Efficient estimation of present-value distributions for long-dated contracts and functionals in the multivariate case

Alexandra Tsimbalyuk

Department of Statistics
London School of Economics and Political Science

This dissertation is submitted for the degree of Doctor of Philosophy

Abstract

The first chapter of this thesis focuses on the problem of estimating the joint law of a discrete-time perpetuity and underlying factors which govern the cash flow rate, in an ergodic Markovian environment. Our approach is based upon the so-called time-reversal technique which allows us to identify the joint law as a stationary distribution of an ergodic multidimensional Markov chain. Furthermore, a central limit theorem (CLT) for an estimator of the joint law is provided for a specific example of the perpetuity. Our proof of the CLT rests upon the geometric ergodicity property, which is also provided and is of independent interest. We further provide a justification for the Monte Carlo methods for approximating the joint law by sampling a single path of the reversed process.

The second chapter of this thesis deals with the estimation of linear functionals in multidimensional spaces. We consider two ubiquitous statistical models: a regression model with one-sided errors and a Poisson point process (PPP) model. We consider two estimation approaches: a block-wise approach, when the estimator is an aggregate of local estimators, and a maximum likelihood approach. First, we assume the regularity of the underlying function in both models to be known. We combine the block-wise approach with martingale stopping time arguments and the PPP geometry to derive the unbiased estimators. We show that the rates of convergence of the mean squared risks match the lower bounds for the risks in both models, which are also provided and are of independent interest. In the PPP model, we show that the maximum likelihood estimator is unbiased with minimal variance among all unbiased estimators. Finally, we sketch ideas for a proof of the CLT for the estimator in the PPP model in multidimensional case and provide illustrative simulations.

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it). The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without my prior written consent. I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party. I declare that my thesis consists of less than 100,000 words.

I confirm that Chapter 1 was jointly co-authored with Professor Konstantinos Kardaras and I contributed 80% of this work. Chapter 2 is a result of my independent work. Chapter 1 and 2 constitutes part of work I intend to submit for publication in the near future.

Alexandra Tsimbalyuk
December 2020

Acknowledgements

I am all too aware that the completion of my thesis would not have been possible were it not for the great support of so many. I would like to express my deepest gratitude to all who have contributed to making this journey a unique, meaningful and life-changing experience for me. Firstly, I wish to thank my supervisor — Kostas Kardaras, for all the time he has given. He not only introduced me to the fascinating world of research but his passion for the subject has enthused me; It was a huge honour to have the opportunity to learn from his immense mathematical knowledge. I wish also to thank Umut Cetin for sharing his exceptional insights into some challenging probability open problems, as well as for his dedication and personal support. I would like to express my gratitude to Angelos Dassios and LSE' directors for putting their trust in me when they accepted me onto what has been a fantastic PhD programme. I also want to thank Penny Montague a lot for all the support that I received from her throughout my studies at LSE. Viet Dang, my sister, thank you for all the inspiration and that dragon energy.

I am very grateful for all the guidance on my work I received from Albert Shiryaev and all the fruitful conversations we had at different conferences.

I would like to thank Yuri Kabanov for giving me the opportunity to present the results of Chapter 1 at the Bachelier Colloquium on Mathematical Finance and Stochastic Calculus in Metabief, that was a very fruitful experience for me.

Special thanks go to my wonderful sister, Nastya, for always being there for me and mainly for reminding me what is important in life and what is not.

Contents

List of Notation

$\mathbb{R}$	The set of real numbers
$\mathbb{R}^d$	A d-dimensional Euclidian space
$\mathbb{R}_+$	The set of positive real numbers
A^c	The complement of $A \subseteq \mathbb{R}$
$\lvert A \rvert$	Lebesgue measure of Lebesgue measurable set A
x_+	The maximum of 0 and x
$\lvert I \rvert$	The number of elements in a finite set I
$\lVert \bullet \rVert_{TV}$	The total variation distance
$\lVert\ x\ \rVert_p$	l_p-norm of the vector x
n	Size of a sample
$L^2([0,1]^d)$	$\int_{[0,1]^d} f(x)^2 dx < \infty$, for $f(x) \in L^2([0,1]^d)$
i.i.d.	Independent and identically distributed
$\mathcal{L}$, log	The distinguished and the natural logarithms
$\xrightarrow{d}$	Classical convergence in distribution
$\overset{d}{=}$	Equality in distribution
$A \subseteq B$	A set A is a subset of a set B
$S \times I$	The product space of two spaces S and I
$\mathcal{S} \otimes \mathcal{B}(\mathbb{R})$	The product of two sigma algebras $\mathcal{S}$ and $\mathcal{B}(\mathbb{R})$
$\mathbb{I}(\bullet)$	The indicator function
$\langle \pi, g \rangle$	The integral $\int g \mathrm{d}\pi$ for a measurable function g and a probability measure π
$\max, \vee$	Maximum
$\min, \wedge$	Minimum
$\equiv$	Equality
$a_n \asymp b_n$	$a_n / b_n \to 1$
$f(x) = \mathcal{O}(g(x))$	$f(x)$ is bounded by a constant multiple of $g(x)$
$f(x) \lesssim g(x)$	$f(x)$ is bounded by a constant multiple of $g(x)$
$f(x) \sim g(x)$	if $f(x) \lesssim g(x)$ as well as $g(x) \lesssim f(x)$

Chapter 1

Discrete-time perpetuities and time reversal

1.1 Introduction to estimation of the distribution of discrete-time perpetuities

This chapter studies the main properties of a random variable, the so-called *perpetuity* or infinite horizon stochastic discounted rewards of the following form

$$X_0 = \sum_{k=0}^{\infty} \left(\prod_{0<\ell<k} b(Z_\ell) \right) a(Z_k), \tag{1.1.1}$$

where a and b are well-defined functions which will be specified further and $(Z_n)_{n\in\mathbb{Z}}$ is a stationary and ergodic discrete-time process which takes values on some measurable space $(S, \mathcal{S})$.

In many applications of operations and economics research, it is of the main interest to identify the expected value of X_0. There is also a range of research areas where it is essential to identify the entire distribution of a random variable X_0 as it plays a major role in the areas of financial and insurance mathematics, see [7, 10, 22, 24, 54, 69] for a review on possible applications. One such example is the infinite horizon net present value of an investment strategy, see [45, 46] and references therein for a more detailed presentation. Here we just touch upon this example to give more flavour for the motivation of studying the distributional properties of perpetuities. If we consider the rates of return defined as a sequence $(r_k)_{k\geq 1}$, which are constants on the corresponding time intervals $[k-1, k)$, as well as a sequence of random variables $(C_k)_{k\geq 1}$ which represents a sequence of cash payments paid at the beginning of the corresponding time interval $[k, k+1)$. Further, the initial wealth will be denoted as S_0 and the sequence of the accumulated wealth as $(S_k)_{k\geq 1}$. The present value of cash flows up to the time k is referred to as $(X_k)_{k\geq 1}$. Then one can

write down the recursive equation for S_k, $k \geq 1$ in the following way

$$S_k = (1 + r_k)S_{k-1} + C_k.$$

Oftentimes the equation above refers to as annuity equation. In finance, annuities and cash flow sequences are usually priced using their present value X_k, which is the value of all payments received up to and including the moment k. Then the present value corresponding to the scenario above follows

$$X_k = S_k \prod_{i=1}^{k} \frac{1}{1 + r_k},$$

which leads to the recursive equation

$$X_k = X_{k-1} + C_k \prod_{i=1}^{k} \frac{1}{1 + r_i}.$$

Starting from $X_0 = S_0$ we obtain

$$X_k = S_0 + \sum_{j=1}^{k} C_j \prod_{i=1}^{j} \frac{1}{1 + r_i}.$$

We thus can arrive to a perpetuity by taking a limit of the present value of the corresponding annuity. When starting from zero, i.e. $S_0 = 0$ and placing $D_k = 1/(1 + r_k)$, for $k \geq 1$ the perpetuity is then equal to

$$X_\infty = \sum_{j=1}^{\infty} C_j \prod_{i=1}^{j} D_j. \tag{1.1.2}$$

Consider the following two examples where the distribution of X_∞ can be inferred explicitly. For the two examples below we assume that the sequences $(C_k)_{k\geq 1}$ and $(D_k)_{k\geq 1}$ are formed from i.i.d. random variables as well as mutually independent.

Example 1.1.1. If C is arbitrary and $D \sim Ber(p)$, $q = 1 - p > 0$. This example illustrates the situation, when after some time, there is a probability that all the payments will stop. Then, the present value for a given k can be written as

$$X_k = \begin{cases} \sum_{j=1}^{m} C_j & \text{with probability } p^m q, \text{ for } 0 \leq m \leq k-1 \\ \sum_{j=1}^{k} C_j & \text{with probability } p^k \quad \text{otherwise.} \end{cases}$$

The characteristic function for X_k is

$$\begin{aligned}\mathbb{E}(e^{itX_k}) &= \sum_{m=0}^{k-1} \left[qp^m \mathbb{E}\left(e^{it\sum_{j=1}^{m} C_j}\right)\right] + p^k \mathbb{E}\left(e^{it\sum_{j=1}^{k} C_j}\right) \\ &= q\sum_{m=0}^{k-1} [p^m \psi(t)^m] + p^k \psi(t)^k = q\frac{1-p^k\psi(t)^k}{1-p\psi(t)} + p^k\psi(t)^k,\end{aligned}$$

where $\psi(t)$ is the characteristic function for C. Since for every t from $\mathbb{R}$ it holds $|\psi(t)| \leq 1$, it follows that $p^k\psi(t)^k \to 0$ as $k \to \infty$. From where we can conclude that

$$\mathbb{E}(e^{itX_\infty}) = \frac{q}{1-p\psi(t)},$$

which corresponds to the compound geometric distribution.

Example 1.1.2. Let's now assume that the discounting factor $D = d \in (0,1)$ is a constant. In case C is finite a.s. the resulted perpetuity is convergent. Consider $C \sim \mathcal{N}(\mu, \sigma^2)$ then

$$\begin{aligned}\mathbb{E}(e^{itX_k}) &= E\left(e^{it\sum_{j=1}^{k} d^j C_j}\right) = \prod_{j=1}^{k} \psi(td^j) \\ &= \exp\left\{it\mu\sum_{j=1}^{k} d^j - \frac{t^2}{2}\sigma^2\sum_{j=1}^{k} d^{2j}\right\}\end{aligned}$$

Putting $k \to \infty$ one can get

$$\mathbb{E}(e^{itX_\infty}) = exp\left\{it\frac{d\mu}{1-d} - \frac{t^2}{2}\sigma^2\frac{d^2\sigma^2}{1-d^2}\right\}.$$

The characteristic function above represents a normal distribution, namely, $\mathcal{N}(\frac{d\mu}{1-d}, \frac{d^2\sigma^2}{1-d^2})$.

