Estimating nonparametric functionals efficiently under one-sided errors

Markus Reiß
Institute of Mathematics
Humboldt-Universität zu Berlin
mreiss@mathematik.hu-berlin.de

and

Leonie Selk
Institute of Mathematics
Universität Hamburg
leonie.selk@math.uni-hamburg.de

July 17, 2014

Abstract

For nonparametric regression with one-sided errors and a related continuous-time model for Poisson point processes we consider the problem of efficient estimation for linear functionals of the regression function. The optimal rate is obtained by an unbiased estimation method which nevertheless depends on a Hölder condition or monotonicity assumption for the underlying regression function.

We first construct a simple blockwise estimator and then build up a nonparametric maximum-likelihood approach for exponential noise variables and the point process model. In that approach also non-asymptotic efficiency is obtained (UMVU: uniformly minimum variance among all unbiased estimators). In addition, under monotonicity the estimator is automatically rate-optimal and adaptive over Hölder classes. The proofs rely essentially on martingale stopping arguments for counting processes and the point process geometry. The estimators are easily computable and a small simulation study confirms their applicability.

Key words and Phrases: frontier estimation, support estimation, Poisson point process, sufficiency, completeness, UMVU, nonparametric MLE, shape constraint, monotone boundary, optional stopping.

AMS subject classification: 62G08, 62G15, 62G32, 62M30, 60G55

*Financial support by the DFG through Research Unit FOR 1735 Structural Inference in Statistics: Adaptation and Efficiency is gratefully acknowledged.
1 Introduction

For regression models

\[ y_i = g(i/n) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1.1} \]

the estimation of linear functionals of the regression function \( g \) is well understood if \((\varepsilon_i)\) are uncorrelated with mean zero and variance \( \sigma^2 > 0 \). Then the discrete functionals

\[ \vartheta^{(n)} = \frac{1}{n} \sum_{i=1}^{n} g(i/n)w(i/n) \text{ for some function } w : [0, 1] \to \mathbb{R} \tag{1.2} \]

can be estimated by the plug-in version \( \hat{\vartheta}_n = \frac{1}{n} \sum_{i=1}^{n} y_i w(i/n) \) without bias and with variance \( \frac{\sigma^2}{n^2} \sum_{i=1}^{n} w(i/n)^2 \). By the Gauß-Markov theorem \( \hat{\vartheta}_n \) has minimal variance among all linear and unbiased estimators. In the Gaussian case \( \hat{\vartheta}_n \) is even UMVU (uniformly minimum variance among all unbiased estimators). In the corresponding continuous-time signal-in-white-noise model \( dY(t) = g(t)dt + \sigma n^{-1/2}dW_t, \quad t \in [0, 1] \), with a Brownian motion \( W \) and some \( g \in L^2([0, 1]) \) the plug-in estimator \( \hat{\vartheta} = \int_0^1 w(t)dY(t) \) is equally an unbiased estimator of

\[ \vartheta = \int_0^1 g(t)w(t) \, dt \text{ for some } w \in L^2([0, 1]) \tag{1.3} \]

of variance \( \frac{\sigma^2}{n} \int_0^1 w(t)^2 \, dt \). By the Riesz representation theorem, we can thus estimate any linear \( L^2 \)-continuous functional of \( g \) with parametric rate \( n^{-1/2} \).

In certain applications the function \( g \) is, however, determined as the boundary or frontier function of the observations, which can be modeled equivalently by one-sided errors \((\varepsilon_i)\). The prototypical case is that \((\varepsilon_i)\) are i.i.d. with \( \varepsilon_i \geq 0 \) and for some \( \lambda > 0 \)

\[ P(\varepsilon_i \leq x) = \lambda x + O(x^2) \text{ as } x \downarrow 0, \tag{1.4} \]

e.g. \( \varepsilon_i \sim \text{Exp}(\lambda) \). In that case the parametric rate for the location model (i.e. assuming \( g \) to be constant) is with \( n^{-1} \) much faster than in the regular case. These irregular statistical models have also found considerable theoretical interest, e.g. in the recent work by Baraud and Birgé (2014). A rate-optimal estimator is given by the extreme value statistics \( \min_i y_i \). For the nonparametric problem of estimating the entire function \( g \) in \( L^2 \)-loss, the optimal rate is \( n^{-\beta/(\beta+1)} \) for \( g \) in a Hölder ball of regularity \( \beta \in (0, 1] \) and radius \( R > 0 \):

\[ g \in C^\beta(R) = \left\{ f : [0, 1] \to \mathbb{R} \mid \forall x, y \in [0, 1] : |f(y) - f(x)| \leq R|y - x|^\beta \right\}. \tag{1.5} \]

This is achieved by a local polynomial estimator \( \hat{g}_{n,h} \) as in the regular case, see e.g. Jirak, Meister, and Reiß (2014) for a construction and a survey of the large literature on that topic. A plug-in estimator \( \hat{\vartheta}_n := \int_0^1 \hat{g}_{n,h_n}(x)w(x) \, dx \) with optimal bandwidth \( h_n \) to estimate \( \vartheta \) in \((1.3)\) can achieve at best the rate \( n^{-\beta/(\beta+1/2)} \). This rate, however, is not optimal (for \( \beta < 1/2 \) it is even slower than \( n^{-1/2} \) in the regular case).
Here we show that the optimal estimation rate for $\vartheta(n)$ under one-sided errors is $n^{-(\beta+1/2)/(\beta+1)}$ for $g \in C^\beta(R)$. The improvement over the plug-in estimator is achieved by an unbiased estimation procedure. The bias is exactly zero for the case of exponentially distributed errors and it is asymptotically negligible under (1.4) for $\beta > 1/2$. Compared with standard nonparametric results it is remarkable that an unbiased estimator can be constructed whose rate is nevertheless worse than the parametric rate ($n^{-1}$ in this case). The risk bound comes from a trade-off between two terms in the variance instead of the usual bias-variance trade-off.

As for mean regression with the signal-in-white-noise model, also for one-sided errors an analogous continuous-time model is most useful in exhibiting the main statistical structure. It is given by observing a Poisson point process (PPP) on $[0,1] \times \mathbb{R}$ of intensity

$$\lambda_g(x,y) = n \mathbf{1}(y \geq g(x)), \quad x \in [0,1], \ y \in \mathbb{R}, \quad (1.6)$$

see e.g. Karr (1991) for point process properties and Figure 1 below for an illustration. For sufficiently regular $g$ this model can be shown to be asymptotically equivalent to the regression-type model (1.1) with $\lambda = 1$ in (1.4), cf. Meister and Reiß (2013). Here it serves as a fundamental model for developing the methods, which will then be transferred explicitly to the discrete model (1.1). The functional of interest is $\vartheta$ from (1.3). We then show that $\vartheta$ can not only be estimated without bias, but we can even construct an estimator of $\vartheta$ which is UMVU. This non-asymptotic efficiency result is based on a nonparametric maximum-likelihood approach, where the maximum-likelihood estimator (MLE) $\hat{g}^{\text{MLE}}$ is not only explicit, but also forms a sufficient and complete statistics. In parallel with Gaussian mean regression we thus have the UMVU-property of the estimator, but its construction depends on the smoothness parameter $\beta$ and its asymptotic rate is worse than for parametric location estimation. Still, we are able to prove its asymptotic normality and to provide a self-normalising CLT such that asymptotic inference is feasible.

We extend the theory to the class of monotone functions $g$. There the nonparametric MLE approach even allows to construct an estimator which is UMVU among all estimators of monotone functions and simultaneously rate-optimal for any Hölder regularity $\beta$. The additional shape constraint thus yields an adaptive estimation method.

The regression-type model (1.1) with one-sided errors and the PPP model (1.6) exhibit a similar structure as density support estimation or image boundary recovery problems. Let us review briefly the literature on functional estimation for these statistical models. Many asymptotic results for the expected area of the convex hull for i.i.d. observations are based on the classical results by Rényi and Sulanke (1964). In a recent article by Groeneboom (2012) the asymptotic joint distribution of the points on the convex hull and a remaining area is determined. For image recovery problems Korostelev and Tsybakov (1993) describe already the rate $n^{-(\beta+1/2)/(\beta+1)}$ obtained for the functional $\int_0^1 g(x)dx$. The upper bound is based on a localisation step and loses a logarithmic factor. By threefold sample splitting Gayraud (1997) has constructed an
estimator achieving this rate exactly for the related density support area estimation. All these estimators are analysed asymptotically. They lack the non-asymptotic unbiasedness and UMVU property we have found here. Many other estimators are concerned with the estimation of the density support set or the regression-type function itself, not of the area or other functionals, let us mention the work by Mammen and Tsybakov (1995) for connections to classification problems. Specifically, a nonparametric MLE approach under monotonicity has been developed by Korostelev, Simar, and Tsybakov (1995) for the asymptotically exact risk in estimating the density support set in Hausdorff distance. In Gaussian mean regression a nonparametric MLE over regular function classes is equivalent to a least-squares approach with roughness penalty, leading e.g. to smoothing splines. Under shape constraints the MLE is a well studied object, see e.g. Groeneboom and Wellner (1992), but usually results are derived asymptotically.

In the next section we shall develop a simple block-wise estimator and derive explicit bounds on its bias and variance. Based on optional stopping for an intrinsic martingale we prove that it is unbiased under the PPP model and under exponential noise in the regression-type model. For more general regression noise the required compensation cannot be achieved exactly, but it comes close to the model with corresponding exponential noise. The last part of that section is devoted to the lower bound showing that the rate is indeed optimal. The nonparametric MLE approach is presented in Section 3, first for the class of Hölder functions, then for monotone functions. The derivation of the completeness of the nonparametric MLE and the stopping arguments for the intrinsic martingale are intriguing. For the MLE under Hölder conditions we obtain central limit theorems where for Lipschitz functions the asymptotic variance is completely explicit.

In Section 4 we discuss major implications of the results, in particular concerning adaptive estimation and estimating coefficients in a series or projection estimator approach. Extensions and limitations are mentioned and a small simulation study shows that all estimators are numerically feasible and have good finite-sample properties. Most proofs are instructive and reveal some beautiful interplay between statistics, probability and geometry such that in the Appendix we only provide some technical lemmata (some of independent interest) and the more involved proof of the CLT. The notation follows the usual conventions. We write $a_n \lesssim b_n$ or $a_n = O(b_n)$ to say that $a_n$ is bounded by a constant multiple of $b_n$ and $a_n \sim b_n$ for $a_n \lesssim b_n$ as well as $b_n \lesssim a_n$. Moreover, $a_n = o(b_n)$ means $a_n/b_n \to 0$ and $a_n \asymp b_n$ means $a_n/b_n \to 1$.

## 2 Simple rate-optimal estimation

### 2.1 Block-wise estimation in the PPP model

Let $(X_j, Y_j)_{j \geq 1}$ denote the observations of the Poisson point process (PPP) with intensity $\lambda_0$. We shall see here that we can estimate $\vartheta$ from (1.3) without any bias. To
grasp the main idea, suppose that \( w(x) = 1 \) holds and that we know a deterministic function \( \tilde{g} : [0, 1] \to \mathbb{R} \) with the property \( \tilde{g} \geq g \) (pointwise). Then the number of PPP observations below the graph of \( \tilde{g} \) is Poisson-distributed with intensity equal to \( n \) times the area between \( g \) and \( \tilde{g} \):

\[
\sum_{j \geq 1} 1(Y_j \leq \tilde{g}(X_j)) \sim \text{Pois}(n \int_0^1 (\tilde{g} - g)(x) \, dx).
\]

Hence, we obtain

\[
\tilde{\vartheta} := \int_0^1 \tilde{g}(x) \, dx - \frac{1}{n} \sum_{i \geq 1} 1(Y_i \leq \tilde{g}(X_i)) \Rightarrow \mathbb{E}[\tilde{\vartheta}] = \int_0^1 (\tilde{g} - (\tilde{g} - g))(x) \, dx = \vartheta.
\]

The variance of \( \tilde{\vartheta} \) is the larger, the larger the area between the graphs is. The main idea is now to find an empirical substitute for \( \tilde{g} \), which by stopping time arguments keeps the unbiasedness.

