}\Bigg\}\,. \tag{11.5}$$

The factor (11.4) is asymptotically dominated by the rate $r(a)r(b)^{-1}$ as can be seen by inserting $r(\beta)r(\beta)^{-1}$ so that

$$(11.4)\le r(a)r(b)^{-1}\sup_{\alpha\in[a,b]}\sup_{\gamma\in\Gamma(\alpha)} r(\alpha)^{-1}\Big(\mathbb{E}_\gamma\Big[\Big(\sup_{x\in I,\theta\in\Theta}\big\|S_n(\theta,x)-S(\theta,x;\gamma)\big\|\Big)^2\Big]\Big)^{\frac12}\,,$$

where the supremum is asymptotically bounded by C^{**} according to **(A5)**. The second factor (11.5) 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 of (11.4). □

11.3. Proofs of Theorems 11.1 and 11.2

Proof of Theorem 11.1. We have that

$$\begin{aligned}
0 &\le \sup_{x\in I}\Big[M\big(\hat\theta_n(x),x;\gamma\big)-M\big(\theta_*(x;\gamma),x;\gamma\big)\Big]\\
&\le \sup_{x\in I}\Big[M\big(\hat\theta_n(x),x;\gamma\big)-M_n\big(\hat\theta_n(x),x\big)\Big]\\
&\quad+\sup_{x\in I}\Big[M_n\big(\hat\theta_n(x),x\big)-M\big(\theta_*(x;\gamma),x;\gamma\big)\Big]\\
&\le \sup_{\theta\in\Theta}\sup_{x\in I}\big|M(\theta,x;\gamma)-M_n(\theta,x)\big|\\
&\quad+\sup_{x\in I}\Big[M_n\big(\theta_*(x;\gamma),x\big)-M\big(\theta_*(x;\gamma),x;\gamma\big)\Big] \qquad (\hat\theta_n(x)\text{ minim. of }M_n)\\
&\le 2\sup_{\theta\in\Theta}\sup_{x\in I}\big|M(\theta,x;\gamma)-M_n(\theta,x)\big|\,. \qquad (11.6)
\end{aligned}$$

Fix $\varepsilon>0$. Because of $(*)$ there is an $\eta>0$ so that for any $\gamma\in\Gamma$, the inequality

$$\sup_{x\in I}\Big[M(\hat\theta_n(x),x;\gamma)-M(\theta_*(x;\gamma),x;\gamma)\Big]<\eta$$

implies

$$\sup_{x\in I}\big\|\hat\theta_n(x)-\theta_*(x;\gamma)\big\|<\varepsilon\,,$$

which implies

$$\begin{aligned}
\Big\{\sup_{x\in I}\big\|\hat\theta_n(x)-\theta_*(x;\gamma)\big\|\ge\varepsilon\Big\}&\subset\Big\{\sup_{x\in I}\Big[M(\hat\theta_n(x),x;\gamma)-M(\theta_*(x;\gamma),x;\gamma)\Big]\ge\eta\Big\}\\
&\subset\Big\{\sup_{\theta\in\Theta}\sup_{x\in I}\big|M(\theta,x;\gamma)-M_n(\theta,x)\big|\ge\eta/2\Big\}\,,
\end{aligned}$$

where we used (11.6) in the second step. Thus, by uniform consistency of the random functions M_n,

$$\lim_{n\to\infty}\sup_{\gamma\in\Gamma}\mathbb{P}_\gamma\Big(\sup_{x\in I}\big\|\hat\theta_n(x)-\theta_*\big\|\ge\varepsilon\Big)=0\,.$$

□

The proof of Theorem 11.2 is similar to (van der Vaart and Wellner, 1996, Theorem 3.2.5).

Proof of Theorem 11.2. For every $n\in\mathbb{N}$, $x\in I$, $\gamma\in\Gamma$, we define a partition of Θ by $\bigcup_{j\in\mathbb{Z}}S_{jnx\gamma}$, where

$$S_{jnx\gamma}=\Big\{\theta\in\Theta:2^{j-1}<r_{n,\gamma}\|\theta-\theta_*\|\le 2^j\Big\}\,.$$

Let us define for any $N, n \in \mathbb{N}$, $\gamma \in \Gamma$ the sets

$$A_{Nn\gamma} := \left\{ r_{n,\gamma} \sup_{x \in I} \left\| \hat{\theta}_n(x) - \theta_* \right\| > 2^N \right\}$$

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

$$\begin{aligned} \mathbb{P}_\gamma(A_{Nn\gamma}) \leq & \sum_{\substack{j \geq N \\ 2^j \leq \eta r_{n,\gamma}}} \mathbb{P}_\gamma\left(\sup_{x\in I} \sup_{\theta \in S_{jnx\gamma}} \left[M_n(\theta_*, x) - M_n(\theta, x) \right] \geq 0 \right) \\ & + \mathbb{P}_\gamma\left(2 \sup_{x \in I} \left\| \hat{\theta}_n(x) - \theta_* \right\| \geq \eta \right) . \end{aligned} \tag{11.7}$$

Therefore, let

$$\begin{aligned} \omega \in & \bigcap_{\substack{j \geq N \\ 2^j \leq \eta r_{n,\gamma}}} \left\{ \sup_{x\in I} \sup_{\theta \in S_{jnx\gamma}} \min_{\theta_* \in \mathfrak{S}_{x;\gamma}} \left[M_n(\theta_*, x) - M_n(\theta, x; \gamma) \right] < 0 \right\} \\ & \cap \left\{ 2 \sup_{x \in I} \left\| \hat{\theta}_n(x) - \theta_* \right\| < \eta \right\} . \end{aligned} \tag{11.8}$$