Identifying the distribution of a perpetuity with non-restrictive assumptions is an involved task and, often, it is impossible to provide its explicit form. At the same time, all the known results in the literature for approximating the distribution are often slow. Numerical methods based on partial differential equations fail because of the lack of information about boundary conditions. Monte-Carlo-based simulations typically take a prohibitive amount of time. We propose an alternative simulation method, using ergodicity and time-reversal, which leads to significantly better results; in effect, reducing the entire simulation to sampling a single path, thus, allowing an efficient use of Monte Carlo simulation.

1.1.1 Former results

1. As it was shown in [27], the distribution of perpetuities arises as a key factor in risk management theory as it is used to calculate rates for estimating the pension fund stability so that it is managed in a balanced way with respect to its actuarial liabilities. For further details on the pension funding theory, refer to [9].

2. The distribution of perpetuities is widely studied in non-pension fund insurance settings as well. Consider the following ruin model from risk insurance theory, first studied in [53], as an illustration and for further stimulation of our research interest. When working with insurance products, it is of high importance to be able to answer the question as to what the probability of collapse is. The usual approach is to study the distribution of the difference between the present value of the aggregated benefits and the present value of the aggregated premiums received. Consider an insurance risk model with the aggregated paid claim process A_t at time t, modelled by a compound Poisson process. If the risk reserve of a company R_t are being invested in at a rate r, then the discounted risk reserve satisfies the following equation

$$R_t = R_0 + c\int_0^t e^{-ru}du - \int_0^t e^{-ru}dA_u,$$

where $c > 0$ is a constant and represents the rate at which the premiums are accumulated. It was shown in [53] that the ruin probability, namely $\mathbb{P}(\inf_{t>0} R_t < 0)$, can be calculated in terms of the distribution of the following random variable

$$Y_0 = \int_0^\infty e^{-N_u}dF_u, \tag{1.1.3}$$

where $N_t = rt$ is the aggregated discount at time t and $F_t = ct - A_t$ is the value of the surplus process at time t. In a more realistic framework, it would be natural to assume that both the discounting rate r and the reward rate h are governed by an ergodic continuous-time Markov chain $(M_t)_{t\in\mathbb{R}_+}$ with the finite state space I. It was shown in [92, 94] that the key to calculating the ruin probabilities in a more general framework is computing the distribution of the following perpetuity

$$Y_0 = \int_0^\infty \exp\left(-\int_0^t r(M_{u-})du\right) h(M_t)dt. \tag{1.1.4}$$

3. The distribution of perpetuities plays a fundamental role in the theory of differential equations as well. As an example, consider the following ordinary differential equation

$$dY_t = \alpha_t Y_t dt + \lambda_t dt, \tag{1.1.5}$$

where $\alpha_t, \lambda_t \in \mathbb{R}$. When $Y_0 = 0$, the solution to (1.1.5) is given by

$$\begin{aligned} Y_t &= \int_0^t \exp\Big(\int_u^t \alpha_s ds\Big)\lambda_u du \\ &= \int_{-t}^0 \exp\Big(\int_v^0 \alpha_{r+t} dr\Big)\lambda_{v+t} dv. \end{aligned}$$

If $((\lambda_t, \alpha_t)) : t \in (-\infty, \infty)$ is strictly stationary with $\alpha_t < 0$ and $\lambda_t > 0$, then the following result holds

$$\begin{aligned} Y_t &\stackrel{d}{=} \int_{-t}^0 \exp\Big(\int_v^0 \alpha_r dr\Big)\lambda_v dv \\ &\to \int_{-\infty}^0 \exp\Big(\int_v^0 \alpha_s ds\Big)\lambda_v dv \stackrel{d}{=} Y_\infty, \end{aligned}$$

as $t \to \infty$. When $\alpha_t = -r_t$ and $\lambda_t = h_t$ the form of Y_∞ corresponds to the random variable X_0. This implies that in order to calculate the equilibrium distribution for the solution of the differential equation in (1.1.5), it suffices to compute the distribution of the perpetuity.

4. Another interesting application of the perpetuity is originated from a study of the running time of the so-called *Quickselect algorithm*, see [55, 58].

   Quickselect algorithm (Quickselect(n, m)) is a recursive algorithm to find the item of a given rank $m \geq 1$ in a given array with $n \geq m$ distinct numbers, these numbers are usually called "keys". This algorithm was invented by Hoare, refer to [55], and is based on partitioning of the array into two subarrays around a pivot element. It is one of the most uninvolved and efficient algorithms in practice for finding a specified order statistics for a given sequence. As a first step, we select the pivot element, which is chosen at random uniformly from the array of elements. Then every key is compared to the pivot and the rank of the pivot, i.e. i. If a randomly selected pivot element happened to have the rank $i = m$, then (Quickselect(n, m)) will return the pivot. If, however, $i > m$, then we apply (Quickselect$(i - 1, m)$) on a smaller array of the elements smaller than the pivot. Alternatively, if $i < m$ then (Quickselect$(n - i, m - i)$) is applied. The properties of the fundamental quantities for (Quickselect(n, m)), such as, the number of comparisons between data points and the number of recursive calls of the algorithm has been extensively studied in the literature, see, for example [50, 68, 70, 85] and references therein. One of the property of interest is the distribution of the number of key comparisons that are required by the call of Quickselect(n, m). This distribution is not available in closed form for general finite values n and m. If we denote the number of key comparisons as $C(n, m)$ and the rank of the first pivot that was chosen as Z_n, then we can write

down the following recursive equation

$$C(n,m) \stackrel{d}{=} (n-1) + \mathbb{I}_{Z_n>m} C(Z_n - 1, m) + \mathbb{I}_{Z_n<m} C^*(n - Z_n, m - Z_n),$$

where $C(n,m)$ is a random variable which is for every fixed n and m is distributed as the number of comparisons needed for Quickselect(n,m) and $C^*(n,m)$ is an independent copy of $C(n,m)$. The random variable Z_n follows a uniform distribution on the set of numbers from 1 to n and is independent of $C(n,m)$ and $C^*(n,m)$ as well as $C(n,m)$ is independent of $C^*(n,m)$. When turning into asymptotics where m is fixed and $n \to \infty$ one can show that the random variable $Y(m,n)$, which is defined as follows

$$Y(m,n) := \frac{C(n,m)}{n} - 1,$$

has a limiting distribution that satisfies the following fixed-point equation

$$Y \stackrel{d}{=} U(1+Y), \tag{1.1.6}$$

where Y and U are two independent random variables and U is uniformly distributed on $[0,1]$, see [85] for the result. If now we introduce the random variable Y as

$$Y = U_1 + U_1U_2 + U_1U_2U_3 + \dots \tag{1.1.7}$$

for some sequence $U_1, U_2, U_3, \dots$ of independent and identically distributed random variables such that $U \sim U([0,1])$. Then, Y is also referred to as a perpetuity. Under some non-restrictive conditions it can be shown that the perpetuity following (1.1.7) satisfies the distributional fixed-point equation of the form (1.1.6) as well. The law of Y is known as the Dickman distribution in the literature, see [25, 32] for further detail and proofs of the results above. There is a number of different application settings including number theory with topics related to the largest prime factors, as well as combinatorics, the theory on the longest cycles in permutations, in which the Dickman distribution naturally arises, see [58]. Another special case of (1.1.7) is when the sequence $U_1, U_2, U_3, \dots$ of random variables is independent and identically distributed as $W = U^{1/\beta}$ with $U \sim U([0,1])$ and $\beta > 0$. This special case of perpetuity is called Vervaat perpetuity and it has been extensively studied in the literature. In [112] it was shown that Vervaat perpetuity is infinitely divisible, and obtained the Lévy Khintchine representation.

The distributional properties of different types of perpetuities have been extensively studied in the literature. The method of identifying the distribution of perpetuities from the

fixed point equation of the type (1.1.6) is one of such widely studied approach, see [27, 112]. The following examples illustrate the situation where the distribution of a perpetuity can be identified by using fixed point equations. The type of perpetuities considered below represents the class of Vervaat perpetuities.

Example 1.1.3. Consider again the perpetuity of the type 1.1.2 and assume that the discounting factor D follows the exponential function with the rate of return defined the same way as in the previous section as r. This rate of return is assumed to be exponentially distributed with the parameter $\lambda > 0$. Then we have $D = \exp(-r) \stackrel{d}{=} U^{1/\lambda}$, where $U \sim U([0,1])$ and the corresponding density function

$$f_D(x) = \lambda x^{\lambda-1} \mathbb{I}_{[0,1]}(x).$$

Under some appropriate assumptions, refer to [11, 112], one can show that for the corresponding perpetuity the fixed point equation holds

$$X \stackrel{d}{=} D(X + C).$$

As a result

$$\begin{aligned}\mathbb{E}\big(e^{itX}\big) = \mathbb{E}\big(e^{itD(X+C)}\big) &= \int_{-\infty}^{\infty} f_D(x)\mathbb{E}\big(e^{itx(X+C)}\big)dx \\ &= \int_0^1 \lambda x^{\lambda-1} E\big(e^{itxX}\big)E\big(e^{itxC}\big)dx = \int_0^1 \lambda x^{\lambda-1}\varphi(tx)\psi(tx)dx,\end{aligned}$$

where $\varphi(t)$ is the characteristic function of X. By multiplying by t^λ the following equation

$$\varphi(t) = \lambda t^{-\lambda} \int_0^t u^{\lambda-1}\varphi(u)\psi(u)du,$$

we arrive to the differential equation of the form

$$t^\lambda \varphi' + \lambda t^{\lambda-1}\varphi = \lambda t^{\lambda-1}\varphi\psi,$$

with the corresponding solution

$$\varphi(t) = \exp\left\{\lambda \int_0^t \frac{\psi(u) - 1}{u} du\right\}.$$

There are several well known examples when the integral above can be calculated. One of such examples of Vervaat perpetuity is when C follows an exponential distribution with the parameter α. Then, $\psi(t) = \alpha(\alpha - it)^{-1}$ and as a consequence $\varphi(t) = \left(\frac{\alpha}{\alpha - it}\right)^\lambda$, which corresponds to the gamma distribution, i.e. $\Gamma(\lambda, \alpha)$. If we assume that C is distributed

according to $Laplace(0, \alpha)$ then this corresponds to the situation when the payments' sizes are exponentially distributed, however we are not aware of whether we will receive a payment or not. Then $\psi(t) = \alpha^2(\alpha^2 + t^2)^{-1}$ and the corresponding characteristic function for X_∞ is given by

$$\varphi(t) = \Big(\frac{\alpha^2}{\alpha^2 + t^2}\Big)^{\lambda/2},$$

which serves as a characteristic function for Variance-gamma distribution. Let's now consider a slightly more involved case of Vervaat perpetuity. Namely, we assume that $C = ZW - (1 - Z)V$, where $Z \sim Ber\big(b/(a + b)\big)$, $W \sim Exp(a)$, $V \sim Exp(b)$, as well as W and V are independent and Z is independent of both W and V. In this set up, we can easily calculate the characteristic function of C

$$\psi(t) = \mathbb{E}\Big(e^{it(ZW-(1-Z)V)}\Big) = \frac{b}{a+b}\frac{a}{a-it} + \frac{a}{a+b}\frac{b}{b+it}.$$