To this end, we partition \([0, 1]\) in subintervals \( I_k = [kh, (k + 1)h) \) of length \( h \) and note that the block-wise minimum \( Y_k^* := \min_{j : X_j \in I_k} Y_j \) satisfies \( Y_k^* \geq \min_{x \in I_k} g(x) \). By the Hölder property of \( g \) we conclude that \( g(x) \leq Y_k^* + Rh^\beta \) holds for all \( x \in I_k \) and thus \( Y_k^* + Rh^\beta \) is a local upper bound for \( g \), see also Figure 1. We thus estimate the functional locally on these blocks by

\[
\hat{\vartheta}_k := (Y_k^* + Rh^\beta)\bar{w}_k - \frac{1}{nh} \sum_{i \geq 1} 1(X_i \in I_k, Y_i \leq Y_k^* + Rh^\beta) w(X_i)
\]

where the true local parameter is \( \vartheta_k := \frac{1}{n} \int_{I_k} g(x)w(x) \, dx \) and \( \bar{w}_k = \frac{1}{n} \int_{I_k} w(x) \, dx \).

2.1 Theorem. The estimator \( \hat{\vartheta}^{\text{block}} = \sum_{k=0}^{h^{-1} - 1} \hat{\vartheta}_k h \) satisfies with \( \|w\|_{L^2}^2 = \int_0^1 w(x)^2 \, dx \)

\[
\mathbb{E}[\hat{\vartheta}^{\text{block}}] = \vartheta, \quad \text{Var}(\hat{\vartheta}^{\text{block}}) \leq \frac{2R\|h\|^{-1} + (nh)^{-1}}{n}\|w\|_{L^2}^2.
\]
In particular, the asymptotically optimal block size \( h \propto (2\beta R n)^{-1/(\beta+1)} \) yields
\[
\limsup_{n \to \infty} \sup_{g \in \mathcal{C}_\beta(\mathbb{R})} n^{(2\beta+1)/(\beta+1)} \text{Var}(\hat{\vartheta}^{\text{block}}) \leq \frac{\beta + 1}{\beta} (2\beta R n)^{1/(\beta+1)} \| w \|_{L_2}^2.
\]

Proof. Let us study the weighted counting process
\[
N(t) := \sum_{i \geq 1} \mathbf{1}(X_i \in I_k, Y_i \leq t) w(X_i), \quad t \in \mathbb{R}.
\]
The pure counting process \( \sum_{i} \mathbf{1}(X_i \in I_k, Y_i \leq t) \) is a point process in \( t \) with deterministic intensity \( \lambda_t = n \int_{I_k} (t - g(x))_+ dx. \) Hence, \( (N(t), t \in \mathbb{R}) \) is a process with independent increments satisfying (e.g. via Prop. 2.32 in Karr (1991))
\[
\mathbb{E}[N(t)] = \int_{I_k} n(t - g(x))_+ w(x) dx, \quad \text{Var}(N(t)) = \int_{I_k} n(t - g(x))_+ w(x)^2 dx.
\]
In particular, \( M(t) = N(t) - \mathbb{E}[N(t)] \) is a càdlàg martingale with respect to the filtration
\[
\mathcal{F}_t = \sigma((X_i, Y_i) \mathbf{1}(Y_i \leq t), i \geq 1), \quad t \in \mathbb{R}, \tag{2.1}
\]
with mean zero and predictable quadratic variation \( \langle M \rangle_t = \text{Var}(N(t)). \)

Now note that \( \tau := Y_k^* + Rh^\beta \) is an \( (\mathcal{F}_t) \)-stopping time with
\[
P(\tau \geq t) = \exp \left( -n \int_{I_k} (t - Rh^\beta - g(x))_+ dx \right)
\leq \exp \left( -nh(t - \max_{x \in I_k} g(x) - Rh^\beta) \right) \tag{2.2}
\]
for \( t \geq \max_{x \in I_k} g(x) + Rh^\beta. \) In particular, \( \tau \) has finite expectation and Lemma 5.1 on optional stopping yields
\[
\mathbb{E}[M(\tau)] = 0 \Rightarrow \mathbb{E}[N(\tau)] = n \int_{I_k} \mathbb{E}[(\tau - g(x))_+] w(x) dx
\]
and
\[
\text{Var}(M(\tau)) = \mathbb{E}[\langle M \rangle_\tau] = n \int_{I_k} \mathbb{E}[(\tau - g(x))_+] w(x)^2 dx.
\]
Noting \( \tau \geq g(x) \) we have
\[
\mathbb{E}[(\tau - g(x))_+] = \mathbb{E}[Y_k^*] + Rh^\beta - g(x) \text{ for all } x \in I_k.
\]
The identity
\[
\hat{\vartheta}_k = \tau \bar{w}_k - \frac{1}{nh} N(\tau) = \vartheta_k - \frac{1}{nh} M(\tau)
\]
implies \( \mathbb{E}[\hat{\vartheta}_k] = \vartheta_k \) and
\[
\text{Var}(\hat{\vartheta}_k) = \frac{1}{n^2 h^2} \text{Var}(M(\tau)) = \frac{1}{n^2 h^2} \int_{I_k} \mathbb{E}[Y_k^* + Rh^\beta - g(x)] w(x)^2 dx.
\]
A rough universal bound as in \([2.2]\) yields with a random variable \(E \sim \text{Exp}(nh)\)
\[
E[Y^*_k] \leq E \left[ \max_{x \in I_k} g(x) + E \right] \leq g(x) + Rh^\beta + (nh)^{-1}.
\]
This implies
\[
\text{Var}(\tilde{\vartheta}_k) \leq \frac{2R h^\beta + (nh)^{-1}}{nh^2} \int_{I_k} w(x)^2 dx.
\]
We conclude for the final estimator \(\hat{\vartheta}_{\text{block}} = \sum_{k=0}^{h^{-1}-1} \tilde{\vartheta}_k h\) by the independence of \((\tilde{\vartheta}_k)_k\) that
\[
E[\hat{\vartheta}_{\text{block}}] = \vartheta, \quad \text{Var}(\hat{\vartheta}_{\text{block}}) \leq \frac{2R h^\beta + (nh)^{-1}}{n} \int_0^1 w(x)^2 dx.
\]
Finally, insertion of the asymptotically optimal \(h\) yields the variance bound. \(\square\)

### 2.2 Blockwise estimation in the regression-type model

Let us consider the equi-distant regression model \([1.1]\) where \((\varepsilon_i)\) are i.i.d. satisfying \([1.4]\). Moreover, for the survival function \(F_\varepsilon(y) = P(\varepsilon > y)\) we require \(e^{-c(1+y)^p} \lesssim F_\varepsilon(y) \lesssim (1+y)^{-\delta}\) for some \(c, p, \delta > 0\) and all \(y \geq 0\). The primary example will be \(\varepsilon_i \sim \text{Exp} (\lambda)\), but any distribution on \([0, \infty)\) with a Lipschitz continuous density \(f_\varepsilon\) at zero and \(f_\varepsilon(0) = \lambda\) is covered, as soon as the relatively loose tail bounds at infinity are satisfied.

Since the observation design is discrete, our parameter of interest now is \(\vartheta^{(n)}\) from \([1.2]\). In analogy with the PPP case we build an estimator for \(\tilde{\vartheta}_k = \frac{1}{nh} \sum_{i \in I_k} g(i/n)w(i/n)\) on each block of indices \(I_k := \{ i : kh < \frac{i}{n} \leq (k+1)h\}\), where \(h^{-1}, nh \in \mathbb{N}\):
\[
\tilde{\vartheta}_k := \frac{1}{nh} \sum_{i \in I_k} \left( y_i \wedge (y_k^* + Rh^\beta) - \lambda^{-1} I \left( y_i \leq y_k^* + Rh^\beta \right) \right) w(i/n).
\]
Here, \(y_k^* = \min_{i \in I_k} y_i\) is again the minimal observation on each block. In contrast to the PPP-estimator the empirical upper bound for \(g\) on \(I_k\) is given by the minimum of \(y_k^* + Rh^\beta\) and \(y_i\), which for the rate-optimal choice of \(h\), however, has negligible impact. We obtain the following result where \(\|w\|_p = (\frac{1}{n} \sum_{i=1}^n |w(i/n)|^p)^{1/p}\) denotes the standardised \(\ell^p\)-norm.

**2.2 Theorem.** The estimator \(\tilde{\vartheta}_{n,\text{block}} = \sum_{k=0}^{h^{-1}-1} \tilde{\vartheta}_k h\) satisfies for \(h \to 0\) with \(nh \to \infty\) uniformly in \(n, h, \lambda\)
\[
|E[\tilde{\vartheta}_{n,\text{block}} - \vartheta^{(n)}]| \lesssim \frac{(Rh^\beta + (n\lambda h)^{-1})^2}{\lambda} \|w\|_1, \quad \text{Var}(\tilde{\vartheta}_{n,\text{block}}) \lesssim \frac{Rh^\beta + (n\lambda h)^{-1}}{n\lambda(1 \wedge \lambda/c)} \|w\|^2_2.
\]
In particular, uniformly over \(\beta \geq \beta_0 > 1/2, \lambda \geq \lambda_0 > 0, R \leq R_0 < \infty\) we obtain with the rate-optimal block size \(h \sim (Rn\lambda)^{-1/(\beta+1)}\)
\[
(E[\tilde{\vartheta}_{n,\text{block}} - \vartheta^{(n)}])^2 = o(\text{Var}(\tilde{\vartheta}_{n,\text{block}})), \quad \text{Var}(\tilde{\vartheta}_{n,\text{block}}) \lesssim R^{1/(\beta+1)}(n\lambda)^{-(2\beta+1)/(\beta+1)} \|w\|^2_2.
\]
In the case \( \varepsilon_i \sim \text{Exp}(\lambda) \) we have for any \( \beta \in (0, 1], R, \lambda > 0 \) the more precise result

\[
\mathbb{E}[\tilde{\vartheta}_{\text{block}}] = \vartheta, \quad \text{Var}(\tilde{\vartheta}_{\text{block}}) \leq \frac{2Rh^\beta + (n\lambda h)^{-1}}{n\lambda} \|w\|_2^2.
\]

2.3 Remark. The result and proof for \( \varepsilon_i \sim \text{Exp}(\lambda) \) are exactly as in the PPP model. For other distributions of \( \varepsilon_i \) the estimator is only asymptotically unbiased, but for Hölder regularity \( \beta > 1/2 \) the bias is negligible with respect to the stochastic error. For strong asymptotic equivalence with the PPP model in Le Cam’s sense the necessary minimal regularity is \( \beta > 1 \), see Meister and Reiß (2013).