Then, for all $j \geq N$ with $2^j \leq \eta r_{n,\gamma}$ and all $x \in I$ we have $\hat{\theta}_n(x)(\omega) \notin S_{jnx\gamma}$ because $\hat{\theta}_n(x)$ minimizes $M_n(\cdot, x)$. Hence, for all $x \in I$, either

$$r_{n,\gamma} \left\| \hat{\theta}_n(x)(\omega) - \theta_* \right\| \leq 2^{N-1} \quad \text{or} \quad r_{n,\gamma} \left\| \hat{\theta}_n(x)(\omega) - \theta_* \right\| > 2^{l_\gamma} , \tag{11.9}$$

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

$$r_{n,\gamma} \left\| \hat{\theta}_n(x)(\omega) - \theta_* \right\| > 2^{J-1} .$$

Then,

$$2^{J-1} < r_{n,\gamma} \left\| \hat{\theta}_n(x)(\omega) - \theta_* \right\| < r_{n,\gamma} \eta / 2 < 2^{J-1} ,$$

according to the right-hand side of (11.8), a contradiction. Hence,

$$r_{n,\gamma} \sup_{x \in I} \left\| \hat{\theta}_n(x)(\omega) - \theta_* \right\| \leq 2^{N-1} \leq 2^N$$

according to (11.9), giving $\omega \in A^c_{Nn\gamma}$ and via subadditivity we deduce (11.7). The second summand on the right-hand side of (11.7) converges uniformly over all $\gamma \in \Gamma$ to zero for all $\eta > 0$ according to assumption (ii), i.e.

$$\lim_{n\to\infty} \sup_{\gamma\in\Gamma} \mathbb{P}_\gamma \left(2 \sup_{x\in I} \left\| \hat{\theta}_n(x) - \theta_* \right\| \geq \eta \right) = 0 . \tag{11.10}$$

Hence, it remains to handle the first term in (11.7). Choose $\eta > 0$ so that Assumption (i) of the theorem 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

$$\begin{aligned}
&\sup_{x\in I}\sup_{\theta\in S_{jnx\gamma}}\Big[M(\theta_*,x;\gamma)-M(\theta,x;\gamma)\Big]\\
&=\sup_{x\in I}\sup_{\varepsilon\in\left(2^{j-1}/r_{n,\gamma},2^j/r_{n,\gamma}\right]}\sup_{*}\Big[M\big(\theta_*(x;\gamma),x;\gamma\big)-M(\theta,x;\gamma)\Big]\\
&=\sup_{\varepsilon\in\left(2^{j-1}/r_{n,\gamma},2^j/r_{n,\gamma}\right]}\sup_{x\in I}\sup_{*}\Big[M\big(\theta_*(x;\gamma),x;\gamma\big)-M(\theta,x;\gamma)\Big]\\
&\leq C_1\sup_{\varepsilon\in\left(2^{j-1}/r_{n,\gamma},2^j/r_{n,\gamma}\right]}-\varepsilon^2=-C_1\left(\frac{2^{j-1}}{r_{n,\gamma}}\right)^2
\end{aligned}$$

according to the first part of (i), where the suprema indexed with $*$ are taken over $\big\{\theta\in\Theta:\|\theta-\theta_*\|=\varepsilon\big\}$. Now, (11.7), (11.10), the display above and Markov's inequality give

$$\begin{aligned}
&\lim_{N\to\infty}\limsup_{n\to\infty}\sup_{\gamma\in\Gamma}\mathbb{P}_\gamma(A_{Nn\gamma})\\
&=\lim_{N\to\infty}\limsup_{n\to\infty}\sup_{\gamma\in\Gamma}\mathbb{P}_\gamma\Big(r_{n,\gamma}\sup_{x\in I}\big\|\hat\theta_n(x)-\theta_*\big\|\geq 2^N\Big)\\
&\leq\lim_{N\to\infty}\limsup_{n\to\infty}\sup_{\gamma\in\Gamma}\sum_{\substack{j\geq N\\ 2^j\leq\eta r_{n,\gamma}}}\mathbb{P}_\gamma\Big(\sup_{x\in I}\sup_{\theta\in S_{jnx\gamma}}\Big[M_n(\theta_*,x)-M_n(\theta,x)\Big]\geq 0\Big)\\
&=\lim_{N\to\infty}\limsup_{n\to\infty}\sup_{\gamma\in\Gamma}\sum_{\substack{j\geq N\\ 2^j\leq\eta r_{n,\gamma}}}\mathbb{P}_\gamma\Big(\sup_{x\in I}\sup_{\theta\in S_{jnx\gamma}}\Big[W_n(\theta_*,x;\gamma)-W_n(\theta,x;\gamma)\\
&\qquad\qquad+M(\theta_*,x;\gamma)-M(\theta,x;\gamma)\Big]\geq 0\Big)\\
&\leq\lim_{N\to\infty}\limsup_{n\to\infty}\sup_{\gamma\in\Gamma}\sum_{\substack{j\geq N\\ 2^j\leq\eta r_{n,\gamma}}}\mathbb{P}_\gamma\Big(\sup_{x\in I}\sup_{\theta\in S_{jnx\gamma}}|W_n(\theta_*,x;\gamma)-W_n(\theta,x;\gamma)|\\
&\qquad\qquad+\sup_{x\in I}\sup_{\theta\in S_{jnx\gamma}}\Big[M(\theta_*,x;\gamma)-M(\theta,x;\gamma)\Big]\geq 0\Big)\\
&\leq\lim_{N\to\infty}\limsup_{n\to\infty}\sup_{\gamma\in\Gamma}\sum_{\substack{j\geq N\\ 2^j\leq\eta r_{n,\gamma}}}\mathbb{P}_\gamma\Big(\sup_{x\in I}\sup_{\theta\in S_{jnx\gamma}}|W_n(\theta_*,x;\gamma)-W_n(\theta,x;\gamma)|\geq C_1\frac{2^{2j-2}}{r_{n,\gamma}^2}\Big)\\
&\leq\lim_{N\to\infty}\limsup_{n\to\infty}\sum_{j\geq N}\sup_{\gamma\in\Gamma}\mathbb{1}_{\frac{2^j}{r_{n,\gamma}}\leq\eta}\frac{r_{n,\gamma}^2}{C_1 2^{2j-2}}\,\mathbb{E}_\gamma\left[\sup_{x\in I}\sup_{\theta\in S_{jnx\gamma}}|W_n(\theta_*,x;\gamma)-W_n(\theta,x;\gamma)|\right].
\end{aligned}\tag{11.11}$$