From where it follows that

$$\varphi(t) = \Big(\frac{a}{a-it}\Big)^{\lambda b/(a+b)}\Big(\frac{b}{b+it}\Big)^{\lambda a/(a+b)}.$$

If $\Gamma_1 \sim \Gamma(\lambda b/(a + b), a)$ and $\Gamma_2 \sim \Gamma(\lambda a/(a + b), b)$ and Γ_1 is independent from Γ_2 then

$$\varphi(t) = \mathbb{E}\big(e^{it\Gamma_1}\big)\mathbb{E}\big(e^{-it\Gamma_2}\big) = \mathbb{E}\big(e^{it(\Gamma_1-\Gamma_2)}\big),$$

implying that $X \stackrel{d}{=} \Gamma_1 - \Gamma_2$. Even though we did not manage to identify the distribution of the perpetuity explicitly the difference equation is still of high importance as it helps with sampling the distribution of the perpetuity.

There are other methods for identifying solutions to the fixed point distributional equations. One of such methods is based on the properties of Beta and Gamma distributions, see [28–30] and references therein for more details on this method.

The stability of iterations of random linear maps in a discrete Markovian environment has received additional attention in the statistical literature in the last decade. In [1] the authors characterised all feasible limit distributions as solutions to a particular Markovian stochastic fixed-point equation and also furnished necessary and sufficient conditions for the a.s. and convergence in law of the iterative scheme. It was shown in [102] that the distribution tail of the stationary solution to the linear recursion with stationary Markov-dependent coefficients has a power law decay. In [23] based on the renewal theory, the author provided the description of the situation where the tail of stationary solution to the stochastic difference equation exhibits power law behaviour. The local limit theorem and the renewal theorem for partial sums of the stochastic recursion with Markov-dependant

coefficient was proved in [15].

The distribution of Y_0 following (1.1.3) when N stands for a Brownian motion with negative drift and $F_t = t$ was calculated in [96]; a more general result showing that Y_0 follows an inverse gamma distribution was obtained in [27]. An explicit form of the distribution of Y_0 in the case when both N and F follow particular types of Levy processes was provided in [93]. Computing the distribution of a perpetuity in a general setting numerically is often infeasible. We refer to [95] for a survey of possible ways of approximating the distribution of perpetuities using Markov chain Monte Carlo simulations, some important examples where the distribution can be calculated explicitly are also provided in that work.

In this chapter, we are interested in identifying the distribution of (1.1.1) given the initial value of the process $(Z_n)_{n\in\mathbb{Z}}$ is some z_0 from S, or the joint law π of (X_0, Z_0). It is often an involved task and in many application settings it is impossible to find the exact distribution of (X_0, Z_0). We hence propose an alternative simulation method that is based on ergodic theory and time-reversal techniques and motivated by the results from [65], where the authors consider the problem of estimating the joint distribution of a continuous-time perpetuity and the underlying factors which govern the cash flow rate, in an ergodic Markovian model. One of the approach they use is based on techniques of time reversal, which helps them to identify the joint law as the stationary distribution of an ergodic multidimensional diffusion.

1.1.2 Time-reversal

Consider an irreducible stationary discrete-time process $(X_n)_{n\in\mathbb{Z}}$ which takes values on some measurable space $(S, \mathcal{S})$. One can associate with $(X_n)_{n\in\mathbb{Z}}$ the so-called reversed process $(X_{\tau-n})_{n\in\mathbb{Z}}$, where τ is some appropriate random time. If the reversed process and the original process are statistically indistinguishable, one says that the original process is time-reversible (or simply reversible). By observing that for a time-reversible process $(X_n)_{n\in\mathbb{Z}}$, it holds that $(X_0, X_1) \stackrel{d}{=} (X_1, X_0)$, we obtain a necessary condition for a process to be reversible, which is the original process must be stationary.

It turns out that by studying the reversed process, one can get more insight into the properties of the process itself. For instance, the equilibrium of a complex system can often be derived by "guessing" the reversed process. Time reversal of diffusion processes was first considered in [103]. Kolmogorov in [71, 72] computed the transition kernels of the reversed process, and gave necessary and sufficient conditions for symmetry. The basic role of time reversal and duality in potential theory was recognised by Hunt in [56].

Time-reversal can provide some useful results that can be utilised for our research goal. We start with a pivotal example from financial time-series modelling.

1.1.3 ARCH(1) process

A general ARCH(1) process satisfies the following stochastic recursion equation

$$Y_n = A_n + B_n Y_{n-1}, \quad n \in \mathbb{N}, Y_n \in \mathbb{R}\,, \tag{1.1.8}$$

where the sequence $(A_i, B_i)_{i\geq 1}$ are independent and identically distributed random variables. If the sequence $(A_i, B_i)_{i\geq 1}$ is stationary and ergodic, $\mathbb{E}(\log|B_0|) < 0$, and $\mathbb{E}(\log|A_0|_+) < \infty$, then for any initial random value Y_0, the limiting distribution of Y_n is the same as that of the random variable $R = A_0 + \sum_{n=1}^{\infty} A_{-n} \prod_{i=0}^{n-1} B_{-i}$, and it is the unique initial distribution under which $(Y_n)_{n\geqslant 0}$ is stationary, see [11].

This type of processes is widely used to quantify the logarithmic returns on an investment, exchange rates, inflation, and many other financial and economic time series, see [16, 31, 67]. The process (1.1.8) was studied extensively by many authors, see, e.g., [31, 98, 112]. It was shown in [31, 67], that under some stability conditions, the Markov chain $(Y_n)_{n\in\mathbb{N}}$ with dynamics (1.1.8) has a stationary solution with the distribution which corresponds to the distribution of a perpetuity. In [112] solutions to the aforementioned distributional equation (1.1.8) are obtained based upon $\mathbb{E}(\log|B_0|) < 0$ and $\mathbb{E}(\log|A_0|_+) < \infty$. The tails of the stationary solution as well as the convergence of iterative schemes, are studied in [31]; furthermore, in [44] almost if and only if conditions for the convergence of iterative schemes were obtained.

Let us now consider a special case of the perpetuity (1.1.1)

$$X_0 = \sum_{n=1}^{\infty} \big(\prod_{i=1}^{n} D_i\big) C_n,$$

where $(D_n)_{n\in\mathbb{N}}$ and $(C_n)_{n\in\mathbb{N}}$ are both i.i.d. In financial applications, $(D_n)_{n\in\mathbb{N}}$ usually correspond to the discount factors, whereas $(C_n)_{n\in\mathbb{N}}$ correspond to the cash flows. The time-reversal technique allows us to reduce an analysis of the distribution of the perpetuity X_0 to the analysis of the stationary solution to (1.1.8). For a fixed $N \in \mathbb{N}$ define

$$X_0^{(N)} := \sum_{n=1}^{N} \big(\prod_{i=1}^{n} D_i\big) C_n\,.$$

It follows from the i.i.d. property that $X_0^{(N)}$ has the same distribution as the random variable

$$\hat{X}_N := D_N C_N + D_N D_{N-1} C_{N-1} + \ldots + \big(\prod_{j}^{N} D_j\big) C_1 = \sum_{n=1}^{N} \big(\prod_{i=1}^{n} D_{N-n+1}\big) C_{N-n+1}.$$

Then calculations show that the reversed process $(\hat{X}_n)_{n\in\mathbb{N}}$ satisfies the stochastic recursive

equation $\hat{X}_n = D_n\hat{X}_{n-1} + D_nC_n$, which, with a change of variables, corresponds to (1.1.8). Thus, assuming that $(\hat{X}_n)_{n\in\mathbb{N}}$ converges to a random variable $\hat{X}$ in distribution, $\hat{X}$ must solve the distributional equation

$$\hat{X} = D(\hat{X} + C),$$

where D, C and $\hat{X}$ are independent, D has the same distribution as D_1 and C has the same distribution as C_1. The solutions to the aforementioned distributional equation were studied in [112]. The distribution tail of the random variable X_0 is the topic of, for example, [8, 42, 43, 48, 49, 67].

1.1.4 Organisation of the chapter

In Section 1.3 we construct a discrete-time process $(X_n)_{n\in\mathbb{Z}}$ out of a random variable (1.1.1) by taking the sum up to the point $n \in \mathbb{Z}$, and then prove stationarity and ergodicity for the joint process $(Z_n, X_n)_{n\in\mathbb{Z}}$. In Section 1.4 we introduce the time-reversal process $(\zeta_n, \chi_n)_{n\in\mathbb{Z}}$ of the joint process $(Z_n, X_n)_{n\in\mathbb{Z}}$ and identify the dynamics for $(\chi_n)_{n\in\mathbb{Z}}$ in the form of a recursive equation of the type similar to (1.1.8) but with Markov-dependent coefficients. We also show that the joint process $(\zeta_n, \chi_n)_{n\in\mathbb{Z}}$ preserves the property of stationarity and ergodicity implying that the limiting law of $(\zeta_n, \chi_n)_{n\in\mathbb{Z}}$ coincides with the joint law of (Z_0, X_0), which we will refer to as π. We hence consider the following empirical measure as an estimator of π

$$\hat{\pi}_n[A] := \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_A(\zeta_i, \chi_i), \quad A \in \mathcal{S}\otimes\mathcal{B}(\mathbb{R}),\, n \in \mathbb{N}.$$

We further study the stability properties of $(\zeta_n, \chi_n)_{n\in\mathbb{Z}}$, namely, the property of geometric ergodicity which allows us to estimate the rate of convergence of the empirical measure to π, in Section 1.6.2.