Proof. Fix a block with index \( k \) and consider for \( t \in \mathbb{R} \)

\[
M(t) := \sum_{i \in I_k} \left( 1(y_i \leq t) + \log \tilde{F}_\varepsilon(y_i \land t - g(i/n)) \right) w(i/n).
\]

With respect to the filtration \( \mathcal{F}_t = \sigma(y_i1(y_i \leq t), i \in I_k) \), \( (M(t), t \in \mathbb{R}) \) defines a martingale (just note that on \( \{y_i > t\} \) the compensator equals the integrated hazard function \( \int_0^{y_i-g(i/n)} \tilde{F}_\varepsilon(s)^{-1}d\tilde{F}_\varepsilon(s) \) with \( \mathbb{E}[M(t)] = 0 \) and quadratic variation \( \langle M \rangle_t = \sum_{i \in I_k} (-\log \tilde{F}_\varepsilon(y_i \land t - g(i/n)))w(i/n)^2 \). Moreover, \( \tau := \gamma_k^* + Rh^\beta \) is a stopping time with respect to \( (\mathcal{F}_t) \). From the representation

\[
\vartheta_k - \vartheta = \frac{1}{n\lambda h} \left( \sum_{i \in I_k} \left( \lambda(y_i \land \tau - g(i/n)) + \log \tilde{F}_\varepsilon(y_i \land \tau - g(i/n)) \right) w(i/n) - M(\tau) \right)
\]

and the stopping Lemma 5.1 in combination with \( \mathbb{E}[\tau] < \infty \) due to the moment bound from Lemma 5.2 we conclude

\[
\mathbb{E}[\tilde{\vartheta}_k - \vartheta] = \frac{1}{n\lambda h} \sum_{i \in I_k} \mathbb{E} \left[ (\lambda z + \log \tilde{F}_\varepsilon(z))|z=y_i,\tau=g(i/n) \right] w(i/n).
\]

In the case \( \varepsilon_i \sim \text{Exp}(\lambda) \) we have \( \tilde{F}_\varepsilon(z) = e^{-\lambda z} \) and the estimator is unbiased.

In general, \( \lambda z + \log \tilde{F}_\varepsilon(z) = O(z^2 + z^p) \) holds, using \( \log \tilde{F}_\varepsilon(z) = -\lambda z + O(z^2) \) for \( z \) in a neighbourhood of zero and the exponential growth bound on \( \tilde{F}_\varepsilon \) for larger \( z \), where we may assume \( p \geq 2 \). By Lemma 5.2 this leads to

\[
|\mathbb{E}[^{\tilde{\vartheta}_k} - \vartheta]| \leq \sum_{i \in I_k} \mathbb{E} \left[ \frac{(\tau - g(i/n))^2 + (\tau - g(i/n))^p}{n\lambda h} |w(i/n)| \right]
\]

\[
\leq \frac{(Rh^\beta + (n\lambda h)^{-1})^2}{n\lambda} \sum_{i \in I_k} |w(i/n)|.
\]

This implies the assertion for the bias of \( \tilde{\vartheta}_{\text{block}} \). For the variance bound we use \( \text{Var}(A + B) \leq 2 \text{Var}(A) + 2 \text{Var}(B) \) and the stopping Lemma 5.1 to obtain:

\[
\text{Var}(\tilde{\vartheta}_k) = \frac{1}{(n\lambda h)^2} \text{Var} \left( \sum_{i \in I_k} (\lambda z + \log \tilde{F}_\varepsilon(z))|z=y_i,\tau=g(i/n) w(i/n) - M(\tau) \right)
\]
≤ \frac{2}{(n\lambda h)^2} \sum_{i \in I_k} E \left[ \left( nh(\lambda z + \log \tilde{F}_\varepsilon(z))^2 - \log \tilde{F}_\varepsilon(z) \right) \right] w(i/n)^2 \\
\leq \frac{1}{(n\lambda h)^2} \sum_{i \in I_k} \left( nh \left( (\tau - g(i/n))^2 + (\tau - g(i/n))^p \right) \right. \\
+ E \left[ \lambda(\tau - g(i/n)) + c(\tau - g(i/n))^p \right) \right] w(i/n)^2 \\
\leq \frac{(\lambda \vee c)(R^\beta + (n\lambda h)^{-1})}{(n\lambda h)^2} \sum_{i \in I_k} w(i/n)^2.

For the global estimator we infer \( \text{Var}(\tilde{\varrho}^\text{block}_n) \leq \frac{R^\beta + (n\lambda h)^{-1}}{n\lambda(1+\lambda/c)} \|w\|_2^2 \) by independence of \((\tilde{\varrho}_i)\). It remains to insert the rate-optimal choice of \( h \) and to note that \( n^{-4\beta/(\beta+1)} = o(n^{-2(\beta+1)/(\beta+1)}) \) holds for \( \beta > 1/2 \).

Finally, in the case \( \varepsilon_i \sim \text{Exp}(\lambda) \) we have \( E[Y_k^* - \max_i g(i/n)] \leq (n\lambda h)^{-1} \) and \( \text{Var}(\tilde{\varrho}_k) = \frac{E[(M)^2]}{(n\lambda h)^2} \). Consequently,

\[
\text{Var}(\tilde{\varrho}_k) \leq \sum_{i \in I_k} \frac{E[Y_k^* + R^\beta - g(i/n)]}{\lambda(nh)^2} w(i/n)^2 \leq \frac{2R^\beta + (n\lambda h)^{-1}}{\lambda(nh)^2} \sum_{i \in I_k} w(i/n)^2,
\]

which by independence gives the asserted bound for \( \text{Var}(\tilde{\varrho}^\text{block}_n) \).

\[ \square \]

2.3 Rate optimality

We prove that the rate \( R^{1/(2\beta+2)}n^{-(\beta+1/2)/(\beta+1)} \) is optimal in a minimax sense over \( C^\beta(R) \). The proof is conducted for the PPP model, the regression case with \( \varepsilon_i \sim \text{Exp}(\lambda) \) can be treated analogously.

2.4 Theorem. For estimating \( \varrho = \int_0^1 g(x)w(x) dx, \ w \in L^2([0,1]) \), in the PPP model with parameter class \( C^\beta(R) \), \( \beta \in (0,1] \), \( R > 0 \), the following asymptotic lower bound holds:

\[
\liminf_{n \to \infty} \inf_{\varrho_n} \sup_{g \in C^\beta(R)} R^{1/(2\beta+1)}n^{(2\beta+1)/(\beta+1)} \|w\|_{L^2}^{-2} \text{E}_g[(\hat{\varrho}_n - \varrho)^2] > 0.
\]

The infimum extends over all estimators from the PPP model with intensity \( n \).

Proof. The proof is based on a Bayesian risk bound, which clearly provides a lower bound for the minimax risk, see Korostelev and Tsybakov (1993) for similar approaches.

Take an independent Bernoulli sequence \((\varepsilon_k)\), i.e. \( P(\varepsilon_k = 1) = p, \ P(\varepsilon_k = 0) = 1 - p \) with \( p \in (0,1) \), and set for a triangular kernel \( K(y) = 2\min(y,1-y)1_{[0,1]}(y) \)

\[
g(x) = \sum_{k=0}^{h^{-1}-1} \varepsilon_k g_k(x) \text{ with } g_k(x) = cR^\beta K((x-kh)/h),
\]

where \( h \in (0,1) \) with \( h^{-1} \in \mathbb{N} \) will be chosen later. Then for \( c > 0 \) sufficiently small, we have \( g \in C^\beta(R) \) for all \( h \) and all realisations of \( (\varepsilon_k) \). We interpret this specification
of \( g \) as a prior on \( \mathcal{E}^\beta(R) \) and we shall make use of the independence of prior as well as the observation laws on different blocks \( I_k = [kh, (k + 1)h) \). For each \( k \) we obtain from the Bayes formula the posterior probability given the observations of the PPP in interval \( I_k \) (cf. the likelihood derivation in (3.1) below)

\[
\hat{\varepsilon}_k := P(\varepsilon_k = 1 \mid (X_i, Y_i)_{i \geq 1}) = \frac{p e^n f g_k \mathbf{1}(\forall X_i \in I_k : Y_i \geq g_k(X_i))}{1 - p + p e^n f g_k}.
\]

Using that \( \varepsilon_k \) are 0-1-valued, we have \( \hat{\varepsilon}_k = \mathbb{E}[\varepsilon_k | (X_i, Y_i)_{i \geq 1}] \) and \( \text{Var}(\varepsilon_k | (X_i, Y_i)_{i \geq 1}) = \hat{\varepsilon}_k (1 - \hat{\varepsilon}_k) \). Therefore the Bayes-optimal estimator of \( \vartheta \) under squared loss is given by

\[
\hat{\vartheta} = \sum_{k = 0}^{h^{-1} - 1} \hat{\varepsilon}_k \int_{I_k} g_k(x) w(x) \, dx.
\]

Using independence and \( \mathbb{E}[\hat{\varepsilon}_k - \varepsilon_k] = 0 \), its Bayes risk is calculated as

\[
\mathbb{E}[(\hat{\vartheta} - \vartheta)^2] = \sum_{k = 0}^{h^{-1} - 1} \text{Var}(\hat{\varepsilon}_k - \varepsilon_k) \left( \int_{I_k} g_k(x) w(x) \, dx \right)^2
\]

\[
= \sum_{k = 0}^{h^{-1} - 1} \mathbb{E} \left[ \text{Var}(\varepsilon_k \mid (X_i, Y_i)_{i \geq 1}) \right] \left( \int_{I_k} g_k(x) w(x) \, dx \right)^2
\]

\[
= \sum_{k = 0}^{h^{-1} - 1} \frac{p(1 - p)}{1 - p + p e^n f g_k} \left( \int_{I_k} g_k(x) w(x) \, dx \right)^2.
\]

We now choose \( h = [(cRn)^{1/(\beta + 1)}]^{-1} \) such that \( n \int_{I_k} g_k(x) \, dx = cRh^{\beta + 1}n \leq 1 \) holds. Then the Bayes risk is bounded in order by

\[
\mathbb{E}[(\hat{\vartheta} - \vartheta)^2] \gtrsim R^2 h^{2\beta + 1} \sum_{k = 0}^{h^{-1} - 1} \left( \int_{I_k} K((x - kh)/h) \|K((\bullet - kh)/h)\|_{L^2} w(x) \, dx \right)^2.
\]

The same argument over the shifted blocks \( I'_k = [(k + 1/2)h, (k + 3/2)h) \) implies that the minimax risk is bounded by the maximum (and thus the average) over the respective Bayes risks:

\[
\inf_{\vartheta_n} \sup_{g \in \mathcal{E}^\beta(R)} \mathbb{E}_g[(\hat{\vartheta}_n - \vartheta)^2] \gtrsim R^2 h^{2\beta + 1} \sum_{k = 0}^{2h^{-1} - 2} \left( \int_{I_k} \|K((x - kh/2)/h)\|_{L^2} w(x) \, dx \right)^2.
\]

The tent functions \( K((x - kh/2)/h), k = 0, 1, \ldots, 2h^{-1} - 2 \) form a Riesz basis for their linear span \( V_h \) (see e.g. Example 2.1 in Wojtaszczyk (1997)), which is the space of all linear splines with knots at \( kh/2 \), vanishing at the boundary. This means that the sum in the last display is larger than a constant times the \( L^2([0, 1]) \)-norm of the orthogonal projection of \( w \) onto \( V_h \). As \( \bigcup_{h > 0} V_h \) is dense in \( L^2([0, 1]) \), the \( L^2 \)-norm of the projections of \( w \) onto \( V_h \) converges for \( h \to 0 \) to the \( L^2 \)-norm of \( w \). Insertion of \( h \sim (Rn)^{-1/(\beta + 1)} \) yields the desired lower bound rate \( R^{1/(\beta + 1)} n^{-(2\beta + 1)/(\beta + 1)} \|w\|_{L^2}^2 \).
3 Nonparametric Maximum-Likelihood

3.1 The MLE over $C^\beta(R)$

Let us study the nonparametric maximum-likelihood estimator (MLE) in the class $C^\beta(R)$. Denote by $P_g$ the law of the observations in the PPP model with intensity $\lambda_g(x,y) = n1(y \geq g(x))$. Then the Radon-Nikodym-derivative $\frac{dP_g}{dP_{g_0}}$ for $g \geq g_0$ is by (a minor generalisation of) Thm. 1.3 in Kutoyants (1998)

$$\frac{dP_g}{dP_{g_0}} = \exp\left(n \int_0^1 (g - g_0)(x) \, dx\right)1\{\forall i : Y_i \geq g(X_i)\}.$$ (3.1)