We will use the second point in Assumption (i) of the theorem in order to treat the $\limsup_{n\to\infty}$ term in (11.11) 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 (11.11) 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

$$\limsup_{n\to\infty} \sup_{\gamma\in\Gamma} \sup_{\substack{j\geq N \\ \frac{2^j}{r_{n,\gamma}}\leq\eta}} \frac{t_{n,\gamma}}{\phi_n(2^j/r_{n,\gamma})} \mathbb{E}_\gamma \left[\sup_{x\in I} \sup_{\theta\in S_{jnx\gamma}} |W_n(\theta_*, x; \gamma) - W_n(\theta, x; \gamma)| \right] \leq C_2.$$

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

$$\mathbb{E}_\gamma \left[\sup_{x\in I} \sup_{\theta\in S_{jnx\gamma}} |W_n(\theta_*, x; \gamma) - W_n(\theta, x; \gamma)| \right] \leq \frac{\big(C_2 + \kappa\big)\phi_n(2^j/r_{n,\gamma})}{t_{n,\gamma}} .$$

Hence, for every $\gamma \in \Gamma$, $n \geq n_0$, $j \geq N$ with $2^j \leq \eta r_{n,\gamma}$, the summands in (11.11) can be treated by

$$\begin{aligned} &\frac{r_{n,\gamma}^2}{C_1 2^{2j-2}} \mathbb{E}_\gamma \left[\sup_{x\in I} \sup_{\theta\in S_{jnx\gamma}} |W_n(\theta_*, x; \gamma) - W_n(\theta, x; \gamma)| \right] \\ &\leq \frac{r_{n,\gamma}^2}{C_1 2^{2j-2}} \frac{\big(C_2 + \kappa\big)\phi_n(2^j/r_{n,\gamma})}{t_{n,\gamma}} . \end{aligned} \tag{11.12}$$

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

$$(11.12) \leq \frac{r_{n,\gamma}^2\big(C_2 + \kappa\big)2^{j\alpha}\phi_n(1/r_{n,\gamma})}{C_1 2^{2j-2} t_{n,\gamma}} \leq \frac{4\big(C_2 + \kappa\big)}{C_1} \cdot \left(\frac{1}{2^{2-\alpha}}\right)^j ,$$

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

$$\begin{aligned} &(11.11) \\ &\leq \lim_{N\to\infty} \sum_{j\geq N} \limsup_{n\to\infty} \sup_{\gamma\in\Gamma} \mathbb{1}_{\frac{2^j}{r_{n,\gamma}}\leq\eta} \frac{r_{n,\gamma}^2}{C_1 2^{2j-2}} \mathbb{E}_\gamma \left[\sup_{x\in I} \sup_{\theta\in S_{jnx\gamma}} |W_n(\theta_*, x; \gamma) - W_n(\theta, x; \gamma)| \right] \\ &\leq \lim_{N\to\infty} \sum_{j\geq N} \frac{4\big(C_2 + \kappa\big)}{C_1} \cdot \left(\frac{1}{2^{2-\alpha}}\right)^j = 0 . \end{aligned}$$

□

12. Proofs for Section 8.2

Proof of Theorem 8.3. We will drop dependence of the bandwidth parameters on α and n for convenience but point out where it comes into play. Throughout the proof we 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 or A. All calculations below hold for each $\gamma \in \Gamma$.

We use the classical Hoeffding decomposition to write the centered U-statistic $M_n(\theta,\cdot;h) - \mathbb{E}_\gamma[M_n(\theta,\cdot;h)]$ as a canonical U-statistic and a linear process as follows

$$\begin{aligned} M_n(\theta,x;h) - \mathbb{E}_\gamma[M_n(\theta,x;h)] =& \frac{1}{n(n-1)} \sum_{1\leq j\neq k\leq n} U_n\big(Z_j, Z_k, \theta, x; h\big) \\ &+ \frac{2}{n}\sum_{j=1}^n \Big[u_n^*(Z_j,\theta,x;h) - \mathbb{E}_\gamma\big[u_n(Z_1,Z_2,\theta,x;h)\big]\Big] \\ =:& \, T_n^1(\theta,x;h) + T_n^2(\theta,x;h)\,, \end{aligned} \tag{12.1}$$

where for $z = (z_1, z_2^\top)^\top, w = (w_1, w_2^\top)^\top \in \mathbb{R}\times J$,

$$\begin{aligned} U_n(z,w,\theta,x;h) :=& \, u_n(z,w,\theta,x;h) - u_n^*(z,\theta,x;h) - u_n^*(w,\theta;x;h) \\ &+ \mathbb{E}_\gamma[u_n(Z_1,Z_2,\theta,x;h)]\,, \\ u_n(z,w,\theta,x;h) :=& \, \tau(z_1,w_1,\theta) K_h(z_2-x) K_h(w_2-x)\,, \\ u_n^*(z,\theta,x;h) :=& \, \mathbb{E}_\gamma[u_n(Z_1,z,\theta,x;h)] = \mathbb{E}_\gamma\big[\tau(z_1,Y_1,\theta)K_h(X_1-x)\big]\cdot K_h(z_2-x)\,. \end{aligned}$$