In Section 1.7.1 we state the central limit theorem for the process with recursive dynamics of type (1.1.8) in a Markovian environment. We show that a central limit result of the following form holds

$$\sqrt{n}\big(\langle g, \hat{\pi}_n\rangle - \langle g, \pi\rangle\big) \xrightarrow{\mathrm{d}} N(0, \sigma_g^2), \tag{1.1.9}$$

where $\sigma_g^2 < \infty$ is some asymptotic variance to be specified and g is a bounded and $\mathcal{S}\otimes\mathcal{B}(\mathbb{R})$-measurable function. The Markov chain theory about convergence of marginals is connected with the one of central limit theorems, see Section 1.7. The law of large numbers can hold for a Markov chain even though marginal distributions do not converge. In the matter of central limit theorems it is essential to control the rate of convergence

of marginals. There are different approaches to proving the result, and all of them do require some conditions on the rate of convergence. These assumptions allow to reprove a lot of classical CLT results using an intuitive probabilistic approach of the regeneration theory of Markov chains, which helps to avoid many technicalities in the original proofs. Finally, we provide assumptions under which the CLT result of the form (1.1.9) holds, and we demonstrate that the geometric ergodicity of the reversed Markov chain $(\zeta_n, \chi_n)_{n\in\mathbb{N}}$ a sufficient condition for its existence.

1.2 Main contributions

As the main contributions of this chapter, the following results are provided

- An estimator of the joint law of the discrete-time perpetuity (X_0, Z_0) of the form (1.1.1) and its underlying factors based on the time-reversal technique.
- Sufficient conditions for a central limit theorem for a normalised version of the estimator of the joint law.
- An illustrative example demonstrating the computational efficiency of the estimator.

1.3 Probabilistic framework

All random elements below live on a fixed probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Expectations with respect to $\mathbb{P}$ are denoted by $\mathbb{E}$.

Let $Z \equiv (Z_n)_{n\in\mathbb{Z}}$ be a stationary and ergodic process taking values on the measurable space $(S, \mathcal{S})$, and let π^Z denote the stationary probability on $\mathcal{S}$. Furthermore, consider two $\mathcal{S}$-measurable functions $b : S \mapsto (0, \infty)$ and $a : S \mapsto \mathbb{R}$. To state that the random variable (1.1.1) is well-defined, we need the following assumption to hold.

Assumption 1.3.1. For a stationary process $(Z_n)_{n\in\mathbb{Z}}$ defined on a measurable space $(S, \mathcal{S})$ and two $\mathcal{S}$-measurable functions $b : S \mapsto (0, \infty)$ and $a : S \mapsto \mathbb{R}$ the following inequalities hold

$$\mathbb{E}\left[\log_+ b(Z_0)\right] < \mathbb{E}\left[\log_- b(Z_0)\right] \in [0, \infty], \quad \mathbb{E}\left[|a(Z_0)|\right] < \infty.$$

Remark 1.3.2. This type of assumption is often used in the literature as a sufficient condition for stochastic recursive difference equations of the type (1.1.8) with stationary and ergodic coefficients to have a unique stationary solution, see [11] and references therein.

Definition 1.3.1. With a slight abuse of notation, we sometimes write $\mathbb{T}(B; X_k) = \mathbb{P}\left[X_{k+1} \in B \mid X_k\right]$ or $\mathbb{T}(B; x) = \mathbb{P}\left[X_{k+1} \in B \mid X_k = x\right]$, for the transition kernel $\mathbb{T}$ of a stationary Markov chain Z living on some $(D, \mathcal{D})$, for all $B \in \mathcal{D}$.

Whenever we consider a Markovian environment we assume that Z is a stationary ergodic Markov chain with the associated transition kernel $\mathbb{T}$, i.e. $\mathbb{T}(B; Z_k) = \mathbb{P}[Z_{k+1} \in B \mid Z_k] = \mathbb{P}\left[Z_{k+1} \in B \mid \mathcal{F}_k^Z\right]$ for all $B \in \mathcal{S}$, where $\mathcal{F}_n^Z := \sigma(Z_k;\, k \leq n)$ for $n \in \mathbb{Z}$. Further, let $\mathbb{T}^n$ denote the n-step Markov transition kernel corresponding to $\mathbb{T}$, i.e. $\mathbb{T}^n(B; Z_k) = \mathbb{P}[Z_{n+k} \in B \mid Z_k] = \mathbb{P}\left[Z_{n+k} \in B \mid \mathcal{F}_k^Z\right]$.

Remark 1.3.3. In the literature the definition of ergodicity for a Markov chain differs from the one for a dynamical system, intuitively meaning that there is only one invariant set of positive measure. For the avoidance of doubt, by ergodicity of a Markov chain we mean that it is irreducible, aperiodic and positive recurrent. This type of chains corresponds to an ergodic system, but the converse does not hold without some additional assumptions, see [106].

Remark 1.3.4. Whenever we refer to the concept of irreducibility, we sometimes write that either a Markov chain is irreducible or the associated transition kernel is.

1.4 Convergence to the stationary distribution

We start this section by proving that Assumption 1.3.1 guarantees that the random variable (1.1.1) is well-defined.

Lemma 1.4.1. *Under Assumption 1.3.1, it holds that*

$$\sum_{k=n}^{\infty} \left(\prod_{n \leq \ell < k} b(Z_\ell) \right) |a(Z_k)| < \infty, \quad \forall n \in \mathbb{Z}, \quad \mathbb{P}\text{-a.s.}$$

Proof. By stationarity, it suffices to argue for $n = 0$. The fact that $\mathbb{E}\left[\log_+ b(Z_0)\right] < \mathbb{E}\left[\log_- b(Z_0)\right]$ and the ergodic theorem imply the existence of $\epsilon \in (0,1)$ and an $\mathbb{N}$-valued random variable M such that $\prod_{n \leq \ell < k} b(Z_\ell) \leq (1-\epsilon)^k$ holds for $k \geq M$; Then,

$$\mathbb{E}\left[\sum_{k=M}^{\infty} \left(\prod_{0 \leq \ell < k} b(Z_\ell) \right) |a(Z_k)| \right] \leq \mathbb{E}\left[\sum_{k=0}^{\infty} (1-\epsilon)^k |a(Z_k)| \right] = \frac{\mathbb{E}\left[|a(Z_0)|\right]}{\epsilon} < \infty,$$

which implies that, $\mathbb{P}$-a.s., $\sum_{k=M}^{\infty} \left(\prod_{0 \leq \ell < k} b(Z_\ell) \right) |a(Z_k)| < \infty$. □

Remark 1.4.2. Products over empty sets, like $\prod_{n \leq \ell < n} a(Z_\ell)$, are by convention equal to one.

By Lemma 1.4.1, under Assumption 1.3.1 the random variables

$$X_n := \sum_{k=n}^{\infty} \left(\prod_{n \leq \ell < k} b(Z_\ell) \right) a(Z_k), \quad n \in \mathbb{Z},$$

are well-defined. Furthermore, [64, Lemma 9.5] implies that the process $(Z_n, X_n)_{n\in\mathbb{Z}}$ is stationary and ergodic.

Remark 1.4.3. Suppose that, additionally, $(S, \mathcal{S})$ is a Borel space and Z is a Markov process, and let $(\mathcal{F}_n^Z)_{n\in\mathbb{Z}}$ be the natural filtration of Z, defined as above $\mathcal{F}_n^Z := \sigma(Z_k;\, k \le n)$ for $n \in \mathbb{Z}$. Furthermore, let $\pi^{X|Z} : \mathcal{B}(\mathbb{R}) \times S \mapsto [0,1]$ denote the conditional law of X_0 given Z_0, such that $\pi(B \times C) = \int_C \pi^{X|Z}(B; s)\pi^Z[\mathrm{d}s]$ holds for all $B \in \mathcal{B}(\mathbb{R})$ and $C \in \mathcal{S}$. It then follows that the conditional law of X_n given $\mathcal{F}_n^Z$ coincides with the conditional law of X_n given Z_n, the latter being $\pi^{X|Z}(\bullet; Z_n)$; in other words,

$$\mathbb{P}\left[X_n \in B \mid \mathcal{F}_n^Z\right] = \pi^{X|Z}(B; Z_n), \quad \mathbb{P}\text{-a.s., } \forall B \in \mathcal{B}(\mathbb{R}).$$

Even though we shall not be assuming the Markovian property of Z in this chapter in what follows, this is clearly the most interesting case.

By definition (see Theorem 2 in [11]), we have

$$X_n = a(Z_n) + b(Z_n)X_{n+1}, \quad \forall n \in \mathbb{Z}. \tag{1.4.1}$$

Let $(\zeta_n, \chi_n) = (Z_{-n}, X_{-n})$ for all $n \in \mathbb{Z}$, so that the process $(\zeta_n, \chi_n)_{n\in\mathbb{Z}}$ is the time-reversal of $(Z_n, X_n)_{n\in\mathbb{Z}}$. Time-reversal is especially widely used in when the coefficients in (1.4.1) form independent and identically distributed sequences of random variables. In this simplified case, when the property of independence between the random variables can be utilised, applying time reversal usually leads to a fixed point stochastic equation of the form similar to (1.1.6) from where the distributional properties of the process can be derived, see [27] and references therein. In the presence of Markovian dependency, applying time-reversal does not bring us to the same fixed point stochastic equation, but it helps to construct a new process with more manageable dynamics than (1.4.1). That allows us to infer the main distributional properties of the reversed process while preserving all of them. Lemma 1.4.4 formulated below proves that the reversed process (ζ_n, χ_n) is stationary and ergodic. In its proof we use the notion of shift operator ϑ. If we consider a random sequence $\alpha = (\alpha_1, \alpha_2, \ldots)$ on some measurable space $(S', \mathcal{S}')$ than the shift operator is acting on $\tilde{S} = (S')^\infty$ in the following way $\vartheta(x_0, x_1, \ldots) = (x_1, x_2, \ldots)$.