A simple probability measure $P_0$ dominating all $P_g, g \in C^\beta(R)$, is given by the PPP model with intensity $\lambda_0(x,y) = n(e^y \wedge 1)$ and yields again by Thm. 1.3 in Kutoyants (1998) the likelihood

$$L(g) = \frac{dP_g}{dP_0} = \left(\prod_{j \geq 1} \frac{n1(Y_j \geq g(X_j))}{n(e^{Y_j} \wedge 1)}\right) \exp\left(-n \int_0^1 \int_{-\infty}^{\infty} (1(y \geq g(x)) - e^y \wedge 1) \, dy \, dx\right)$$

$$= \left(\prod_{j \geq 1} e^{(-Y_j)_+} 1(Y_j \geq g(X_j))\right) \exp\left(-n \int_0^1 (-1 - g(x)) \, dx\right)$$

$$= \exp\left(n + \sum_{j \geq 1} (-Y_j)_+\right) \exp\left(n \int_0^1 g(x) \, dx\right)1\{\forall j \geq 1 : Y_j \geq g(X_j)\}. \quad (3.1)$$

The first factor is independent of $g$ and we obtain thus the same structure as under $P_{g_0}$ above. The MLE over $C^\beta(R)$ is the function $\hat{g}$ that maximizes $\int_0^1 g$ over all $g \in C^\beta(R)$ with $g(X_j) \leq Y_j$ for all $j$. We can write explicitly

$$\hat{g}^{MLE}(x) = \min_{j \geq 1} \left(Y_j + R|x - X_j|^\beta\right),$$

since the right-hand side even maximises $g(x)$ over the considered class of $g$, see also Figure 2. The corresponding likelihood (with respect to $n$-dimensional Lebesgue measure) in the regression-type model (1.1) with $\varepsilon_i \sim \text{Exp}(\lambda)$ i.i.d. is given by

$$L^{reg}(g) = \lambda^n \exp\left(-\lambda \sum_{i=1}^n \varepsilon_i\right) \exp\left(\lambda \sum_{i=1}^n g(i/n)\right)1\{\forall i = 1, \ldots, n : Y_i \geq g(i/n)\}.$$ (3.2)

The maximum-likelihood estimator over $C^\beta(R)$ is then similarly given by

$$\hat{g}^{MLE-reg}(x) = \min_{i=1, \ldots, n} \left(Y_i + R|x - i/n|^\beta\right), \quad x \in [0, 1],$$

see Figure 2 for an illustration of the two constructions of the MLE. They are both quickly determined numerically. In the sequel, we shall focus on the MLE in the PPP.
3.1 Proposition. The nonparametric MLE \( \hat{g}^{MLE}(x), x \in [0, 1] \) is a sufficient and complete statistics for \( \mathcal{C}^\beta(R) \).

Proof. By definition of \( \hat{g}^{MLE} \) the likelihood (3.1) can be written as

\[
\mathcal{L}(g) = \exp \left( n + \sum_{j=1}^{\infty} (-Y_j)_+ \right) \exp \left( n \int_0^1 g(x) \, dx \right) 1 \left( g \leq \hat{g}^{MLE} \right)
\]

such that by Neyman’s factorisation criterion (e.g. Lehmann and Romano (2006)) \( \hat{g}^{MLE} \) is a sufficient statistics for this parameter class.

To establish completeness, let us first remark that by definition \( \hat{g}^{MLE} \) is an element of \( \mathcal{C}^\beta(R) \). Since \( \mathcal{C}^\beta(R) \), equipped with its \( \mathcal{C}^\beta \)-norm, is not separable, we equip it with the Borel \( \sigma \)-algebra generated by the uniform (supremum) norm, which is generated by all point evaluations. Measurability of the estimator \( \hat{g}^{MLE} \) is then easily established since all point evaluations \( \hat{g}^{MLE}(x), x \in [0, 1] \), are measurable as minima of countably many random variables.

For completeness we now consider any statistic \( T : \mathcal{C}^\beta(R) \to \mathbb{R} \) satisfying \( \mathbb{E}_g[T(\hat{g}^{MLE})] = 0 \) for all \( g \in \mathcal{C}^\beta(R) \), which is Borel measurable with respect to the uniform norm. For \( g \in \mathcal{C}^\beta(R) \) denote by \( [g, \infty) := \{ h \in \mathcal{C}^\beta(R) \mid h \geq g \} \) the ’bracket’ between \( g \) and \( \infty \), that is all functions above \( g \). Noting \( [g, \infty) \cap [h, \infty) = [g \lor h, \infty) \) where the maximum \( g \lor h \) is again in \( \mathcal{C}^\beta(R) \), the family \( \{ [g, \infty) \mid g \in \mathcal{C}^\beta(R) \} \) is an \( \cap \)-stable generator of the uniform Borel \( \sigma \)-algebra in \( \mathcal{C}^\beta(R) \): for any \( x_0 \in [0, 1], y_0 \in \mathbb{R} \) we have \( \{ h \in \mathcal{C}^\beta(R) \mid h(x_0) \geq y_0 \} = [y_0 - R[\bullet - x_0]^\beta, \infty) \) by the Hölder condition and \( \{ [y_0, \infty) \mid y_0 \in \mathbb{R} \} \) generates the Borel \( \sigma \)-algebra on \( \mathbb{R} \).
Let us define another weighted counting process
\[ \hat{N}(t) = \sum_{j \geq 1} 1\{ Y_j \leq t \wedge \min_{i \geq 1} (Y_i + R|X_j - X_i|^{\beta}) \} w(X_j), \quad t \in \mathbb{R}. \]

Note that the minimum can be taken over the index set \( \{ i : Y_i < t \} \) only such that \( \bar{N} \) is again adapted to \( (\mathcal{F}_t) \) from [2.1]. Given the stochastic intensity \( \lambda_t = n \int_0^1 \int_{[y(x), t]} 1(\min_{i \leq s} (Y_i + R|x - X_i|^{\beta}) \geq s) ds dx \) for the pure counting process, we obtain by compensation (cf. Prop. 2.32 in Karr (1991)) the \( (\mathcal{F}_t) \)-martingale
\[ \bar{M}(t) = \bar{N}(t) - n \int_0^t \int_{[y(x), t]} 1\left( \min_{s \leq s} (Y_i + R|x - X_i|^{\beta}) \geq s \right) ds w(x) dx. \]
The main observation is the identity

$$
\lim_{t \to \infty} (N(t) - \bar{M}(t)) = n \int_0^1 \int_{g(x)}^\infty 1 \left( \min_{i \geq 1} (Y_i + R|x|^{\beta}) \geq s \right) \, ds \, w(x) \, dx
$$

$$
= n \int_0^1 \int_{g(x)}^\infty 1(\hat{g}^{MLE}(x) \geq s) \, ds \, w(x) \, dx
$$

$$
= n \int_0^1 (\hat{g}^{MLE}(x) - g(x))w(x) \, dx,
$$

which tells us that

$$
\bar{N}(\infty) := \lim_{t \to \infty} N(t) = \sum_{j \geq 1} 1(\hat{g}^{MLE}(x_j) \geq Y_j)w(x_j) = \sum_{j \geq 1} 1(\hat{g}^{MLE}(x_j) = Y_j)w(x_j)
$$

simultaneously counts the weighted number of points \((x_j, Y_j)\) on the graph of \(\hat{g}^{MLE}\) and equals the scaled bias \(n \int_0^1 \hat{g}^{MLE}(x)w(x) \, dx - \beta\) up to a martingale term. We thus have

$$
\hat{M}(\infty) = \sum_j 1(Y_j \leq \hat{g}^{MLE}(x_j))w(x_j) - \int_0^1 (\hat{g}^{MLE}(x) - g(x))w(x) \, dx
$$

is the a.s. and \(L^2\)-limit of the \(L^2\)-bounded martingale \(\bar{M}\) with

$$
\langle \bar{M} \rangle_t = n \int_0^1 \int_{[g(x), t]} 1 \left( \min_{i : Y_i < s} (Y_i + R|x|^{\beta}) \geq s \right) \, ds \, w(x)^2 \, dx
$$

$$
\uparrow n \int_0^1 (\hat{g}^{MLE}(x) - g(x))w(x)^2 \, dx =: \langle \bar{M} \rangle_\infty \text{ as } t \uparrow \infty.
$$

We obtain from \(\mathbb{E}[\hat{g}^{MLE} - \beta] = \frac{1}{n} \mathbb{E}[-\bar{M}(\infty)]\), \(\text{Var}(\hat{g}^{MLE}) = \frac{1}{n^2} \text{Var}(\bar{M}(\infty))\) the result (use Lemma 5.1 with \(\tau = \infty\))

$$
\mathbb{E}[\hat{g}^{MLE}] = \beta \text{ and } \text{Var}(\hat{g}^{MLE}) = \frac{1}{n} \mathbb{E}[\langle \bar{M} \rangle_\infty] = \frac{1}{n} \int_0^1 \mathbb{E}[\hat{g}^{MLE}(x) - g(x)]w(x)^2 \, dx.
$$

Hence, \(\hat{g}^{MLE}\) is an unbiased estimator and by the Lehmann-Scheffé Theorem \(\hat{g}^{MLE}\), derived from a sufficient and complete statistics, is uniformly minimum variance among all unbiased estimators (e.g. Lehmann and Romano (2006)).

To bound the variance we use a universal, but somewhat rough deviation bound for \(s \geq 0\) and \(x \in [0, 1]\):

$$
P(\hat{g}^{MLE}(x) - g(x) \geq s) = \exp \left( - n \int_0^1 (s - R|x|^{\beta} + g(x) - g(x_\xi))_+ \, d\xi \right)
$$

$$
\leq \exp \left( - n \int_0^1 (s - 2R|x|^{\beta})_+ \, d\xi \right)
$$

$$
\leq \begin{cases} 
\exp(-n\frac{2R}{\beta+1}(s/2R)^{\beta+1}/\beta), & s \in [0, 2R], \\
\exp(-n(s - 2R/(\beta + 1))), & s > 2R.
\end{cases} \quad (3.3)
$$
In the first step we have evaluated the probability that no observation lies in \( \{(\xi, \eta) \mid \eta + RX - \xi < g(\bar{x}) + s\} \) using the PPP property. Integrating this survival function bound, we obtain directly

\[
\mathbb{E}[\hat{g}\text{MLE}(x) - g(x)] = \int_0^\infty P(\hat{g}\text{MLE}(x) - g(x) \geq s) \, ds
\]

\[
\leq \int_0^{2R} \exp \left( -n \frac{2R}{\beta + 1} \frac{(s/2R)^{(\beta+1)/\beta}}{\beta + 1} \right) ds + \int_{2R}^{\infty} e^{-n(s-2R/(\beta+1))} ds
\]

\[
= \Gamma(\beta/(\beta + 1)) \beta(2R/(\beta + 1))^{1/(\beta+1)} n^{-\beta/(\beta+1)} + \frac{1}{n} e^{-2\beta R n/(\beta+1)}. 
\]

Insertion and a numerical evaluation then yield (the maximal constant being attained for \( \beta \to 0 \)) \( \text{Var}(\hat{g}\text{MLE}) \leq (2 + o(1))R^{1/(\beta+1)}||w||_2^2 n^{-2(\beta+1)/(\beta+1)}. \)

3.3 Remark. The MLE \( \hat{g}\text{MLE-regr} \) from (3.2) for the regression-type model is by the same (or simpler) arguments a sufficient and complete statistics over \( \mathcal{C}^\beta(R) \). It gives rise to the analogue estimator

\[
\hat{\vartheta}_n\text{MLE-regr} = \frac{1}{n} \sum_{i=1}^n \left( \hat{g}\text{MLE-regr}(i/n) - \lambda^{-1} \mathbb{1}(\hat{g}\text{MLE-regr}(i/n) = Y_i) \right) w(i/n).
\]

Then for Exp(\( \lambda \))-distributed errors \( \hat{\vartheta}_n\text{MLE-regr} \) is an unbiased estimator of \( \vartheta^{(n)} = \frac{1}{n} \sum_{i=1}^n g(i/n)w(i/n) \) with \( \text{Var}(\hat{\vartheta}_n\text{MLE-regr}) = \frac{1}{n^2} \sum_{i=1}^n \mathbb{E}[\hat{g}\text{MLE-regr}(i/n) - g(i/n)]w(i/n)^2. \) This follows analogously from using the corresponding counting process \( \bar{N}(t) \), replacing \( X_j \) in the PPP case by \( j/n. \) The asymptotic upper bound for the regression model as \( n \to \infty \) is the same as for the PPP model, but with the noise level \( 1/n \) replaced by \( 1/(n\lambda) \), provided \( w^2 \) is Riemann-integrable.