Since $s \mapsto s^\rho$ is convex, we have that

$$\begin{aligned} \big|M_n(\theta,x;h) - \mathbb{E}_\gamma[M_n(\theta,x;h)]\big|^\rho &\leq 2^\rho \left|\frac{1}{2}T_n^1(\theta,x;h) + \frac{1}{2}T_n^2(\theta,x;h)\right|^\rho \\ &\leq 2^{\rho-1}\Big(|T_n^1(\theta,x;h)|^\rho + |T_n^2(\theta,x;h)|^\rho\Big)\,. \end{aligned} \tag{12.2}$$

Let us deal with the term T_n^1 in (12.2). The linear process T_n^2 in (12.2) is dealt with similarly by using the classical Bernstein inequality. Now, for a sequence $\delta_n \to 0$ specified below, there are nets $\Theta_n \times J_n \subset \Theta \times J$ so that

$$\sup_{x\in J}\inf_{y\in J_n}\|x-y\| < \delta_n\,, \qquad \sup_{\vartheta\in\Theta}\inf_{\theta\in\Theta_n}\|\vartheta-\theta\| < \delta_n\,, \qquad \#(\Theta_n\times J_n) \leq C_{\Theta,J}\delta_n^{-d-m}\,, \tag{12.3}$$

where $C_{\Theta,J}$ is independent of n. Then

$$\begin{aligned} &\sup_{x\in J,\theta\in\Theta}|T_n^1(\theta,x;h)|^\rho \\ &\leq 2^{\rho-1}\Bigg(\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta \\ \|x-y\|,\|\vartheta-\theta\|\leq\delta_n}}|T_n^1(\theta,x;h) - T_n^1(\vartheta,y;h)|\Bigg)^\rho + 2^{\rho-1}\Big(\sup_{x\in J_n,\vartheta\in\Theta_n}|T_n^1(\vartheta,x;h)|\Big)^\rho. \end{aligned} \tag{12.4}$$

For the first term in (12.4) we have that

$$\mathbb{E}_\gamma\left[\left(\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\le\delta_n}}|T_n^1(\theta,x;h)-T_n^1(\vartheta,y;h)|\right)^\rho\right]$$

$$\le 3^{\rho-1}\mathbb{E}_\gamma\left[\left(\frac{1}{n(n-1)}\sum_{1\le j\ne k\le n}\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\le\delta_n}}\left|u_n(Z_j,Z_k,\theta,x;h)-u_n(Z_j,Z_k,\vartheta,y;h)\right|\right)^\rho\right] \tag{12.5}$$

$$+3^{\rho-1}\mathbb{E}_\gamma\left[\left(\frac{2}{n}\sum_{j=1}^{n}\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\le\delta_n}}\left|u_n^*(Z_j,\theta,x;h)-u_n^*(Z_j,\vartheta,y;h)\right|\right)^\rho\right] \tag{12.6}$$

$$+3^{\rho-1}\left(\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\le\delta_n}}\left|\mathbb{E}_\gamma\left[u_n(Z_1,Z_2,\theta,x;h)-u_n(Z_1,Z_2,\vartheta,y;h)\right]\right|\right)^\rho. \tag{12.7}$$

Let us bound these terms. The summands in (12.5) are bounded by

$$\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\le\delta_n}}\left|\tau(Y_j,Y_k,\vartheta)\cdot\Big(K_h(X_j-x)K_h(X_k-x)-K_h(X_j-y)K_h(X_k-y)\Big)\right|$$

$$+\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\le\delta_n}}\left|\Big[\tau(Y_j,Y_k,\vartheta)-\tau(Y_j,Y_k,\theta)\Big]\cdot K_h(X_j-y)K_h(X_k-y)\right|,$$

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

$$\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\le\delta_n}}\left|\tau(Y_j,Y_k,\vartheta)\cdot\big(K_h(X_j-x)K_h(X_k-x)-K_h(X_j-y)K_h(X_k-y)\big)\right|$$

$$\le 2\|\tau\|_\infty L_K\|K\|_\infty\frac{\delta_n}{h^{2d+1}}.$$

By using the Lipschitz continuity of τ in its third argument, we derive for the second summand that

$$\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\le\delta_n}}\left|\Big[\tau(Y_j,Y_k,\vartheta)-\tau(Y_j,Y_k,\theta)\Big]\cdot K_h(X_j-y)K_h(X_k-y)\right|$$

$$\le L_\tau\delta_n\sup_{\substack{x,y\in J\\ \|x-y\|\le\delta_n}}\left|K_h(X_j-y)K_h(X_k-y)\right|\le L_\tau\|K\|_\infty^2\frac{\delta_n}{h^{2d}}.$$

Hence,

$$(12.5) \leq 3^{\rho-1}\left(2\|\tau\|_\infty L_K \|K\|_\infty \frac{\delta_n}{h^{2d+1}} + L_\tau \|K\|_\infty^2 \frac{\delta_n}{h^{2d}}\right)^\rho \lesssim \frac{\delta_n^\rho}{h^{(2d+1)\rho}} .$$

Using similar arguments, we observe that the summands in (12.6) are bounded by

$$\sup_{\substack{x\in J,\vartheta,\theta\in\Theta \\ \|\vartheta-\theta\|\leq\delta_n}} \left|u_n^*(Z_j,\vartheta,x;h) - u_n^*(Z_j,\theta,x;h)\right| + \sup_{\substack{x,y\in J,\theta\in\Theta \\ \|x-y\|\leq\delta_n}} \left|u_n^*(Z_j,\theta,x;h) - u_n^*(Z_j,\theta,y;h)\right|$$
$$\leq \frac{\|K\|_\infty L_\tau \|\ell_\gamma\|_\infty \delta_n}{h^d} + \frac{\|\tau\|_\infty \|\ell_\gamma\|_\infty L_K \delta_n}{h^{2d}} + \frac{\|\tau\|_\infty \|K\|_\infty^2 L_K \delta_n}{h^{2d+1}} .$$

Thus, we conclude $(12.6) \lesssim \frac{\delta_n^\rho}{h^{(2d+1)\rho}}$.