Lemma 1.4.4. *Under Assumption 1.3.1, the process $(\zeta_n, \chi_n)_{n\in\mathbb{Z}}$ is stationary and ergodic.*

Proof. Let f_{rev} be the time-reversal mapping on $S^Z \times \mathbb{R}^Z$, such that $(\zeta, \chi) = f_{rev}(Z, X)$. Then, for the shift operator $\vartheta_{\zeta,\chi}$ on $S^Z \times \mathbb{R}^Z$ given by $\vartheta_{\zeta,\chi}((\zeta_0, \chi_0), (\zeta_1, \chi_1), ...) = ((\zeta_1, \chi_1), (\zeta_2, \chi_2), ...)$, we have $\vartheta_{\zeta,\chi}(\zeta, \chi) = \vartheta_{\zeta,\chi} f_{rev}(Z, X) = f_{rev} \circ \vartheta(Z, X) = f_{rev}(Z, X)(\zeta, \chi)$, because (Z, X) is stationary, see [64, Lemma 9.5]. The ergodicity of (ζ, χ) follows from $f_{rev} \circ \vartheta = \vartheta_{\zeta,\chi} \circ f_{rev}$ and the ergodicity (Z, X) similar to the proof of [64, Lemma 9.5]. □

Hence, the stationary law of (ζ, χ) clearly coincides with the one of (Z, X), i.e. it is π. An immediate consequence of the result is the following version of the law of large numbers.

COROLLARY 1.4.1. For every bounded and $\mathcal{S} \otimes \mathcal{B}(\mathbb{R})$-measurable g, it follows that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} g(\zeta_n, \chi_n) = \langle \pi, g \rangle \equiv \int g \mathrm{d}\pi, \quad \mathbb{P}\text{-a.s.} \tag{1.4.2}$$

Moreover, the equation (1.4.1) translates into

$$\chi_n = a(\zeta_n) + b(\zeta_n)\chi_{n-1}, \quad \forall n \in \mathbb{Z}\,, \tag{1.4.3}$$

using the time-reversal argument.

To simplify some of the proofs of the results below in a similar fashion, for any fixed $x \in \mathbb{R}$, we define the process $\chi^x \equiv (\chi^x_n)_{n\in\mathbb{N}}$ following the recursion

$$\chi^x_n = a(\zeta_n) + b(\zeta_n)\chi^x_{n-1},$$

for all $n \in \mathbb{N} \setminus \{0\}$ with $\chi^x_0 = x$. The process χ^x differs from the second component of the process (ζ_n, χ_n) only by the starting state, as the initial position for χ^x is fixed at some point x, whereas for χ it is a random state. Lemma 1.4.5 below states that these two processes are close to each other asymptotically as $n \to \infty$.

Lemma 1.4.5. *For any fixed $x \in \mathbb{R}$, it follows that $|\chi^x_n - \chi_n| \to 0$, $\mathbb{P}$-a.s. as $n \to \infty$.*

Proof. Given the value of χ_0, it follows that

$$\chi_n = \left(\prod_{0<k\le n} b(\zeta_k) \right) \chi_0 + \sum_{k=1}^{n} \left(\prod_{k<\ell\le n} b(\zeta_\ell) \right) a(\zeta_k), \quad n \in \mathbb{N}. \tag{1.4.4}$$

It then follows that

$$\chi^x_n = \left(\prod_{0<k\le n} b(\zeta_k) \right) x + \sum_{k=1}^{n} \left(\prod_{k<\ell\le n} b(\zeta_\ell) \right) a(\zeta_k), \quad n \in \mathbb{N}.$$

In other words, we have

$$\chi^x_n = \chi_n + \left(\prod_{0<k\le n} b(\zeta_k) \right) (x - \chi_0)\,, \quad n \in \mathbb{N}.$$

Since $\prod_{k>0} b(\zeta_k) = \lim_{n\to\infty} \prod_{0<k\le n} b(\zeta_k) = 0$ (see proof of Lemma 1.4.1), the result follows. □

Remark 1.4.6. This implies that the ergodic limit (1.4.2) will carry also when χ (which is not practically computable, since the value of χ_0 is unknown) is replaced by χ^x (which is easily computable in practice), when g has some continuity in the second argument.

Lemma 1.4.7. *Let g be bounded and $\mathcal{S}\otimes\mathcal{B}(\mathbb{R})$-measurable, and such that $g(s,\bullet)$ is continuous for π^Z-a.e. $s\in S$. Then,*

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N} g(\zeta_n,\chi_n^x)=\big\langle\pi,g\big\rangle,\quad \forall x\in\mathbb{R},\quad \mathbb{P}\text{-a.s.} \tag{1.4.5}$$

Proof. First, assume that g is bounded and K-Lipschitz, i.e., such that $\sup_{s\in S}|g(s,x)-g(s,y)|\le K|x-y|$ holds for $K\in[0,\infty)$ and all $(x,y)\in\mathbb{R}^2$. Then, it follows that

$$\sum_{n=1}^{\infty}|g(\zeta_n,\chi_n^x)-g(\zeta_n,\chi_n)|\le K|x-\chi_0|\sum_{n=1}^{\infty}\left(\prod_{0<k\le n}b(\zeta_k)\right).$$

Since $\sum_{n=1}^{\infty}\big(\prod_{0<k\le n}b(\zeta_k)\big)<\infty$ holds $\mathbb{P}$-a.s., with the proof identical to the one of Lemma 1.4.1, (1.4.5) follows from (1.4.2) and the above estimate. When g is bounded and only assumed continuous in the second variable, one may find a non-decreasing sequence $(g_m^{\uparrow})_{m\in\mathbb{N}}$ and a non-increasing sequence $(g_m^{\downarrow})_{m\in\mathbb{N}}$, consisting of m-Lipschitz functions that converge point-wise to g, at least for π^Z-a.s. $s\in S$. It then follows that

$$\big\langle\pi,g_m^{\uparrow}\big\rangle\le\liminf_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}g(\zeta_n,\chi_n^x)\le\limsup_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}g(\zeta_n,\chi_n^x)\le\big\langle\pi,g_m^{\downarrow}\big\rangle,\quad\forall x\in\mathbb{R},\quad\forall m\in\mathbb{N}$$

in the $\mathbb{P}$-a.s. sense from the result obtained earlier. By the monotone convergence theorem as $m\to\infty$, we obtain (1.4.5). □

Remark 1.4.8. The same procedure can be applied in the original forward sequence $(Z_n,X_n)_{n\in\mathbb{Z}}$. In particular, the recursion $X_{n+1}=X_n/b(Z_n)-a(Z_n)/b(Z_n)$, valid for all $n\in\mathbb{Z}$, would imply that

$$X_n=\left(\prod_{0<k\le n}b(Z_{k-1})\right)^{-1}X_0-\sum_{k=1}^{n}\left(\prod_{k<\ell\le n}b(Z_{\ell-1})\right)^{-1}a(Z_{k-1}),\quad n\in\mathbb{N}.$$

For any fixed $x\in\mathbb{R}$, define the process $X^x\equiv(X_n^x)_{n\in\mathbb{N}}$ via $X_0^x=x$ and the recursion $X_{n+1}^x=X_n^x/b(Z_n)-a(Z_n)/b(Z_n)$ for all $n\in\mathbb{N}$. It then follows that

$$X_n^x=\left(\prod_{0<k\le n}b(Z_{k-1})\right)^{-1}x-\sum_{k=1}^{n}\left(\prod_{k<\ell\le n}b(Z_{\ell-1})\right)^{-1}a(Z_{k-1}),\quad n\in\mathbb{N}.$$

Therefore,

$$X_n^x = X_n + \left(\prod_{0<k\leq n} b(Z_{k-1}) \right)^{-1} (x - X_0)\,, \quad n \in \mathbb{N}.$$

Since $\prod_{0<k<\infty} b(Z_{k-1}) = 0$ holds $\mathbb{P}$-a.s. and $(X_n)_{n\in\mathbb{N}}$ is stationary, it follows that $(X_n^x)_{n\in\mathbb{N}}$ is divergent on the event $\{X_0 \neq x\}$. Contrary to the case of the reversed process χ and χ^x that are mixing (exponentially fast, pathwise), the forward processes X and X^x are diverging (exponentially fast, pathwise).

Remark 1.4.9. Let g be bounded and $\mathcal{S} \otimes \mathcal{B}(\mathbb{R})$-measurable, and such that $\sup_{s\in S}|g(s,x) - g(s,y)| \leq K|x-y|$ holds for some $K \in [0,\infty)$ and all $(x,y) \in \mathbb{R}^2$. Suppose that a CLT result holds:

$$\lim_{N\to\infty} \sqrt{N} \left(\frac{1}{N} \sum_{n=1}^{N} g(\zeta_n, \chi_n) - \langle \pi, g \rangle \right) = \sigma_g \mathcal{N}, \tag{1.4.6}$$

where the above is weak convergence in law, $\sigma_g \in [0,\infty)$ and $\mathcal{N}$ denotes the standard Gaussian law. Then, since

$$\sqrt{N}|\frac{1}{N} \sum_{n=1}^{N} g(\zeta_n, \chi_n^x) - \frac{1}{N} \sum_{n=1}^{N} g(\zeta_n, \chi_n)| \leq \frac{K|x-\chi_0|}{\sqrt{N}} \sum_{n=1}^{\infty} \left(\prod_{0<k\leq n} b(\zeta_k) \right),$$

as was shown in the proof of Lemma 1.4.7, implying that the quantity on the left-hand-side above converges $\mathbb{P}$-a.s. to zero as $N \to \infty$, we obtain

$$\lim_{N\to\infty} \sqrt{N} \left(\frac{1}{N} \sum_{n=1}^{N} g(\zeta_n, \chi_n^x) - \langle \pi, g \rangle \right) = \sigma_g \mathcal{N}.$$

1.5 Pivotal application

1.5.1 Discretisation of continuous-time Markov chains

In the next two sections, we introduce the main concepts and properties of Markov chains which are used throughout, and discuss connections between discrete- and continuous-time Markov chains. The main difference between discrete- and continuous-time Markov chains is that the latter stays in a given state for a period of time that is exponentially distributed and the rate of that distribution usually depends on the current state of the chain.

A continuous-time Markov chain on a finite state space is a continuous-time Markov process whose sample paths are right-continuous and piecewise-constant of finite lengths. The number of jumps in any finite time should be finite as well. This type of Markov processes is represented by the sequence of states it visits and the corresponding times of these visits.