While the MLE as an UMVU estimator enjoys very desirable finite sample properties of its risk, for inference questions we are also in need of distributional properties, at least asymptotically. A priori, in our non-regular frontier models it might not be clear whether the limiting distribution is Gaussian, but in fact this is the case since we average implicitly over the interval \([0,1]\). The proof of the following central limit theorems is slightly more technical and therefore given in the appendix.

3.4 Theorem. Suppose that \( \sup_{x \in [0,1]} |w(x)| < \infty \) and that \( \text{Var}(\hat{\vartheta}_n\text{MLE}) \sim n^{-(2\beta+1)/(\beta+1)}, \) indicating the dependence of \( \hat{\vartheta}\text{MLE} \) on \( n \). Then the following central limit theorems hold as \( n \to \infty \):

\[
\frac{n^{1/2}(\hat{\vartheta}_n\text{MLE} - \vartheta)}{\left( \int_0^1 \mathbb{E}[\hat{g}\text{MLE}(x) - g(x)]w(x)^2 \, dx \right)^{1/2}} \Rightarrow N(0,1),
\]

\[
\frac{n^{1/2}(\hat{\vartheta}_n\text{MLE} - \vartheta)}{\left( \int_0^1 (\hat{g}\text{MLE}(x) - g(x))w(x)^2 \, dx \right)^{1/2}} \Rightarrow N(0,1).
\]

Furthermore, the following self-normalising version is valid:

\[
\frac{n^{1/2}(\hat{\vartheta}_n\text{MLE} - \vartheta)}{\left( \frac{1}{n} \sum_{j \geq 1} \mathbb{1}(\hat{g}\text{MLE}(X_j) = Y_j)w(X_j)^2 \right)^{1/2}} \Rightarrow N(0,1).
\]
3.5 Corollary. Under the assumptions of Theorem $\beta.4$

\[ J_n := \left[ \hat{\sigma}_n^{MLE} - \bar{\sigma}_n q_{1-\alpha/2}, \hat{\sigma}_n^{MLE} + \bar{\sigma}_n q_{1-\alpha/2} \right], \hat{\sigma}_n^2 := \frac{1}{n^2} \sum_{j \geq 1} 1(\hat{g}_n^{MLE}(X_j) = Y_j)w(X_j)^2, \]

with \( q_{1-\alpha/2} \) the \((1 - \alpha/2)\)-quantile of \( N(0, 1) \), is a confidence interval for \( \vartheta \) with asymptotic level \( 1 - \alpha \).

Proof. The selfnormalising CLT shows \( \lim_{n \to \infty} P_\vartheta(\hat{\sigma}_n^{-1} |\hat{\vartheta}^{MLE} - \vartheta| \leq q_{1-\alpha/2}) = 1 - \alpha \).

3.6 Remark. A lower estimate in (3.3) above shows that \( \mathbb{E}[\hat{g}_n^{MLE}(x) - g(x)] \geq n^{-\beta/(\beta+1)} \) holds as soon as the function \( g \in \mathcal{C}^\beta(R) \) satisfies \( |g(y) - g(x)| \leq R'|y - x|\beta \) for \( R' < R \) and \( y \) in a neighbourhood of \( x \). This means that \( \text{Var}(\hat{\vartheta}^{MLE}) \sim n^{-(2\beta+1)/(\beta+1)} \) and the CLTs above are applicable whenever \( g \) has a local \( \beta \)-Hölder constant smaller than \( R \), at least on some small interval. Note that in this case we also get 'for free' the nice geometric result that the number of observations on the graph of \( \hat{g}_n^{MLE} \) is of order \( n^{1/(\beta+1)} \) (in mean) because of

\[ \frac{1}{n} \sum_{j \geq 1} 1(\hat{g}_n^{MLE}(X_j) = Y_j) \sim \int_0^1 \mathbb{E}[\hat{g}_n^{MLE}(x) - g(x)] dx \sim n^{-\beta/(\beta+1)}. \]

The standard deviation is of smaller order as the proof of Theorem $\beta.4$ shows.

In the case \( \beta = 1 \) the asymptotic variance can be determined explicitly.

3.7 Proposition. For \( \beta = 1 \), i.e. \( g \in \mathcal{C}^1(R) \), we obtain

\[ \text{Var}(\hat{\vartheta}_n^{MLE}) = \left( \frac{\sqrt{\pi}}{2} + o(1) \right)n^{-3/2} \int_0^1 \sqrt{(R^2 - g'(x)^2)/Rw(x)^2} dx, \]

where \( g' \) denotes the weak derivative of the Lipschitz function \( g \), and thus

\[ n^{3/4}(\hat{\vartheta}_n^{MLE} - \vartheta) \Rightarrow N(0, \int_0^1 \sqrt{(R^2 - g'(x)^2)/Rw(x)^2} dx). \]

Proof. A Lipschitz function \( g \) is absolutely continuous, hence a.e. differentiable and necessarily \( |g'(x)| \leq R \) holds a.e. For \( x \in (0, 1) \) where \( g'(x) \) exists we obtain, arguing by dominated convergence using (3.3),

\[ P\left( n^{1/2}(\hat{g}_n^{MLE}(x) - g(x)) \geq z \right) \]

\[ = \exp \left( -n \int_0^1 (g(x) + zn^{-1/2} - R|\xi - x| - g(\xi))_+ d\xi \right) \]

\[ = \exp \left( -\int_{-n^{1/2}}^{n^{1/2}} (n^{1/2}(g(x) - g(x + n^{-1/2}w)) + z - R|w|)_+ dw \right) \]

\[ \to \exp \left( -\int_{-\infty}^{\infty} (z - R|w| - g'(x)w)_+ dw \right) = \exp \left( -\frac{R}{R^2 - g'(x)^2} z^2 \right). \]
By integrating this survival function and applying dominated convergence due to the uniform bound \(3.3\), we conclude

\[
n_{1/2} \mathbb{E}[g_n^{(MLE)}(x) - g(x)] \to \sqrt{(R^2 - g'(x)^2)/R \pi/2}.
\]

Integration over \(x\) yields by another application of dominated convergence the asymptotic expression for \(\text{Var}(\hat{\vartheta}^{MLE})\).

3.8 Remark.

1. The proposition shows that for constant \(g\) the asymptotic variance (rescaled by \(n^{3/2}\)) is with \(\sqrt{R\pi}\|w\|_{L^2}^2\) largest while for linear \(g\) with slope \(\pm R\) the rescaled asymptotic variance vanishes, i.e. the convergence rate is faster than \(n^{-3/2}\). In Figure 2 we see indeed that \(\hat{g}^{MLE}\) is closest to \(g\) where \(g\) has largest slope.

2. By the same arguments, we derive for \(\beta \in (0, 1)\) and functions \(g \in C^\beta(R)\) with \(\lim_{h \to 0} (g(x + h) - g(x))/h^\beta = g^{(\beta)}(x) \in [-R, R]\) for Lebesgue-almost all \(x \in \mathbb{R}\) that \(n^{2\beta+1}/(\beta+1)\) \(\text{Var}(\hat{g}_n^{MLE})\) converges to

\[
\Gamma\left(\frac{2\beta+1}{\beta+1}\right) \int_0^1 \left(\frac{(R + g^{(\beta)}(x))-1/\beta + (R - g^{(\beta)}(x))-1/\beta}{1 + 1/\beta}\right)^{-\beta/(\beta+1)} w(x)^2 dx.
\]

It is not clear, however, how large the class of \(g \in C^\beta(R)\) is whose limit \(g^{(\beta)}\) exists almost everywhere.

3. Our method of proof does not permit to require only \(w \in L^2\) because we need to control the difference to a blockwise MLE. It does also not extend to the monotone MLE introduced next.

3.2 MLE under monotonicity

Let us consider the general nonparametric class

\[\mathcal{M} := \{g : [0, 1) \to \mathbb{R} \mid g \text{ is increasing and left-continuous}\}\]

of monotone, that is (not necessarily strictly) increasing functions. Since monotone \(g\) have at most countably many jumps, the observations for left- and right-continuous version of \(g\) are a.s. identical. Then the nonparametric MLE for the PPP model over this class is given by

\[\hat{g}^{Mon}(x) = \min_{i: X_i \geq x} Y_i, \quad x \in [0, 1),\]

which is obvious from the fact that any \(g \in \mathcal{M}\) with \(g(X_i) \leq Y_i\) for all \(i\) necessarily satisfies \(g \leq \hat{g}^{Mon}\), see also Figure 3. Note that a.s. \(\hat{g}^{Mon}(x) < \infty\) holds for \(x \in [0, 1)\), but \(\lim_{x \uparrow 1} \hat{g}^{Mon}(x) = \infty\).
Figure 3: Construction of the estimator $\hat{g}^{Mon}$ in the PPP model (left) and in the regression-type model (right).

3.9 Proposition. The nonparametric MLE $(\hat{g}^{Mon}(x), x \in [0,1])$ is a sufficient and complete statistics for $\mathcal{M}$.

Proof. Sufficiency follows again from the likelihood representation

$$\mathcal{L}(g) = \exp(n + \sum_{j \geq 1} (-Y_j)_+) \exp \left(n \int_0^1 g(x) \, dx \right) \mathbf{1}(g \leq \hat{g}^{Mon}),$$

using $g \in L^1$ because of $g(x) \in [g(0), g(1)]$ by monotonicity, and the factorisation criterion. For completeness we equip $\mathcal{M}$ with the ball $\sigma$-algebra for the uniform norm, cf. Examples 1.7.3, 1.7.4 in Van Der Vaart and Wellner (1996), which is generated by the point evaluations $f \mapsto f(x), x \in [0,1)$. In particular, this implies that $\hat{g}^{MLE}$ is measurable because its point evaluations are measurable. For fixed $x_0 \in [0,1), y_0 \in \mathbb{R}$ we have the bracket representation

$$\{g \in \mathcal{M} \mid g(x_0) \geq y_0\} = \bigcup_{n \in \mathbb{N}} \left[ y_0 - n \mathbf{1}_{[0,x_0]}(x), \infty \right).$$

Noting that maxima of monotone functions are again monotone, the brackets form an $\cap$-stable generator of the ball $\sigma$-algebra. The proof now follows exactly that of Proposition 3.1.

In analogy with the MLE over $C^\beta(R)$ we build the estimator

$$\hat{g}^{Mon} := \int_0^1 \hat{g}^{Mon}(x) w(x) \, dx - \frac{1}{n} \sum_{j \geq 1} \mathbf{1}(\hat{g}^{Mon}(X_j) = Y_j) w(X_j)$$

that will enjoy similar nice properties. We have to consider, however, weight functions $w$ whose support stays away from $x = 1$ in order to avoid problems arising from $\hat{g}^{Mon}(x) \uparrow \infty$ as $x \uparrow 1$. 