Together these estimates give the following bound on the discretization error in (12.4)

$$\mathbb{E}_\gamma\left[2^{\rho-1}\left(\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta \\ \|x-y\|,\|\vartheta-\theta\|\leq\delta_n}} |T_n^1(\theta,x;h) - T_n^1(\vartheta,y;h)|\right)^\rho\right] \lesssim \frac{\delta_n^\rho}{h^{(2d+1)\rho}} .$$

Now consider the second term in (12.4). 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, p. 15), which we state as Lemma 12.1 below. In order to derive the terms A, B, C described in (12.9), 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.

$$\mathbb{E}_\gamma\left[\left|K_h(X_1-x)\right|\right] \leq \|\ell_\gamma\|_\infty , \quad \mathbb{E}_\gamma\left[K_h^2(X_1-x)\right] \leq \frac{1}{h^d}\|\ell_\gamma\|_\infty \int K^2 \lesssim \frac{1}{h^d} .$$

This yields

$$A = \|U_n\|_\infty \lesssim \|u_n\|_\infty \leq \|\tau\|_\infty \frac{\|K\|_\infty^2}{h^{2d}} \lesssim \frac{1}{h^{2d}} ,$$
$$B^2 = n\left\|\mathbb{E}_\gamma\left[U_n^2(Z_1,\cdot,\theta,x;h)\right]\right\|_\infty \lesssim n\left\|\mathbb{E}_\gamma\left[u_n^2(Z_1,\cdot,\theta,x;h)\right]\right\|_\infty \lesssim \frac{n}{h^{3d}} .$$

The same arguments apply to C^2 defined in Lemma 12.1, giving

$$\begin{aligned}
C^2 =& = n(n-1)\Big(\mathbb{E}_\gamma[u_n^2(Z_1,Z_2,\theta,x;h)] + 4\,\mathbb{E}_\gamma[u_n(Z_1,Z_2,\theta,x;h)u_n^*(Z_1,\theta,x;h)] \\
&\quad + 4\left(\mathbb{E}_\gamma[u_n(Z_1,Z_2,\theta,x;h)]\right)^2 + 4\,\mathbb{E}_\gamma[{u_n^*}^2(Z_1,\theta,x;h)] \\
&\quad + 4\,\mathbb{E}_\gamma[u_n^*(Z_1,\theta,x;h)u_n^*(Z_2,\theta,x;h)]\Big) \\
&\lesssim n(n-1)\mathbb{E}_\gamma\big[\big|u_n(Z_1,Z_2,\theta,x;h)\big|\big]\,\|u_n\|_\infty \lesssim \frac{n^2}{h^{2d}}\,.
\end{aligned}$$

Now, Lemma 12.1 and the monotonicity of the exponential function give for any θ, x, h, γ and any $\omega > 0$ that

$$\begin{aligned}
&\mathbb{P}_\gamma\big(|T_n^1(\theta,x;h)| > \omega\big) \\
&= \mathbb{P}_\gamma\Big(\Big|\sum_{1\le j\ne k\le n} U_n(Z_j,Z_k,\theta,x;h)\Big| > n(n-1)\omega\Big) \\
&\lesssim T\exp\Big(-\frac{1}{T}\min\Big\{\frac{n(n-1)h^d\omega}{n}, \Big(\frac{n(n-1)h^{\frac{3d}{2}}\omega}{\sqrt{n}}\Big)^{\frac{2}{3}}, \big(n(n-1)h^{2d}\omega\big)^{\frac12}\Big\}\Big) \\
&\le T\exp\Big(-\frac{1}{T}nh^d\omega\Big)\mathbb{1}_{\omega\in[0,1)} + T\exp\Big(-\frac{1}{T}nh^d\omega^{\frac12}\Big)\mathbb{1}_{\omega\in[1,\infty)}
\end{aligned}$$

for $T > 0$ some universal constant. We apply the estimate

$$\mathbb{E}[|W|^\rho] \le a^\rho + \int_a^\infty \rho\omega^{\rho-1}\mathbb{P}(|W| > \omega)\,\mathrm{d}\omega\,, \quad a > 0$$

to $W = \sup_{x\in J_n,\theta\in\Theta_n}|T_n^1(\theta,x;h)|\big(\frac{nh^d}{\log n}\big)^{\frac12}$ and obtain

$$\begin{aligned}
&\mathbb{E}_\gamma\Big[\sup_{x\in J_n,\theta\in\Theta_n}|T_n^1(\theta,x;h)|^\rho\Big] \\
&\le \Big(\frac{\log n}{nh^d}\Big)^{\frac\rho2}\Big[a^\rho + \int_a^\infty \rho\omega^{\rho-1}\sum_{x\in J_n,\theta\in\Theta_n}\mathbb{P}_\gamma\Big(|T_n^1(\theta,x;h)| > \Big(\frac{\log n}{nh^d}\Big)^{\frac12}\omega\Big)\,\mathrm{d}\omega\Big] \\
&\le \Big(\frac{\log n}{nh^d}\Big)^{\frac\rho2}\Big[a^\rho + C_{\Theta,J}\delta_n^{-d-m}\int_a^\infty \rho\omega^{\rho-1}T\exp\Big(-\frac{1}{T}nh^d\Big(\frac{\log n}{nh^d}\Big)^{\frac12}\omega\Big)\mathbb{1}_{\omega\in\left[0,\left(\frac{nh^d}{\log n}\right)^{1/2}\right)}\,\mathrm{d}\omega \\
&\quad + C_{\Theta,J}\delta_n^{-d-m}\int_a^\infty \rho\omega^{\rho-1}T\exp\Big(-\frac{1}{T}nh^d\Big(\frac{\log n}{nh^d}\Big)^{\frac14}\omega^{\frac12}\Big)\mathbb{1}_{\omega\in\left[\left(\frac{nh^d}{\log n}\right)^{1/2},\infty\right)}\,\mathrm{d}\omega\Big] \\
&\lesssim \Big(\frac{\log n}{nh^d}\Big)^{\frac\rho2}\Big[a^\rho + C_{\Theta,J}\delta_n^{-d-m}\int_a^{\max\left\{\left(\frac{nh^d}{\log n}\right)^{1/2},a\right\}}\rho\omega^{\rho-1}T\exp\Big(-\frac{1}{T}\log(n)\omega\Big)\,\mathrm{d}\omega \\
&\quad + C_{\Theta,J}\delta_n^{-d-m}\int_{\max\left\{\left(\frac{nh^d}{\log n}\right)^{1/2},a\right\}}^\infty \rho\omega^{\rho-1}T\exp\Big(-\frac{1}{T}\log(n)\omega^{\frac12}\Big)\,\mathrm{d}\omega\Big]\,, \quad (12.8)
\end{aligned}$$