Definition 1.5.1 (*Continuous-time Markov chain*)**.** A continuous-time Markov chain with a finite or countable state space I is a family $\{Y_t = Y(t)\}_{t\in\mathbb{R}_+}$ of I-measured random variables such that

1. The paths $t \to Y(t)$ are right-continuous step functions.

2. For any set of times $t_i < t_{i+1} = t_i + s_{i+1}$, such that, $t_0 = 0$ and set of states $y_i \in I$

$$\mathbb{P}(Y(t_{k+1}) = y_{k+1}|Y(t_i) = y_i \ \ \forall i \leq k) = \mathbb{P}(Y(s_{k+1}) = y_{k+1}|Y(0) = y_k). \qquad (1.5.1)$$

Discrete-time Markov chains are mainly used for simulation purposes as it is much easier to construct algorithms in discrete time. Apart from that, they are quite often implemented as crude models of physical, biological, social, and financial processes. However, when dealing with models describing physical and biological worlds where time runs continuously, discrete-time dynamical systems might not be an appropriate choice. In such, continuous-time Markov chains can be a better fit. Often one can identify for a continuous-time Markov chain its discrete-time analogue, with the famous examples of birth-death and Brownian motion processes. These two continuous-time processes correspond to random walks in discrete time. If we rewrite 1.5.1 in the following way

$$\mathbb{P}(Y(t_i) = y_i \ \ \forall 0 \leq i \leq k+1) = \mathbb{P}(Y(0) = y_0) \prod_{i=0}^{k} p_{s_{i+1}}(y_i, y_{i+1}),$$

with $p_t(a, b) := \mathbb{P}(Y(t) = y|Y(0) = y_0)$. This suggests if we take a scalar $\delta > 0$, then the discrete sequence $\{Y(n\delta)\}_{n\leq 0}$ represents a discrete-time Markov chain. The corresponding transition probabilities of this Markov chain are $p_\delta(a, b)$.

1.5.2 Embedded Markov chains

We start this section by describing the structure of a continuous-time Markov chain that is used to construct the main application result in the next section. Let $(Y_t)_{t\in\mathbb{R}_+}$ be an ergodic continuous-time Markov chain with finite state space I constructed in the following way. Define $(\tau_n)_{n\in\mathbb{N}}$ as the jump times of Y (where we set $\tau_0 = 0$), and let $(y_n)_{n\in\mathbb{N}}$ be the associated embedded discrete time Markov chain, defined via $y_n = Y_{\tau_n}$. The inter-arrival times $\tau_{n+1} - \tau_n$ are exponentially distributed with rate $\lambda(y_n)$, where $\lambda : I \mapsto (0, \infty)$ is a given function; more precisely, the random variables $\lambda(y_n)(\tau_{n+1} - \tau_n)$ are independent and identically distributed with the standard exponential law, and further independent of the discrete-time Markov chain $(y_n)_{n\in\mathbb{N}}$. Therefore, upon defining $u_n := \exp(-\lambda(y_n)(\tau_{n+1} - \tau_n))$, so that $\tau_{n+1} - \tau_n = -(1/\lambda(y_n)) \log u_n$ for all $n \in \mathbb{N}$, it follows that the sequence $(u_n)_{n\in\mathbb{N}}$ consists of independent and identically distributed

random variables with the standard uniform law, and are further independent of the discrete-time Markov chain $(y_n)_{n\in\mathbb{N}}$. The latter Markov chain $(y_n)_{n\in\mathbb{N}}$ has transition matrix $(p(i,j))_{(i,j)\in I\times I}$, where $p(i,i) = 0$ for all $i \in I$. These transition probabilities of the embedded discrete time Markov chain $(y_n)_{n\in\mathbb{N}}$ can be computed in the following way

$$\begin{aligned} p(i,j) &= \lim_{\Delta t\to 0} \mathbb{P}\{Y_{t+\Delta t} = j | Y_{t+\Delta t} \neq i, Y_t = i\} \\ &= \lim_{\Delta t\to 0} \frac{\mathbb{P}\{Y_{t+\Delta t} = j, Y_{t+\Delta t} \neq i | Y_t = i\}}{\mathbb{P}\{Y_{t+\Delta t} \neq i | Y_t = i\}}. \end{aligned}$$

Denote the invariant law for $(y_n)_{n\in\mathbb{N}}$ as p^*, then it follows that $p^*(j) = \sum_{i\in I} p^*(i)p(i,j)$ holds for all $j \in I$. The following result connects the stationary distributions of the continuous-time Markov chain $(Y_t)_{t\in\mathbb{R}_+}$ and its embedded chain $(y_n)_{n\in\mathbb{N}}$. We refer to [104] for its proof. Now we formulate the following well known result for ergodic continuous time Markov chains that allows us to derive the invariant law of $(Y_t)_{t\in\mathbb{R}_+}$.

Proposition 1.5.2. *If $(Y_t)_{t\in\mathbb{R}_+}$ and $(y_n)_{n\in\mathbb{N}}$ are ergodic on I with the corresponding stationary distributions being π and p^*, then*

$$\pi(i) = \frac{p^*(i)/\lambda(i)}{\sum_{j\in I} p^*(j)/\lambda(j)}$$

holds for every $i \in I$.

Based on Proposition 1.5.2 the invariant law for $(Y_t)_{t\in\mathbb{R}_+}$ is such that

$$\mathbb{P}[Y_t = i] = \frac{p^*(i)/\lambda(i)}{\sum_{j\in I} p^*(j)/\lambda(j)}, \quad i \in I, \quad t \in \mathbb{R}_+.$$

1.5.3 Perpetuity driven by a continuous-time Markov chain

In this section, we consider a crucial example of a perpetuity that is driven by a continuous-time Markov chain. Define the following random variable Q_0, that is driven by the ergodic continuous-time Markov chain $(Y_t)_{t\in\mathbb{R}_+}$ defined in the previous section

$$Q_0 = \int_0^\infty \exp\left(-\int_0^t r(Y_{u-})\mathrm{d}u\right) h(Y_t)\mathrm{d}t + \sum_{n>0} \exp\left(-\int_0^{\tau_n} r(Y_{u-})\mathrm{d}u\right) c(Y_{\tau_n-}, Y_{\tau_n}), \tag{1.5.2}$$

where, it is assumed that $r : I \mapsto \mathbb{R}$, $h : I \mapsto \mathbb{R}$ and $c : I \times I \mapsto \mathbb{R}$ are appropriate functions to make the integrals well-defined.[1] We are interested in the conditional law of Q_0 given

[1] In fact, in connection to Assumption 1.3.1, we need to assume that $\sum_{j\in I} r(i)p^*(i)/\lambda(i) > 0$. Simply assuming that $r(i) > 0$ for all $i \in I$ is sufficient.

Y_0, where Q_0 is of the form (1.5.2). The first part of (1.5.2)

$$Q_0^1 = \int_0^\infty \exp\left(-\int_0^t r(Y_{u-})\mathrm{d}u\right) h(Y_t)\mathrm{d}t$$

represents a continuous time perpetuity of rate $h(Y)$ and discounting rate $r(Y)$, while the second part

$$Q_0^2 = \sum_{n>0} \exp\left(-\int_0^{\tau_n} r(Y_{u-})\mathrm{d}u\right) c(Y_{\tau_n-}, Y_{\tau_n})$$

is the discounted payoff of bulk payments made at transition times τ_n, with the payment $c(Y_{\tau_n-}, Y_{\tau_n})$ depending on the states before and at transition. Perpetuities of the form similar to (1.5.2) are derived in [116] from SDEs for the Ornstein-Uhlenbeck process.

1.5.4 The case $c \equiv 0$

We start with the case where $c \equiv 0$ in (1.5.2). In other words, we have

$$Q_0 = Q_0^1 = \int_0^\infty \exp\left(-\int_0^t r(Y_{u-})\mathrm{d}u\right) h(Y_t)\mathrm{d}t.$$

With notation previously set, note that

$$\exp\left(-\int_0^{\tau_k} r(Y_u)\mathrm{d}u\right) = \prod_{0\le\ell<k} \exp\left(-r(y_\ell)(\tau_{\ell+1}-\tau_\ell)\right) = \prod_{0\le\ell<k} u_\ell^{r(y_\ell)/\lambda(y_\ell)}, \quad k \in \mathbb{N},$$

as well as

$$\begin{aligned}
\int_{\tau_k}^{\tau_{k+1}} \exp\left(-\int_0^t r(Y_u)\mathrm{d}u\right) h(Y_t)\mathrm{d}t &= \exp\left(-\int_0^{\tau_k} r(Y_u)\mathrm{d}u\right) \int_{\tau_k}^{\tau_{k+1}} \exp\left(-\int_{\tau_k}^t r(Y_u)\mathrm{d}u\right) h(Y_t)\mathrm{d}t \\
&= \left(\prod_{0\le\ell<k} u_\ell^{r(y_\ell)/\lambda(y_\ell)}\right) \int_{\tau_k}^{\tau_{k+1}} \exp\left(-r(y_k)(t-\tau_k)\right) h(y_k)\mathrm{d}t \\
&= \left(\prod_{0\le\ell<k} u_\ell^{r(y_\ell)/\lambda(y_\ell)}\right) \frac{h(y_k)}{r(y_k)} \left(1-\exp\left(-r(y_k)(\tau_{k+1}-\tau_k)\right)\right) \\
&= \left(\prod_{0\le\ell<k} u_\ell^{r(y_\ell)/\lambda(y_\ell)}\right) \frac{h(y_k)}{r(y_k)} \left(1-u_k^{r(y_k)/\lambda(y_k)}\right), \quad k \in \mathbb{N}.
\end{aligned}$$

Define the product state space as $S = I \times [0,1]$, and introduce two functions a and b in the following way

$$b(i,u) = u^{r(i)/\lambda(i)},$$

$$a(i,u) = \frac{h(i)}{r(i)}\left(1-u^{r(i)/\lambda(i)}\right) = \frac{h(i)}{r(i)}\left(1-b(i,u)\right), \quad (i,u) \in S.$$

Then, we have established that

$$\exp\left(-\int_0^{\tau_k} r(Y_u)\mathrm{d}u\right) = \prod_{0\le\ell<k} b(y_\ell, u_\ell), \quad k \in \mathbb{N},$$

as well as

$$\int_{\tau_k}^{\tau_{k+1}} \exp\left(-\int_0^t r(Y_u)\mathrm{d}u\right) h(Y_t)\mathrm{d}t = \left(\prod_{0\le\ell<k} b(y_\ell, u_\ell)\right) a(y_k, u_k), \quad k \in \mathbb{N}.$$

Therefore, when $c \equiv 0$

$$Q_0 = \sum_{k=0}^{\infty} \left(\prod_{0\le\ell<k} b(y_\ell, u_\ell)\right) a(y_k, u_k).$$