18
3.10 Theorem. Assume \( \text{supp}(w) \subseteq [0,1] \). Then the estimator \( \hat{\theta}^\text{Mon} \) is for each finite sample size \( n \) UMVU over the class \( \mathcal{M} \) with

\[
\text{Var}(\hat{\theta}^\text{Mon}) = \frac{1}{n} \int_0^1 \mathbb{E}[\hat{g}^\text{Mon}(x) - g(x)]w(x)^2dx.
\]

For \( g \in \mathcal{C}^\beta(R) \cap \mathcal{M} \) and \( n \to \infty \) it satisfies

\[
\text{Var}(\hat{\theta}^\text{MLE}) \leq (1.2883 + o(1))R^{1/(\beta+1)}\|w\|^2_{L^2} n^{-(2\beta+1)/(\beta+1)}.
\]

3.11 Remark. The derivation of the upper bound for \( g \in \mathcal{C}^\beta(R) \) will show that the right endpoint \( r \in [0,1] \) of the support of \( w \) only enters in the remainder term (with respect to \( n \)). More precisely, for sequences \( (w_n) \) of weight functions whose support endpoints \( r_n \) satisfy \( n(1 - r_n)^{\beta+1} \gtrsim 1 \) we still have

\[
\text{Var}(\hat{\theta}^\text{MLE}/\|w_n\|_{L^2}) \lesssim R^{1/(\beta+1)} n^{-(2\beta+1)/(\beta+1)}.
\]

In particular, if we know a priori that \( w(x) \) is bounded for \( x \in [1 - \delta, 1] \), we can use the estimator for

\[
w_n(x) = w(x)1_{[0,1-n^{-1/2}]}(x) + w(x + n^{-1/2})1_{[1-2n^{-1/2}, 1-n^{-1/2}]}(x),
\]

whose variance is still of order \( n^{-(2\beta+1)/(\beta+1)} \) while its bias contributes with

\[
\left| \int_{1-n^{-1/2}}^1 (g(x) - g(x - n^{-1/2}))w(x)dx \right| \leq \int_{1-n^{-1/2}}^1 Rn^{-\beta/2}w(x)dx \\
\lesssim n^{-(\beta+1)/2} \lesssim n^{-(\beta+1/2)/(\beta+1)}
\]

at most as much as the stochastic error to the mean squared error.

Proof. The proof follows along the lines of the proof for Theorem 3.2. Here the weighted counting process is

\[
\tilde{N}(t) = \sum_{j \geq 1} 1(Y_j \leq t \land \min_{i, X_i \geq X_j} Y_i)w(X_j), \quad t \in \mathbb{R}.
\]

Its intensity is \( \lambda_t = n \int_0^t \int_{[g(x), t]} 1(\min_{i, X_i \geq x} Y_i \geq s)dsdx \) and compensation yields the corresponding martingale \( \tilde{M}(t) \). The same limiting arguments, by restriction to the support of \( w \), then yield again \( \mathbb{E}[\hat{\theta}^\text{Mon}] = \vartheta \) and

\[
\text{Var}(\hat{\theta}^\text{Mon}) = \frac{1}{n} \int_0^1 \mathbb{E}[\hat{g}^\text{Mon}(x) - g(x)]w(x)^2dx.
\]

It remains to estimate the last expectation by a deviation bound for \( g \in \mathcal{C}^\beta(R) \):

\[
P(\hat{\theta}^\text{Mon}(x) - g(x) \geq s) = \exp \left( -n \int_x^1 (s + g(x) - g(\xi))_+ d\xi \right)
\]
\[ \mathbb{E}[g^{\text{Mon}}(x) - g(x)] \leq \int_{0}^{\infty} e^{-n R^{-1/\beta} s^{(\beta+1)/\beta}} ds + e^{n R^{-1/\beta} (1-x)^{\beta+1}} \int_{(1-x)}^{\infty} e^{-ns(1-x)} ds \\
= \Gamma(\beta/(\beta+1)) \left( \frac{\beta}{\beta+1} \right)^{1/(\beta+1)} R^{1/(\beta+1)} n^{-\beta/(\beta+1)} + \frac{1}{n(1-x)} e^{-n R^{1/(\beta+1)}}. \]

Insertion and a numerical maximisation yield the result. \( \square \)

In the regression-type model the nonparametric MLE over \( \mathcal{M} \) is likewise 
\( \hat{g}^{\text{Mon-regr}}(x) = \min_{i:x \leq i/n} y_i \). Then by the same arguments
\[ \hat{g}_n^{\text{Mon-regr}} := \frac{1}{n} \sum_{i=1}^{n} \left( \hat{g}^{\text{Mon-regr}}(i/n) - \lambda^{-1} \mathbf{1}(\hat{g}^{\text{Mon-regr}}(i/n) = y_i) \right) w(i/n) \]
is an unbiased estimator of \( \frac{1}{n} \sum_{i=1}^{n} g(i/n) w(i/n) \) under \( \text{Exp}(\lambda) \)-noise with 
\( \text{Var}(\hat{g}_n^{\text{Mon-regr}}) = \frac{1}{n \lambda} \sum_{i=1}^{n} \mathbb{E}[\hat{g}^{\text{Mon-regr}}(i/n) - g(i/n)]^2 w(i/n)^2 \). Note that at the right end-point \( \mathbb{E}[\hat{g}^{\text{Mon-regr}}(1) - g(1)] = \lambda^{-1} \) holds, but that summand only contributes \( (n \lambda)^{-2} w(1)^2 \) to the total variance which is usually negligible. Asymptotic variance bounds for \( g \in C^\beta(R) \) can be derived similarly, replacing the noise level 1/n by 1/(n\lambda).

## 4 Discussion

### 4.1 Adaptivity, series estimator and extensions

An important implication of the result for the monotone MLE is that it adapts to the Hölder smoothness of the underlying regression function \( g \), i.e. it satisfies similar risk bounds as in the case of knowing \( g \in C^\beta(R) \) and of using this prior information in the construction of the estimator. In the case of \( \hat{g}^{\text{block}} \) or \( \hat{g}^{\text{MLE}} \) the construction essentially relies on the values of \( \beta \) and \( R \). For \( \hat{g}^{\text{block}} \) a misspecification, using some \( \beta' > \beta \), yields a block-wise upper function \( y_k + Rh^{\beta'} \) and thus induces a bias of maximal size \( R(h^{\beta} - h^{\beta'}) \). The rate-optimal choice \( h \sim n^{-1/(\beta'+1)} \) yields a bias upper bound of order \( O((n^{-\beta/(\beta'+1)} - n^{-\beta'/(\beta'+1)}))_+ \), while the variance remains of order \( n^{-(2\beta+1)/(\beta'+1)} \). In the parameter \( \beta' \) the estimators \( (\hat{g}^{\text{block}}(\beta'), \beta' \in [0,1]) \) are thus ordered in variance and (reversely) in bias. In this situation, the Lepski method offers a natural way to obtain a data-driven choice of the parameter \( \beta' \), see Jirak, Meister, and Reiß (2014) for the case of function estimation with one-sided errors. Due to the fact that we have much more structure here (unbiasedness for \( \beta' \leq \beta \)), we might even hope for a simpler method, but this remains to be explored.
An important application for the estimation of functionals are orthogonal series estimators, also called projection estimators. Let \((\varphi_m)_{m \geq 1}\) be an orthonormal basis of \(L^2([0,1])\). Then we can form the estimator \(\hat{g}_M = \sum_{m=1}^M \hat{\vartheta}_m \varphi_m\) of \(g\) where \(\hat{\vartheta}_m\) estimates the coefficient \(\langle g, \varphi_m \rangle_{L^2}\), i.e. \(w = \varphi_m\) in our notation. Using our estimators for \(g \in \mathcal{C}^\beta(R)\) we thus obtain as stochastic error in the \(L^2\)-risk:

\[
\mathbb{E} \left[ \| \hat{g}_M - \mathbb{E}[\hat{g}_M] \|^2_{L^2} \right] = \sum_{m=1}^M \text{Var}(\hat{\vartheta}_m) \lesssim M n^{-(2\beta+1)/(\beta+1)}.
\]

For \(L^2\)-Sobolev spaces \(H^s\) of regularity \(s\) and standard bases like (trigonometric) polynomials, splines or wavelets we have the bias bound \(\sum_{m>M} \langle g, \varphi_m \rangle^2 \lesssim M^{-2s}\). We always have \(g \in \mathcal{C}^\beta(R) \Rightarrow g \in H^\beta\) such that

\[
\mathbb{E}[\| \hat{g}_M - g \|^2_{L^2}] \lesssim M^{-2\beta} + M n^{-(2\beta+1)/(\beta+1)} \sim n^{-2\beta/(\beta+1)}\text{ for } M \sim n^{1/(\beta+1)}
\]

follows. This seems to be the first rate-optimal estimation result for series estimators in one-sided regression, cf. Jirak, Meister, and Reiβ (2014) for optimal rates and other approaches in the literature. We may, of course, also have \(g \in H^s\) for some \(s > \beta\), but then the derived rate is slower than the optimal \(n^{-2s/(2s+1)}\). The unbiased estimation method essentially relies on a uniform control of the variation of \(g\) and we do not know whether similar results can be obtained for Sobolev (or Besov) instead of Hölder balls.

Concerning the function class \(\mathcal{G}\) over which the nonparametric MLE is feasible and for which the derived estimator of \(\vartheta\) exhibits nice non-asymptotic properties, it was essential for the stopping arguments as well as the completeness property that constants lie in \(\mathcal{G}\) and that for \(g_1, g_2 \in \mathcal{G}\) also \(g_1 \wedge g_2\) and \(g_1 \vee g_2\) are in \(\mathcal{G}\). Thus also \(\mathcal{G} = \mathcal{M} \cap \mathcal{C}^\beta(R)\) or extensions to the multivariate case \(\mathcal{G} = \mathcal{C}^\beta_d(R) = \{g : [0,1]^d \rightarrow \mathbb{R} \mid |g(x) - g(y)| \leq R|x-y|^\beta\}\) are possible. For smoothness degrees \(\beta > 1\) or other shape constraints like convexity, however, our method does not transfer directly.

### 4.2 Monte Carlo experiment

In a small simulation study we investigate the behaviour of the blockwise estimator, the MLE and the monotone MLE for \(\vartheta = \int_0^1 g(x) dx\) on finite samples. We simulate the PPP model as well as the regression-type model with two different monotone regression functions \(g\). The RMSE (root mean squared error) estimate is based on \(M = 200\) Monte Carlo repetitions in each case. On the left-hand side of Figure 4 the RMSE results for \(g(x) = 0.5 \sin(2\pi x) + 4x\) are shown and on the right-hand side those for \(g(x) = \sqrt{x}\). It can be seen that all three estimators work well even for the small sample size \(n = 50\) and that their performances in the PPP and the regression model are comparable. The blockwise estimator, in view of its simplicity, does not perform so much worse than the ML estimators. From our theoretical results this is to be expected: the ratio of the upper bounds for the nonasymptotic variance of \(\hat{g}^{MLE}\) and of \(\hat{g}^{\text{block}}\) is given by

\[
\frac{\Gamma(\beta/(\beta+1))(\beta+1)^{-(1/(\beta+1))}}{\beta^{-(\beta/(\beta+1))(\beta+1)}} = \Gamma((\beta/(\beta+1))(\beta+1)^{(2\beta+1)/(\beta+1)}(\beta+1)^{-(\beta+2)/(\beta+1)}),
\]

21
Figure 4: Monte-Carlo errors for the different estimators and two functions \( g \) which approaches one for \( \beta \downarrow 0 \), has a minimum 0.54 at \( \beta \approx 0.47 \) and then increases to about 0.63 for \( \beta \uparrow 1 \).

5 Appendix

5.1 Technical lemmata

We formulate a stopping theorem for continuous-time martingales, which does not seem readily available in the literature.