where we used

$$nh^d\Big(\frac{\log n}{nh^d}\Big)^{\frac12} = \Big(\frac{nh^d}{\log n}\Big)^{\frac12}\log n\,, \quad nh^d\Big(\frac{\log n}{nh^d}\Big)^{\frac14} = \Big(\frac{nh^d}{\log n}\Big)^{\frac34}\log n$$

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 (12.8) 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

$$\begin{aligned}
&\int_a^{\max\left\{\left(\frac{nh^d}{\log n}\right)^{1/2}, a\right\}} \rho\omega^{\rho-1} T \exp\big(-T^{-1}\log(n)\omega\big)\,\mathrm{d}\omega \\
&\leq \int_a^\infty \rho\omega^l T \exp\big(-T^{-1}\log(n)\omega\big)\,\mathrm{d}\omega \\
&= \rho \frac{T^{l+2} l!}{\log(n)^{l+1}} \sum_{k=0}^{l} \left(\frac{\big(T^{-1}\log(n)a\big)^k}{k!}\right) \cdot \exp(-T^{-1}\log(n)a) \\
&\lesssim \frac{n^{-T^{-1}a}}{\log n}\ .
\end{aligned}$$

By using the transformation $\omega \mapsto \frac{\omega^2 T^2}{\log(n)^2}$, we get

$$\begin{aligned}
&\int_{\max\left\{\left(\frac{nh^d}{\log n}\right)^{\frac{1}{2}}, a\right\}}^\infty \rho\omega^{\rho-1} T \exp\big(-T^{-1}\log(n)\omega^{\frac{1}{2}}\big)\,\mathrm{d}\omega \\
&\leq \int_{\max\left\{\left(\frac{nh^d}{\log n}\right)^{\frac{1}{2}}, a\right\}}^\infty \rho\omega^l T \exp\big(-T^{-1}\log(n)\omega^{\frac{1}{2}}\big)\,\mathrm{d}\omega \\
&= \int_{\max\left\{\left(\frac{nh^d}{\log n}\right)^{\frac{1}{4}}, a^{\frac{1}{2}}\right\} T^{-1}\log n}^\infty \frac{2\rho T^{2l+3}}{(\log n)^{2l+2}} \omega^{2l+1} \exp\big(-\omega\big)\,\mathrm{d}\omega \\
&= \frac{2\rho T^{2l+3}}{(\log n)^{2l+2}} (2l+1)! \sum_{k=0}^{2l+1} \frac{\left(\max\left\{\left(\frac{nh^d}{\log n}\right)^{\frac{1}{4}}, a^{\frac{1}{2}}\right\} T^{-1}\log n\right)^k}{k!} \\
&\quad \cdot \exp\left(-\max\left\{\left(\frac{nh^d}{\log n}\right)^{\frac{1}{4}}, a^{\frac{1}{2}}\right\} T^{-1}\log n\right) \\
&\lesssim \frac{1}{\log n} b_n^{2l+1} n^{-T^{-1}b_n} \\
&\lesssim n^{-c}\ ,
\end{aligned}$$

where $b_n = \max\left\{\left(\frac{nh^d}{\log n}\right)^{\frac{1}{4}}, a^{\frac{1}{2}}\right\}$ 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,

$$\frac{\delta_n^\rho}{h_n(\alpha)^{(2d+1)\rho}} = n^{-\frac{T^{-1}a\rho}{d+m}} h_n(\alpha)^{-(2d+1)\rho} \lesssim \left(\frac{\log n}{n h_n(\alpha)}\right)^{\frac{\rho}{2}},$$

$c > T^{-1}a$ and using that $\sup_\alpha \frac{\log n}{n h_n(\alpha)^d} \to 0$ implies $\sup_\alpha \frac{1}{h_n(\alpha)^d} \lesssim \frac{n}{\log n}$, we get

$$\begin{aligned}
\mathbb{E}_\gamma\Big[\sup_{x\in J,\theta\in\Theta} |T_n^1(\theta,x;h)|^\rho\Big] &\lesssim \frac{\delta_n^\rho}{h^{(2d+1)\rho}} + \left(\frac{\log n}{nh^d}\right)^{\frac{\rho}{2}} + \left(\frac{\log n}{nh^d}\right)^{\frac{1}{2}} \delta_n^{-d-m}\left(\frac{n^{-T^{-1}a}}{\log n} + n^{-c}\right) \\
&\lesssim \left(\frac{\log n}{nh^d}\right)^{\frac{\rho}{2}} + \left(\frac{\log n}{nh^d}\right)^{\frac{\rho}{2}} \cdot \frac{1}{\log n} \\
&\lesssim \left(\frac{\log n}{nh^d}\right)^{\frac{\rho}{2}},
\end{aligned}$$

concluding the considerations of $\sup_{x\in J,\theta\in\Theta} |T_n^1(\theta,x;h)|^\rho$.