Moreover, with $Z_n = (y_n, u_n)$ for each $n \in \mathbb{N}$, and X_n as defined in the previous section, we have that the conditional law of Q_0 given Y_0 coincides with the conditional law of X_0 given y_0.[2] We denote by $(\psi_n)_{n\in\mathbb{N}}$ the time-reversed chain associated to $(y_n)_{n\in\mathbb{N}}$ and, with some abuse of notation, we still denote by $(u_n)_{n\in\mathbb{N}}$ the time-reversed chain of the independent uniforms. Then, the recursion becomes

$$\chi_n = u_n^{r(\psi_n)/\lambda(\psi_n)}\chi_{n-1} + (1 - u_n^{r(\psi_n)/\lambda(\psi_n)})\frac{h(\psi_n)}{r(\psi_n)}. \tag{1.5.3}$$

The above recursion may be also used to obtain a recursive scheme for the conditional laws. For each $n \in \mathbb{N}$, let $F_n(x; i) = \mathbb{P}[\chi_n \le x | \psi_n = i]$. Then,

$$\begin{aligned}
F_n(x;i) = \frac{\mathbb{P}[\chi_n \le x, \psi_n = i]}{\mathbb{P}[\psi_n = i]} &= \frac{1}{p_i^*}\sum_{j\in I} \mathbb{P}[\chi_n \le x, \psi_n = i | \psi_{n-1} = j]\, p_j^* \\
&= \sum_{j\in I} \mathbb{P}[\chi_n \le x | \psi_n = i, \psi_{n-1} = j]\, \frac{\mathbb{P}[\psi_n = i | \psi_{n-1} = j]\, p_j^*}{p_i^*} \\
&= \sum_{j\in I} \mathbb{P}[\chi_n \le x | \psi_n = i, \psi_{n-1} = j]\, p_{ij} \\
&= \sum_{j\in I} \mathbb{P}\left[\chi_{n-1} \le u_n^{-r(i)/\lambda(i)} x - (u_n^{-r(i)/\lambda(i)} - 1)\frac{h(i)}{r(i)} \,\middle|\, \psi_{n-1} = j\right] p_{ij} \\
&= \sum_{j\in I} p_{ij} \int_0^1 F_{n-1}\left[u^{-r(i)/\lambda(i)} x - (u^{-r(i)/\lambda(i)} - 1)\left(h(i)/r(i)\right); j\right] \mathrm{d}u.
\end{aligned}$$

[2]Note, however, that the marginal laws of Y_0 and y_0 are different. This created no problem, since these marginal laws are explicitly known in terms of $p : I \times I \mapsto [0, 1]$ and $\lambda : I \mapsto (0, \infty)$.

1.5.5 The case of general c

With the same notation, observe that

$$\exp\left(-\int_0^{\tau_{k+1}} r(Y_{u-})\mathrm{d}u\right) c(Y_{\tau_{k+1}-}, Y_{\tau_{k+1}}) = \left(\prod_{0\le\ell<k} u_\ell^{r(y_\ell)/\lambda(y_\ell)}\right) u_k^{r(y_k)/\lambda(y_k)} c(y_k, y_{k+1})$$

The state space now becomes $S = I \times I \times (0,1)$, $Z_n = (y_n, y_{n+1}, u_n)$ for all $n \in \mathbb{N}$, the associated functions

$$b(i,j,u) = u^{r(i)/\lambda(i)},$$

$$a(i,j,u) = \frac{h(i)}{r(i)}\left(1 - b(i,u)\right) + b(i,u)c(i,j), \quad (i,j,u) \in S,$$

and the recursion for the time-reversed process becomes

$$\chi_n = u_n^{r(\psi_n)/\lambda(\psi_n)}\left(\chi_{n-1} + c(\psi_n, \psi_{n-1})\right) + (1 - u_n^{r(\psi_n)/\lambda(\psi_n)})\frac{h(\psi_n)}{r(\psi_n)}. \tag{1.5.4}$$

Further, we formulate the following assumption for the joint process $(\psi_n, \chi_n)_{n\in\mathbb{N}}$ and, throughout the rest of this paper, we assume the following condition for the joint process $(\psi_n, \chi_n)_{n\in\mathbb{N}}$ holds. It is introduced for the convenience of the presentation in the following sections. The property of irreducibility and ergodicity holds for $(\psi_n)_{n\in\mathbb{N}}$ automatically since $(y_n)_{n\in\mathbb{N}}$ is assumed to be finite state and ergodic, see remarks on this after the assumption.

Assumption 1.5.3. The Markov chain $(\psi_n)_{n\in\mathbb{N}}$ is irreducible on the finite state space I.

From now on, whenever we refer to the process $(\chi_n)_{n\in\mathbb{N}}$ as a solution to (1.5.3), we assume it is a stationary process.

Remark 1.5.4. The existence of a stationary solution to (1.5.3) is evident from the results of Section 1.4. Moreover, this stationary solution to (1.5.3) is unique under Assumption 1.3.1.

Remark 1.5.5. Since the Markov chain $(y_n)_{n\in\mathbb{N}}$ is assumed to be finite state and ergodic, the associated reversed Markov chain $(\psi_n)_{n\in\mathbb{N}}$ is automatically irreducible.

Assumption 1.5.3 allows us to reduce our analysis of the process $(y_n)_{n\in\mathbb{N}}$ to the analysis of the associated reversed Markov chain $(\psi_n)_{n\in\mathbb{N}}$.

1.6 Geometric ergodicity

Geometric ergodicity refers to a concept of "fast" convergence of a Markov chain to its invariant distribution and is strongly connected with central limit theorems.

Definition 1.6.1 (*Geometric Ergodicity*)**.** A Markov chain X with a transition kernel $\mathbb{T}$ and stationary distribution $\pi(\bullet)$ is called geometrically ergodic if

$$\|\mathbb{T}^n(\bullet\,; x) - \pi(\bullet)\|_{TV} \leq M(x) r^n, \tag{1.6.1}$$

for some function $0 \leq M(x) < \infty$ and some $0 \leq r < 1$, $\pi-$a.s.

Typically, the techniques to establish geometric ergodicity are quite involved, see [88], unless the state space is finite, in which case geometric ergodicity is equivalent to ergodicity. It is known, see [88], that geometric ergodicity is equivalent to the existence of the so-called drift to a *petite* set.

Definition 1.6.2 (*Petite set C*)**.** A subset $C \subseteq \mathcal{S}$ is *ν_a-petite* if the sampled chain satisfies the bound

$$\mathbb{T}_a(\bullet\,; x) \geq \nu_a(\bullet), \qquad x \in C, \tag{1.6.2}$$

i.e. $\mathbb{T}_a(A; x) \geq \nu_a(A)$ for all $x \in C$ and all measurable $A \subseteq \mathcal{S}$, where $\nu_a(A)$ is a non-trivial measure on $\mathcal{S}$.

Definition 1.6.3 (*Geometric Drift Towards C*)**.** The geometric drift condition for a Markov chain X on a state space S holds if there exists a real valued measurable function $V : S \to [1, \infty]$ finite as some one point $x_0 \in S$ such that for some constants $\alpha > 0$, $\beta < \infty$, and a measurable set C,

$$\mathbb{E}[V(X_n)|X_{n-1} = x] - V(x) \leq -\alpha V(x) + \beta \mathbb{I}_C(x), \quad x \in S.$$

1.6.1 Supporting results

In this section we consider a collection of technical results formulated for Markov chains that are required for the proofs of the main results. The first such result that we prove in this section provides us with necessary and sufficient conditions for a Markov chain on general state space to be φ-irreducible and it is stated in Lemma 1.6.6.

Definition 1.6.4 (*φ-irreducibility on general state space*)**.** The Markov chain X with transition kernel $\mathbb{T}$ is called φ-irreducible if, for every measurable A with $\varphi(A) > 0$ and every $x \in X$, there exists an $n \in \mathbb{N}$ (which may depend on x and A) such that $\mathbb{T}^n(A; x) > 0$.

In words, the meaning behind φ-irreducibility is that every set A with $\varphi(A) > 0$ is reachable from any $x \in X$. We say that the chain is φ-reducible if it is not φ-irreducible, meaning that, when there exist x and A with $\varphi(A) > 0$ such that $\mathbb{T}^n(A; x) = 0$ for all $n \in \mathbb{N}$. Next, we need to formulate another definition that is used in Lemma 1.6.6. This definition introduces the notion of a closed set for Markov chains.

Definition 1.6.5 (*Closed set for Markov chains*)**.** For a Markov chain X on a state space $\mathbf{X}$ with transition kernel $\mathbb{T}$ a set $C \in \mathcal{B}(\mathbf{X})$, where $\mathcal{B}(\mathbf{X})$ is the Borel sigma algebra, is called closed if it is non-empty and $\mathbb{T}(C^c; x) = 0$ for all $x \in C$.

From Definition (1.6.5) it follows that a non-empty set $C \subseteq X$ is called closed if, once the chain enters the set C it can not leave this set anymore. As an obvious observation the state space $\mathbf{X}$ is a closed set.

Lemma 1.6.6. *A Markov chain X with transition kernel $\mathbb{T}$ on a general state space $\mathbf{X}$ is φ-reducible if and only if there exists a closed set C such that $\varphi(C^c) > 0$.*

Proof. Let X be a φ-reducible Markov chain on a state space $\mathbf{X}$ with transition kernel $\mathbb{T}$, meaning that there exists a measurable set $B \in \mathcal{B}(\mathbf{X})$ such that $\varphi(B) > 0$ and a point $x \in B^c$ with the property that $\forall n \in \mathbb{N}$ it holds that $\mathbb{T}^n(B; x) = 0$. Define the set A as a union of $A_k = \{x : \mathbb{T}^k(B; x) > 0\}$ with $k \in \mathbb{N}$, then one can show that the set $C := A^c \cap B^c$ is closed in the definition for Markov chains sense. First, note that $\varphi(C^c) > \varphi(A) > 0$. We next prove the closeness of C by contradiction. Suppose that $x^* \in C$ and $\mathbb{T}(C^c; x^*) > 0$, then we should have

$$\mathbb{T}(A; x^*) + \mathbb{T}(B; x^*) \geq \mathbb{T}(A \cup B; x^*) = \mathbb{T}(C^c; x^*) > 0. \tag{1.6.3}$$

Since x^* is in C, it follows that x^* is in A^c and hence A_1^c. From the fact that the chain X is φ-reducible, we know that $\mathbb{T}(B; x^*) = 0$ and hence it must be true that $\mathbb{T}(A; x^*) > 0$, so that (1.6.3) holds. Then there exists $k \in \mathbb{N}$ with $\mathbb{T}(A_k; x^*) > 0$ and as a result

$$\mathbb{T}^{(k+1)}(B; x^*) = \int_{\mathbf{X}} \mathbb{T}^k(B; y)\mathbb{T}(dy; x) = \int_{A_k} \mathbb{T}^k(B; y)\mathbb{T}(dy; x^*) > 0.$$

This however implies that x^* is inside of A_{k+1} which contradicts the fact that x^* is in C. Hence, for all $x \in C$ it must be the case that $\mathbb{T}(C^c; x) = 0$ and one can conclude that the set C is closed in the sense of the definition for Markov chains.
Now, suppose that the set C is closed and $\varphi(C^c) > 0$. By induction one can show that if $\mathbb{T}(C^c; x) = 0\ \forall x \in C$, then $\forall x \in C$ and $\forall n \in \mathbb{N}$

$$\mathbb{T}^{n+1}(C^c; x) = \int_{\mathbf{X}} \mathbb{T}^n(C^c; y)\mathbb{T}(dy; x) = \int_C \mathbb{T}^n(C^c; y)\mathbb{T}(dy; x) = 0,$$

which proves reducibility of the chain. □

1.6.2 Geometric ergodicity of $(\zeta_n, \chi_n)_{n\in\mathbb{N}}$

The property of geometric ergodicity is often based on a slightly "weaker" notion of the so-called φ-irreducibility, which is adopted from the previous section to the considered

setting in the following definition.