5.1 Lemma. Let \((M(t), t \geq t_0)\) be a càdlàg martingale with \( M(t_0) = 0 \) and let \( \tau \) be a stopping time with values in \([t_0, \infty)\), both on some filtered probability space. If \( \mathbb{E}[(M)_\tau] \) is finite, then \( \mathbb{E}[M(\tau)] = 0 \) and \( \mathbb{E}[M(\tau)^2] = \mathbb{E}[(M)_\tau] \) hold.

Proof. From the Burkholder-Davis-Gundy inequality (Thm. 26.12 in Kallenberg (2002)) and the identity \( \mathbb{E}[|M|] = \mathbb{E}[(M)_\tau] \) (e.g. by Prop. 4.50(c) in Jacod and Shiryaev (1987) for \( M^r \)), we conclude \( \mathbb{E}[\sup_{t \geq t_0} M^2_{t \wedge \tau}] \leq \mathbb{E}[(M)_\tau] \). Hence, \( (|M|_{t \wedge \tau})_{t \geq t_0} \) is uniformly integrable and by optional stopping \( \mathbb{E}[M_\tau] = \lim_{t \to \infty} \mathbb{E}[M_{t \wedge \tau}] = 0 \) follows as well as \( \mathbb{E}[M^2_\tau] = \mathbb{E}[(M)_\tau] = \mathbb{E}[(M)_\tau] \). 

A moment bound for the stopping time in the proof of Theorem 2.2 is provided.

5.2 Lemma. Under the assumptions of Theorem 2.2 we have for \( \tau = \gamma_k^* + R h^\beta \)

\[
\mathbb{E}[(\tau - g(i/n))^{1/p}] \leq R h^\beta + (n \lambda h)^{-1}
\]

as \( nh \to \infty \) for any \( p > 0 \).

Proof. The property \( \gamma_k^* \leq \max_{i \in \tilde{I}_k} g(i/n) + \min_{i \in \tilde{I}_k} \varepsilon_i \) implies for \( nh \to \infty \)

\[
P \left( n \lambda h (\gamma_k^* - \max_{i \in \tilde{I}_k} g(i/n)) \geq z \right) \leq F_{\varepsilon}(z/n \lambda h)^n \leq e^{nh \log F_{\varepsilon}(z/n \lambda h)} \to e^{-z}.
\]
Using $F(z/nh)^{n} \leq (1 + z/nh)^{-nh}$, we establish

$$\limsup_{R \to \infty} \sup_{nh} \int_{R}^{\infty} z^{p-1} P\left(n\lambda h \left(\frac{y_i}{\lambda h} - \max_{i \in I_k} g(i/n)\right) \geq z\right) dz = 0$$

for any $p \geq 1$ such that by uniform integrability

$$\limsup_{nh \to \infty} E\left[\left(n\lambda h \left(\frac{y_i}{\lambda h} - \max_{i \in I_k} g(i/n)\right)\right)^{p}\right] \leq \int_{0}^{\infty} z^{p} e^{-z} dz < \infty$$

follows. By the Hölder condition $g$ varies at most by $Rh^\beta$ on each block and thus $E[(\tau - \min_{i \in I_k} g(i/n))^p] \leq (Rh^\beta + (n\lambda h)^{-1})^p$ holds.

We need the following interesting self-normalising property. The constant is certainly not optimal.

**5.3 Lemma.** Suppose that a non-negative random variable $X$ satisfies $P(X \geq x) = e^{-a(x)}$, $x \geq 0$, with a strictly increasing convex function $a$. Then $E[X^2] \leq 6(e+1)E[X]^2$ holds.

**Proof.** The property $P(a(X) \geq a(x)) = e^{-a(x)}$ shows that $Y := a(X)$ is Exp(1)-distributed. Since the inverse $a^{-1}$ of $a$ exists, we may use $a^{-1}(0) = 0$ and the concavity of $a^{-1}$ to calculate

$$E[X^2] = E[a^{-1}(Y)^2] = \int_{0}^{\infty} \int_{0}^{\infty} a^{-1}(x)a^{-1}(y)e^{-x-y} dy dx$$

$$= \int_{0}^{\infty} \int_{0}^{x} a^{-1}(x)a^{-1}(z-x) dx e^{-z} dz$$

$$\geq \int_{0}^{\infty} \int_{0}^{x} \frac{x}{z} a^{-1}(z) - \frac{x}{z} a^{-1}(z) dx e^{-z} dz = \int_{0}^{\infty} \frac{z}{6} a^{-1}(z)^2 e^{-z} dz.$$

By monotonicity, we have $\int_{0}^{1} z a^{-1}(z)^2 e^{-z} dz \geq e^{-1} \int_{0}^{1} a^{-1}(z)^2 e^{-z} dz$. This shows

$$6E[X^2] \geq \frac{1}{e+1} \int_{0}^{\infty} a^{-1}(z)^2 e^{-z} dz = \frac{1}{e+1} E[X^2].$$

\[\Box\]

**5.2 Proof of Theorem 3.4**

Let $r_n \to 0$ such that $r_n^{2\beta+1} \to \infty$ and $r_n^{-1} \in \mathbb{N}$. On each block $J_l = [lr_n, (l+1)r_n)$, $l = 0, \ldots, r_n^{-1} - 1$, we can define the blockwise MLE

$$\hat{g}_l^{MLE} = \min_{i : X_i \in J_l} (Y_i + R|x - X_i|^\beta), \quad x \in J_l.$$  

Note that by definition the blockwise MLE is at least as large as the global MLE, i.e. $\hat{g}_l^{MLE} \geq \hat{\gamma}^{MLE}$. By construction, $(\hat{g}_l^{MLE})_l$ are independent and each

$$\hat{g}_l^{MLE} := \int_{J_l} \hat{g}_l^{MLE}(x) w(x) dx - \frac{1}{n} \sum_{j \geq 1} 1(X_j \in J_l, \hat{g}_l^{MLE}(X_j) = Y_j) w(X_j).$$

23
enjoys the non-asymptotic properties of Theorem 3.2 on \( J_t \), in particular \( \mathbb{E}[\hat{\theta}_t^{MLE}] = \int J_t g(x) w(x) dx \) and \( \text{Var}(\hat{\theta}_t^{MLE}) = \frac{1}{n} \int J_t \mathbb{E}[g_t^{MLE}(x) - g(x)] w(x)^2 dx \). Let us therefore first establish for the blockwise MLE \( \hat{\theta}_n := \sum_{l=0}^{r_n - 1} \hat{\theta}_l^{MLE} \) that

\[
\left( \frac{1}{n} \sum_{l=0}^{r_n - 1} \int J_t \mathbb{E}[g_t^{MLE}(x) - g(x)] w(x)^2 dx \right)^{-1/2} (\hat{\theta}_n - \theta) \Rightarrow N(0, 1). \tag{5.1}
\]

By independence of (\( \hat{\theta}_l^{MLE} \)), for the CLT to hold it suffices to check the 4th moment Lyapunov condition

\[
\sum_{l=0}^{r_n - 1} \frac{\mathbb{E}[(\hat{\theta}_l - \theta)^4]}{\text{Var}(\hat{\theta}_n)^2} \to 0.
\]

For each \( l = 0, \ldots, r_n - 1 \) let \((\hat{M}_{l,t})_t\) be the compensated weighted counting process from the proof of Theorem 3.2 restricted to \( J_t \). The (non-predictable) quadratic variation of \((\hat{M}_{l,t})_t\) is given by the sum of squared jumps:

\[
[\hat{M}_l]_t = \sum_{s \leq t} (\Delta \hat{M}_{l,s})^2 = \sum_{j \geq 1} 1 \left( X_j \in J_t, Y_j \leq t \wedge \min_{i: X_i \in J_t} (Y_i + R_i |X_i - X_j|^\beta) \right) w(X_j)^2.
\]

The Burkholder-Davis-Gundy inequality (e.g. Thm 26.12 in Kallenberg (2002)) then yields by similar arguments as for \((\hat{M})_t\) above

\[
\mathbb{E}[\hat{M}_{l,\infty}^4] \lesssim \mathbb{E}[[\hat{M}_{l,\infty}]^2] = \mathbb{E} \left[ \left( n \int_{J_t} (\hat{g}_t^{MLE} - g) w^2 \right)^2 + n \int_{J_t} (\hat{g}_t^{MLE} - g) w^4 \right].
\]

Using Jensen’s inequality, we find

\[
\sum_{l=0}^{r_n - 1} \mathbb{E}[(\hat{\theta}_l^{MLE} - \theta_l)^4] = \frac{1}{n^4} \sum_{l=0}^{r_n - 1} \mathbb{E}[\hat{M}_{l,\infty}^4] \lesssim \sum_{l=0}^{r_n - 1} \mathbb{E} \left[ n^{-2} \left( \int_{J_t} (\hat{g}_t^{MLE} - g) w^2 \right)^2 + n^{-3} \int_{J_t} (\hat{g}_t^{MLE} - g) w^4 \right] \lesssim \sum_{l=0}^{r_n - 1} \left( n^{-2} r_n \int_{J_t} \mathbb{E}[(\hat{g}_t^{MLE} - g)^2] w^4 + n^{-3} \int_{J_t} \mathbb{E}[\hat{g}_t^{MLE} - g] w^4 \right).
\]

As in \(3.3\) we can bound

\[
P(\hat{g}_t^{MLE}(x) - g(x) \geq s) \leq \begin{cases} 
\exp(-n^{2R/\beta + 1} (s/2R)^{(\beta + 1)/\beta}), & s \in [0, 2Rr_n^\beta], \\
\exp(-n(s r_n - 2Rr_n^{\beta + 1}/(\beta + 1))), & s > 2Rr_n^\beta.
\end{cases} \tag{5.2}
\]

Noting \( r_n^\beta/(\beta + 1) \to \infty \) and \( \|w\|_\infty < \infty \), we apply the moment bound of Lemma 5.3 to \( \hat{g}_t^{MLE}(x) - g(x) \) with \( a(s) = n \int_{J_t} (s - R|\xi - x|^\beta + g(x) - g(\xi))_+ dx \) and integrate over \( s \) to obtain

\[
\sum_{l=0}^{r_n - 1} \mathbb{E}[(\hat{\theta}_l^{MLE} - \theta_l)^4] \lesssim (r_n + n^{-1/(\beta + 1)}) \left( n^{-(2\beta + 1)/(\beta + 1)} \right)^2.
\]
Hence, in view of \( \text{Var}(\hat{\phi}_n) \geq \text{Var}(\hat{\phi}_{n}^{\text{MLE}}) \sim n^{-(2\beta+1)/(\beta+1)} \) the Lyapunov condition is satisfied and the CLT (5.1) follows.

In the second step we show that the difference between \( \hat{\phi}_n \) and \( \hat{\phi}_{n}^{\text{MLE}} \) is of small stochastic order \( o_p(n^{-(\beta+1)/2}) \). First, we note that the above martingale arguments yield

\[
\mathbb{E}[(\hat{\phi}_n - \hat{\phi}_{n}^{\text{MLE}})^2] = \text{Var}(\hat{\phi}_n - \hat{\phi}_{n}^{\text{MLE}}) = n^{-1} \sum_{l=0}^{r_n-1} \int_{J_l} \mathbb{E}[\hat{g}_{l}^{\text{MLE}}(x) - \hat{g}_{l}^{\text{MLE}}(x)] w(x)^2 \, dx.
\]

Introduce the notation \( \hat{g}_{l}^{\text{MLE}}(x) = \min_{i : \mathcal{X}_i \notin J_l} (Y_i + R|X_i|^{\beta}) \) and consider the event

\[
\Omega_n = \left\{ \forall l = 0, \ldots, r_n - 1 \exists x \in J_l : \hat{g}_{l}^{\text{MLE}}(x) = g^{\text{MLE}}(x) \right\}
\]

whose complement is given by \( \Omega_n^c = \bigcup \{ \min_{x \in J_l} (\hat{g}_{l}^{\text{MLE}} - \hat{g}_{l}^{\text{MLE}})(x) > 0 \} \). By independence of \( \hat{g}_{l}^{\text{MLE}} \) and \( \hat{g}_{l}^{\text{MLE}} \) and conditioning on the latter we obtain

\[
P\left( \min_{x \in J_l} (\hat{g}_{l}^{\text{MLE}} - \hat{g}_{l}^{\text{MLE}})(x) > 0 \right) = \mathbb{E} \left[ \exp \left( - n \int_{J_l} (\hat{g}_{l}^{\text{MLE}} - g)(x) \, dx \right) \right] \leq \mathbb{E} \left[ \exp \left( - nr_n \min \left( (\hat{g}_{l}^{\text{MLE}} - g)(lr_n), (\hat{g}_{l}^{\text{MLE}} - g)(l + 1) r_n) \right) \right] \].