□

Lemma 12.1 (Giné *et al.* (2000, p. 15))**.** *Let $(Z_n)_n$ 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 $x, y \in \mathbb{R}^d$*

$$U_n = \sum_{1\leq j\neq k\leq n} \chi(Z_j, Z_k)\,, \qquad \chi(x,y) = \chi(y,x)\,, \qquad \mathbb{E}[\chi(Z_1,x)] = \int_{\mathbb{R}^d} \chi(z,x)\,\mathrm{d}\mathbb{P}_{Z_1}(z) = 0\,.$$

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

$$A := \|\chi\|_\infty, \qquad B^2 := n\big\|\mathbb{E}\big[\chi^2(Z_1,\cdot)\big]\big\|_\infty, \qquad C^2 := n(n-1)\mathbb{E}\big[\chi^2(Z_1,Z_2)\big]\,. \tag{12.9}$$

Proof of Lemma 8.4. For brevity we introduce 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}. Using the decomposition in (12.1) yields

$$\begin{aligned}
\tilde{p}_{lj} &= \mathbb{P}_\gamma\Big(\sup_{x\in J,\theta\in\Theta} \|M_n(\theta,x;h_j) - \mathbb{E}_\gamma[M_n(\theta,x;h_j)]\| > \psi(C_{\mathrm{Lep}})r_l\Big) \\
&\leq \mathbb{P}_\gamma\Big(\sup_{x\in J,\theta\in\Theta} \|T_n^1(\theta,x;h_j)\| > \psi(C_{\mathrm{Lep}})r_l/2\Big) \\
&\quad + \mathbb{P}_\gamma\Big(\sup_{x\in J,\theta\in\Theta} \|T_n^2(\theta,x;h_j)\| > \psi(C_{\mathrm{Lep}})r_l/2\Big) \\
&=: \tilde{p}_{lj}^1 + \tilde{p}_{lj}^2\,.
\end{aligned}$$

We shall focus on the probabilities $\tilde{p}^1_{lj}$, the $\tilde{p}^2_{lj}$ are dealt with similarly and in fact simpler.
We use a discretization as in (12.3) and estimate

$$\begin{aligned}
&\mathbb{P}_\gamma\Big(\sup_{x\in J,\theta\in\Theta}\|T^1_n(\theta,x;h_j)\|>\psi(C_{\mathrm{Lep}})h_l^{\alpha_l}/2\Big)\\
\le&\,\mathbb{P}_\gamma\Big(\sup_{x\in J_n,\theta\in\Theta_n}\|T^1_n(\theta,x;h_j)\|>\psi(C_{\mathrm{Lep}})h_l^{\alpha_l}/4\Big) \qquad (12.10)\\
&+\mathbb{P}_\gamma\Bigg(\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta\\ \|x-y\|,\|\vartheta-\theta\|\le\delta_n}}\|T^1_n(\theta,x;h_j)-T^1_n(\vartheta,y;h_j)\|>\psi(C_{\mathrm{Lep}})h_l^{\alpha_l}/4\Bigg) \qquad (12.11)
\end{aligned}$$

The term (12.10) can by handled by Bernstein's inequality for U-statistics, cf. Lemma 12.1, just like we did in the proof of Theorem 8.3 and the fact that $h_j\ge h_l$, whence

$$\begin{aligned}
&\mathbb{P}_\gamma\Big(\sup_{x\in J_n,\theta\in\Theta_n}\|T^1_n(\theta,x;h_j)\|>\psi(C_{\mathrm{Lep}})h_l^{\alpha_l}/4\Big)\\
\le&\sum_{x\in J_n,\theta\in\Theta_n}\mathbb{P}_\gamma\big(\|T^1_n(\theta,x;h_j)\|>\psi(C_{\mathrm{Lep}})h_l^{\alpha_l}/4\big)\\
\le&\sum_{x\in J_n,\theta\in\Theta_n}\mathbb{P}_\gamma\big(\|U_n(Z_1,Z_2,\theta,x;h_j)\|>n(n-1)\psi(C_{\mathrm{Lep}})h_l^{\alpha_l}/4\big)\\
\le&C_{\Theta,J}\delta_n^{-d-m}\Big\{T\exp\Big(-T^{-1}nh_j^d\psi(C_{\mathrm{Lep}})h_l^{\alpha_l}/4\Big)\mathbb{1}_{\psi(C_{\mathrm{Lep}})h_l^{\alpha_l}/4\in[0,1)}\\
&+T\exp\Big(-T^{-1}nh_j^d\sqrt{\psi(C_{\mathrm{Lep}})h_l^{\alpha_l}/4}\Big)\mathbb{1}_{\psi(C_{\mathrm{Lep}})h_l^{\alpha_l}/4\in[1,\infty)}\Big\}\\
\le&C_{\Theta,J}\delta_n^{-d-m}\Big\{T\exp\Big(-T^{-1}n\psi(C_{\mathrm{Lep}})h_l^{\alpha_l+d}/4\Big)\mathbb{1}_{\psi(C_{\mathrm{Lep}})h_l^{\alpha_l}/4\in[0,1)}\\
&+T\exp\Big(-T^{-1}n\sqrt{\psi(C_{\mathrm{Lep}})}h_l^{\alpha_l/2+d}/2\Big)\mathbb{1}_{\psi(C_{\mathrm{Lep}})h_l^{\alpha_l}/4\in[1,\infty)}\Big\}. \qquad (12.12)
\end{aligned}$$

By using $h_l=\big(\frac{\log n}{n}\big)^{\frac{1}{2\alpha_l+d}}$, we get that

$$nh_l^{\alpha_l/2+d}\ge nh_l^{\alpha_l+d}=n\Big(\frac{\log n}{n}\Big)^{\frac{\alpha_l+d}{2\alpha_l+d}}\ge\frac{n\log n}{n}=\log n$$

yielding

$$\begin{aligned}
(12.12)\le&C_{\Theta,J}T\delta_n^{-d-m}\Big(n^{-T^{-1}\psi(C_{\mathrm{Lep}})/4}+n^{-T^{-1}\sqrt{\psi(C_{\mathrm{Lep}})}/2}\Big)\\
\le&2C_{\Theta,J}T\delta_n^{-d-m}n^{-T^{-1}\sqrt{\psi(C_{\mathrm{Lep}})}/2}, \qquad (12.13)
\end{aligned}$$

because $\psi(C_{\mathrm{Lep}})\ge 4$. The term (12.11) is handled by arguments similar to the ones found in the proof of Theorem 8.3. Using Markov's inequality, $h_j\ge h_l$

and the arguments used to treat (12.5) - (12.7) 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

$$\begin{aligned}
&\limsup_{n\to\infty} \sup_{*} \delta_n^{-1} h_l^{\alpha_l+2d+1} \mathbb{P}_\gamma \left(\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta \\ \|x-y\|,\|\vartheta-\theta\|\le\delta_n}} \left\| T_n^1(\theta,x;h_j) - T_n^1(\vartheta,y;h_j) \right\| > \psi(C_{\mathrm{Lep}}) h_l^{\alpha_l}/4 \right) \\
&\le \limsup_{n\to\infty} \sup_{*} \frac{4\delta_n^{-1} h_j^{2d+1}}{\psi(C_{\mathrm{Lep}})} \mathbb{E}_\gamma \left[\sup_{\substack{x,y\in J,\vartheta,\theta\in\Theta \\ \|x-y\|,\|\vartheta-\theta\|\le\delta_n}} \left\| T_n^1(\theta,x;h_j) - T_n^1(\vartheta,y;h_j) \right\| \right] < \widetilde{C} < \infty ,
\end{aligned} \tag{12.14}$$

where the suprema are taken over $\alpha \in [a,b]$, $\gamma \in \Gamma(\alpha)$, $0 \le l \le j \le k_n(\alpha)$.

Now set

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

where T is the universal constant in Lemma 12.1.
Combining (12.13) and (12.14) yields

$$\begin{aligned}
&\limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \sup_{0\le l\le j\le k_n(\alpha)} n^{u(C_{\mathrm{Lep}})} \tilde{p}_{lj}^1 \\
&\le \limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} 2 C_{\Theta,J} T\, n^{u(C_{\mathrm{Lep}})} \delta_n^{-d-m} n^{-T^{-1}\sqrt{\psi(C_{\mathrm{Lep}})}/2} \qquad (12.15) \\
&\quad + \limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \sup_{0\le l\le j\le k_n(\alpha)} \widetilde{C}\, n^{u(C_{\mathrm{Lep}})} \delta_n h_l^{-\alpha_l-2d-1} , \qquad (12.16)
\end{aligned}$$

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 (12.15), we see that

$$\begin{aligned}
&\log_n \left(n^{u(C_{\mathrm{Lep}})} \delta_n^{-d-m} n^{-T^{-1}\sqrt{\psi(C_{\mathrm{Lep}})}/2} \right) \\
&\le \frac{T^{-1}\sqrt{\psi(C_{\mathrm{Lep}})}}{8(d+m)} + \frac{T^{-1}\sqrt{\psi(C_{\mathrm{Lep}})}}{4} - \frac{T^{-1}\sqrt{\psi(C_{\mathrm{Lep}})}}{2} \\
&= -\frac{\big(2(d+m)-1\big) T^{-1}\sqrt{\psi(C_{\mathrm{Lep}})}}{8(d+m)} \\
&\le -\frac{\big(2(d+m)-1\big) T^{-1}\sqrt{\tilde{c}_1 C_- - \tilde{c}_2}}{8(d+m)}
\end{aligned}$$

$$\leq -1$$

because

$$C_- \geq \tilde{c}_1^{-1}\left[\tilde{c}_2 + T^2\left(\frac{8(d+m)}{2(d+m)-1}\right)^2\right].$$

Hence, for all $C_{\text{Lep}} \geq C_-$, we have

$$\begin{aligned}(12.15) &= \limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} 2C_{\Theta,J} T\, n^{u(C_{\text{Lep}})}\delta_n^{-d-m} n^{-T^{-1}\sqrt{\psi(C_{\text{Lep}})}/2} \\ &\leq \limsup_{n\to\infty} n^{-1} = 0.\end{aligned}$$

Ad (12.16). 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

$$\begin{aligned}&\log_n\left(n^{u(C_{\text{Lep}})}\delta_n h_l^{-\alpha_l-2d-1}\right) \\ &\leq \frac{T^{-1}\sqrt{\psi(C_{\text{Lep}})}}{8(d+m)} - \frac{T^{-1}\sqrt{\psi(C_{\text{Lep}})}}{4(d+m)} + \frac{b+2d+1}{2a+d} \\ &= -\frac{T^{-1}\sqrt{\psi(C_{\text{Lep}})}}{8(d+m)} + \frac{b+2d+1}{2a+d} \\ &\leq -\frac{T^{-1}\sqrt{\tilde{c}_1 C_- - \tilde{c}_2}}{8(d+m)} + \frac{b+2d+1}{2a+d} \leq -1\end{aligned}$$

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 $C_{\text{Lep}} \geq C_-$, we have

$$\begin{aligned}(12.16) &= \limsup_{n\to\infty} \sup_{\alpha\in[a,b]} \sup_{\gamma\in\Gamma(\alpha)} \sup_{0\leq l\leq j\leq k_n(\alpha)} \widetilde{C}\, n^{u(C_{\text{Lep}})}\delta_n h_l^{-\alpha_l-2d-1} \\ &\leq \limsup_{n\to\infty} n^{-1} = 0\,,\end{aligned}$$

concluding the proof.

□