Using \( \hat{g}_{l}^{\text{MLE}} \geq \hat{g}_{l}^{\text{MLE}} \) for \( l' \neq l \), the bound (5.2) yields

\[
P\left( \min_{x \in J_l} (\hat{g}_{l}^{\text{MLE}} - \hat{g}_{l}^{\text{MLE}})(x) > 0 \right) \leq 2 \max_{l', x} \mathbb{E} \left[ \exp(-nr_n(\hat{g}_{l}^{\text{MLE}} - g(x))) \right] \lesssim n^{-1/\beta} r_n^{-(\beta+1)/\beta}.
\]

We conclude \( P(\Omega_n^c) = O(n^{-1/\beta} r_n^{(2\beta+1)/\beta}) \to 0 \) by a union bound and the choice of \( r_n \).

On the event \( \Omega_n \) the left-most point \( L_l \) in \( J_l \) where \( \hat{g}_{l}^{\text{MLE}} \) and \( \hat{g}_{l}^{\text{MLE}} \) coincide is well defined and satisfies for \( l \geq 1 \)

\[
L_l := \inf \{ x \in J_l \mid \hat{g}_{l}^{\text{MLE}}(x) = \hat{g}_{l}^{\text{MLE}}(x) \} = \inf \{ x \in J_l \mid \hat{g}_{l}^{\text{MLE}}(x) \leq \hat{g}_{l-1}^{\text{MLE}}(lr_n) + R(x - lr_n)^{\beta} \}.
\]

Now \( L_l = lr_n \) holds on \( \Omega_n \) if the corresponding right-most point \( R_{l-1} := \sup \{ x \in J_{l-1} \mid \hat{g}_{l-1}^{\text{MLE}}(x) = \hat{g}_{l}^{\text{MLE}}(x) \} \) on \( J_{l-1} \) satisfies \( R_{l-1} < lr_n \) and vice versa \( L_l > lr_n \Rightarrow R_{l-1} = lr_n \). Due to this symmetry we only consider the case \( L_l > lr_n \). For \( z \in (0, r_n] \) and \( l = 1, \ldots, r_n - 1 \) a rough bound yields:

\[
P\left( L_l \geq lr_n + z \mid (X_i, Y_i) \mathbf{1}(X_i < lr_n) \right) \leq \exp \left( - n \int_{lr_n}^{lr_n+z} (\hat{g}_{l-1}^{\text{MLE}}(x) - g(x)) \, dx \right)
\]

\[
\leq \exp \left( - nz(\hat{g}_{l-1}^{\text{MLE}}(lr_n) - g(lr_n)) \right).
\]

Since \( \hat{g}_{l}^{\text{MLE}}(L_l) = \hat{g}_{l-1}^{\text{MLE}}(L_l) \) holds and both functions are in \( \mathcal{O}^{\beta}(R) \), we obtain the bound

\[
\int_{lr_n}^{L_l} (\hat{g}_{l}^{\text{MLE}} - \hat{g}_{l-1}^{\text{MLE}})(x) \, dx \leq \int_{lr_n}^{L_l} 2R(L_l - x)^{\beta} \, dx \lesssim (L_l - lr_n)^{\beta+1}.
\]
By the identity $E[Z^{\beta+1}] = \int_0^\infty (\beta+1) z^\beta P(Z \geq z)dz$ for non-negative random variables $Z$ and by the above probability bound for $L_l$ we obtain further

$$
E \left[ \int_{lr_n}^{L_l} (\hat{g}_l^{MLE} - \hat{g}^{MLE}) \mathbf{1}_{\Omega_n} \right] \lesssim E \left[ \int_0^{r_n} z^\beta e^{-nz(\hat{g}_{l-1}^{MLE}(lr_n) - g(lr_n))} dz \right] 
\lesssim E \left[ \min \left( \frac{1}{n(\hat{g}_{l-1}^{MLE}(lr_n) - g(lr_n))}, r_n \right)^{\beta+1} \right].
$$

Using (5.2) on $J_{l-1}$, we arrive, after suitable substitution inside the integral, at

$$
E \left[ \int_{lr_n}^{L_l} (\hat{g}_l^{MLE} - \hat{g}^{MLE}) \mathbf{1}_{\Omega_n} \right] \lesssim \int_0^\infty \min(u^{-\beta} n^{-1}, r_n^{\beta+1}) e^{-u} du
$$
\lesssim \begin{cases} n^{-1}, & \beta < 1, \\ n^{-1} \log(nr_n^2), & \beta = 1. \end{cases}
$$

Summing over $l$, bounding the alternative case $R_{l-1} < lr_n$ by the same estimate and using $\|w\|_\infty < \infty$, we arrive at

$$
E \left[ \sum_{l=0}^{r_n-1} \int_{J_l} (\hat{g}_l^{MLE} - \hat{g}^{MLE})(x)w(x)^2 dx \mathbf{1}_{\Omega_n} \right] \lesssim (nr_n)^{-1} \log(nr_n^2) = o(n^{-\beta/(\beta+1)}).
$$

This gives the desired result $E[(\tilde{\nu}_n - \hat{\nu}^{MLE})^2 \mathbf{1}_{\Omega_n}] = o(n^{-(2\beta+1)/(\beta+1)})$ with $P(\Omega_n) \rightarrow 1$.

Furthermore, from (5.3) we derive also that

$$
\sum_{l=0}^{r_n-1} \int_{J_l} E[\hat{g}_l^{MLE}(x) - g(x)]w(x)^2 dx - \int_0^1 E[\hat{g}^{MLE}(x) - g(x)]w(x)^2 dx = o(n^{-\beta/(\beta+1)}) + O(\sup_{l,x \in J_l} E[(\hat{g}_l^{MLE}(x) - \hat{g}^{MLE}(x)) \mathbf{1}_{\Omega_n}]).
$$

By the Cauchy-Schwarz inequality, the last term is at most of order $O(n^{-\beta/(\beta+1)} P(\Omega_n^{1/2}) = o(n^{-\beta/(\beta+1)})$. Hence, applying Slutsky’s Lemma twice to the CLT (5.1), we arrive at

$$
\left( \frac{1}{n} \int_0^1 E[\hat{g}^{MLE}(x) - g(x)]w(x)^2 dx \right)^{-1/2} (\hat{\nu}_n^{MLE} - \tilde{\nu}) \Rightarrow N(0, 1).
$$

By Jensen’s inequality, $\text{Var}(\int_{J_l} (\hat{g}_l^{MLE} - g)w^2) \leq r_n \int_{J_l} E[(\hat{g}_l^{MLE} - g)^2]w^4 \leq r_n^2 n^{-2\beta/(2\beta+1)}$, which implies

$$
\text{Var}\left( \sum_{l=0}^{r_n-1} \int_{J_l} (\hat{g}_l^{MLE} - g(x))w(x)^2 dx \right) \leq r_n n^{-2\beta/(2\beta+1)} = o(n^{-2\beta/(\beta+1)}).
$$

We infer $\sum_{l=0}^{r_n-1} \int_{J_l} (\hat{g}_l^{MLE} - g)w^2 \overset{P}{\rightarrow} \sum_{l=0}^{r_n-1} \int_{J_l} E[\hat{g}_l^{MLE} - g]w^2$. Together with (5.3), this yields the CLT

$$
\left( \int_0^1 (\hat{g}^{MLE}(x) - g(x))w(x)^2 dx \right)^{-1/2} n^{-1/2} (\hat{\nu}_n^{MLE} - \tilde{\nu}) \Rightarrow N(0, 1).
$$
Now note that the MLE for the functional $\int gw^2$

$$\hat{\vartheta}_{MLE}^n(w^2) := \int_0^1 \hat{g}_{MLE}(x)w(x)^2 \, dx - \frac{1}{n} \sum_{j \geq 1} 1(\hat{g}_{MLE}(X_j) = Y_j)w(X_j)^2$$

is also unbiased with $\text{Var}(\hat{\vartheta}_{MLE}^n(w^2))^{1/2} \lesssim n^{-(\beta+1)/2}/(\beta+1)$. Since by assumption $\int_0^1 E[\hat{g}_{MLE} - g]w^2 \sim n^{-\beta/(\beta+1)}$ is of larger order, Slutsky’s Lemma permits to replace $\int_0^1 E[\hat{g}_{MLE} - g]w^2$ by $\int_0^1 \hat{g}_{MLE}w^2 - \hat{\vartheta}_{MLE}^n(w^2)$ in the CLT, which gives the desired self-normalising form.

References

BARAUD, Y., AND L. BIRgé (2014): “A new method for estimation and model selection: $\rho$-estimation,” Preprint, Université Nice Sophia Antipolis.

GAYRAUD, G. (1997): “Estimation of functionals of density support,” Mathematical Methods of Statistics, 6(1), 26–46.

GROENEBOOM, P. (2012): “Convex hulls of uniform samples from a convex polygon,” Advances in Applied Probability, 44(2), 330–342.

GROENEBOOM, P., AND J. A. WELLNER (1992): Information bounds and nonparametric maximum likelihood estimation. Oberwolfach Seminars 19, Birkhäuser/Springer.

JACOD, J., AND A. N. SHIRYAEV (1987): Limit theorems for stochastic processes, vol. Grundlehren 288. Springer.

JIRAK, M., A. MEISTER, AND M. REISS (2014): “Adaptive estimation in nonparametric regression with one-sided errors,” The Annals of Statistics, to appear.

KALLENBERG, O. (2002): Foundations of modern probability. Probability and its Applications, Springer.

KARR, A. (1991): Point processes and their statistical inference. Probability Pure and Applied, CRC press.

KOROSTELEV, A. P., L. SIMAR, AND A. B. TSYBAKOV (1995): “Efficient estimation of monotone boundaries,” The Annals of Statistics, 23(2), 476–489.

KOROSTELEV, A. P., AND A. B. TSYBAKOV (1993): Minimax theory of image reconstruction, Lecture Notes in Statistics. Springer, New York.

KUTOYANTS, Y. A. (1998): Statistical inference for spatial Poisson processes. Lecture Notes in Statistics, Springer.

LEHMANN, E. L., AND J. P. ROMANO (2006): Testing statistical hypotheses. Springer.
Mammen, E., and A. B. Tsybakov (1995): “Asymptotical minimax recovery of sets with smooth boundaries,” *The Annals of Statistics*, 23(2), 502–524.

Meister, A., and M. Reiß (2013): “Asymptotic equivalence for nonparametric regression with non-regular errors,” *Probability Theory and Related Fields*, 155(1-2), 201–229.

Rényi, A., and R. Sulanke (1964): “Über die konvexe Hülle von n zufällig gewählten Punkten. II,” *Probability Theory and Related Fields*, 3(2), 138–147.

Van Der Vaart, A. W., and J. A. Wellner (1996): *Weak Convergence*. Springer Series in Statistics.

Wojtaszczyk, P. (1997): *A mathematical introduction to wavelets*. LMS Student Texts 37, Cambridge University Press.