Definition 1.6.7 (*φ-irreducibility*)**.** The joint chain $(\psi_n, \chi_n)_{n\in\mathbb{N}}$ on a state space S is φ-irreducible for some non-trivial measure φ on S, if for any starting point $(i_0, x_0) \in S$ of the joint process $(\psi_n, \chi_n)_{n\in\mathbb{N}}$, any set $\{i\} \times A \in \mathcal{S}$ with $\varphi(\{i\} \times A) > 0$ can be reached by the chain in a finite number of steps with positive probability.

Theorem 1.6.8. *Let Assumptions 1.3.1 and 1.5.3 hold and $r : I \to \mathbb{R}_+$, as well as*

$$h(i)/r(i) \geq \frac{1}{1-\alpha} \quad \textit{for all } i \in I, \tag{1.6.4}$$

where $\alpha = \max_{i\in I} \mathbb{E}\big[u_0^{r(\psi_n)/\lambda(\psi_n)} \big| \psi_{n-1} = i\big]$ and u_0 is random variables with standard uniform law. Then the joint Markov chain $(\psi_n, \chi_n)_{n\in\mathbb{N}}$ is φ-irreducible for some non-trivial measure φ on $I \times \mathbb{R}$.

Proof. It is clear that $(\psi_n)_{n\in\mathbb{N}}$ is an irreducible finite state space Markov chain with transition probabilities $(p_{\mathrm{rev},\,ij})_{i,j\in I}$ satisfying the following equation

$$p_{\mathrm{rev},\,ij} = \frac{\pi_j}{\pi_i} p_{\mathrm{rev},ji}, \quad \text{for all } i, j \in I.$$

Denote the stationary distribution of the chain $(\psi_n)_{n\in\mathbb{N}}$ as p^*_{rev} and let $\varphi = p^*_{\mathrm{rev}} \otimes \lambda^D_{Leb}$, where D is an open connected set of the form $\big(h(i^*)/r(i^*), h(j^*)/r(j^*)\big)$, with $i^* = \arg\min_{i\in I}\{h(i)/r(i)\}$ (assuming it is unique, if not, we should subtract in the argument of the definition for j^* the set of values $\{i^*\}$ for which the function $h/r(\bullet)$ attains its minimal value) and $j^* = \mathrm{argmax}_{j\in I}\{h(j)/r(j)\}$. In the degenerate case, when $h(i)/r(i) = c$ for all $i \in I$ and some constant $c > 0$, the limiting distribution of $(\chi_n)_{n\in\mathbb{N}}$ degenerates at c. In this case the joint process $(\psi_n, \chi_n)_{n\in\mathbb{N}}$ is clearly irreducible.

Assuming that we do not have a degenerate case, namely $i^* \neq j^*$, we prove φ-irreducibility of the joint process $(\psi_n, \chi_n)_{n\in\mathbb{N}}$. Since the first component of the joint process is irreducible Markov chain, for any starting point i_0 from I and any point i again from I there is a finite $k \in \mathbb{N}$ such that in k number of steps the chain reaches i.

Let $p^{(k)}_{i_0,i} = \mathbb{P}\big(\psi_k = i \big| \psi_0 = i_0\big)$, then $p^{(k)}_{i_0,i} > 0$ and

$$\begin{aligned}
&\mathbb{P}\Big((\psi_k, \chi_k) \in \{i\} \times A \Big| (i_0, x)\Big) = p^{(k)}_{i_0,i} \mathbb{P}\Big(\chi_k \in A \Big| \psi_0 = i_0, \psi_k = i, \chi_0 = x\Big) \\
&= p^{(k)}_{i_0,i} \mathbb{P}\Big(u_k^{\frac{r(i)}{\lambda(i)}} \chi_{k-1} + \big(1 - u_k^{\frac{r(i)}{\lambda(i)}}\big) \frac{h(i)}{r(i)} \in A \Big| \psi_0 = i_0, \chi_0 = x\Big) \\
&= p^{(k)}_{i_0,i} \mathbb{P}\Big(\prod_{i_j \in (i_0,\ldots,i),j} u_j^{\frac{r(i_j)}{\lambda(i_j)}} x_0 + \prod_{i_j \in (i_1,\ldots,i_{k-1}),j} u_j^{\frac{r(i_j)}{\lambda(i_j)}} \big(1 - u_1^{\frac{r(i_1)}{\lambda(i_1)}}\big) \frac{h(i_1)}{r(i_1)} + \ldots \\
&+ \big(1 - u_k^{\frac{r(i)}{\lambda(i)}}\big) \frac{h(i)}{r(i)} \in A \Big| \psi_0 = i_0, \chi_0 = x\Big)
\end{aligned}$$

$$= p_{i_0,i}^{(k)} \mathbb{P}\Big(\prod_{i_j \in (i_0,\dots,i),j} u_j^{\frac{r(i_j)}{\lambda(i_j)}} x_0 + \prod_{i_j \in (i_1\dots,i_{k-1}),j} u_j^{\frac{r(i_j)}{\lambda(i_j)}} \big(1 - u_1^{\frac{r(i_1)}{\lambda(i_1)}}\big) \frac{h(i_1)}{r(i_1)} + \dots$$
$$+ \big(1 - u_k^{\frac{r(i)}{\lambda(i)}}\big) \frac{h(i)}{r(i)} \in A \Big)$$

One can find a path for $(\psi_n)_{n\in\mathbb{N}}$ from i_0 to i through i^* and j^*, so that the support of the random variable $\Big(\prod_{i_j \in (i_0,\dots,i),j} u_j^{\frac{r(i_j)}{\lambda(i_j)}} x_0 + \prod_{i_j \in (i_1\dots,i_{k-1}),j} u_j^{\frac{r(i_j)}{\lambda(i_j)}} \big(1 - u_1^{\frac{r(i_1)}{\lambda(i_1)}}\big) \frac{h(i_1)}{r(i_1)} + \dots + \big(1 - u_k^{\frac{r(i)}{\lambda(i)}}\big) \frac{h(i)}{r(i)}\Big)$ is D. As a result, we can conclude that $\mathbb{P}\big((\psi_k, \chi_k) \in \{i\} \times A \big| (i_0, x)\big) > 0$ proving φ-irreducibility of the joint chain.

□

Since the second component of the process $(\zeta_n)_{n\in\mathbb{N}} = (\psi_n, u_n)_{n\in\mathbb{N}}$ forms an independent and identically distributed sequence of uniform random variables, in order to show that the process $(\zeta_n, \chi_n)_{n\in\mathbb{N}}$ is geometrically ergodic it suffices to show that the joint process $(\psi_n, \chi_n)_{n\in\mathbb{N}}$ is geometrically ergodic.

Proposition 1.6.9. *Consider the joint Markov chain $(\psi_n, \chi_n)_{n\in\mathbb{N}}$, where the first component is a finite state space Markov chain defined on I and the second component follows the dynamics* (1.5.3)*. Let Assumptions 1.3.1 and 1.5.3 hold. Assume that for the functions $r : I \to \mathbb{R}+$, $\lambda : I \to (0, \infty)$, and $h : I \to \mathbb{R}_-$ the following conditions are satisfied*

$$\sum_{j\in I} \frac{p_{rev,\,ij}}{1 + r(j)/\lambda(j)} < 1 \quad \textit{for all } i \in I, \tag{1.6.5}$$

where $(p_{rev,\,ij})_{(i,j)\in I\times I}$ denote the transition probabilities of $(\psi_n)_{n\in\mathbb{N}}$. As well as

$$h(i)/r(i) \geq \frac{1}{1-\alpha} \quad \textit{for all } i \in I, \tag{1.6.6}$$

where $\alpha = \max_{i\in I} \mathbb{E}\big[u_0^{r(\psi_n)/\lambda(\psi_n)} \big| \psi_{n-1} = i\big]$ with u_0 - a random variable with standard uniform law. Then the joint Markov chain $(\psi_n, \chi_n)_{n\in\mathbb{N}}$ is geometrically ergodic.

Remark 1.6.10. The condition (1.6.5) guarantees that the contractive and expansive parts of the second component $(\chi_n)_{n\in\mathbb{N}}$ of the joint process balance each other, that is, the distribution of $u_n^{r(\psi_n)/\lambda(\psi_n)}$ has enough mass on $\{u_n^{r(\psi_n)/\lambda(\psi_n)} < 1\}$.

In the proof of this proposition to follow, we show the existence of geometric drift from Definition 1.6.3 for the process $(\psi_n, \chi_n)_{n\in\mathbb{N}}$, i.e. we construct a petite set C on which the geometric drift condition holds. An explicit proof that the constructed set C is petite requires to analyse higher order transition kernels and is rather involved. Instead, we exploit the property of the uniform countable additivity, see [110].

Definition 1.6.11 (*Uniform Countable Additivity*)**.** Let X be a Markov chain defined on a state space S with transition kernel $\mathbb{T}$. A set $C \in \mathcal{S}$ satisfies the uniform countable additivity condition if for any sequence B_n in $\mathcal{S}$ with $B_n \downarrow \emptyset$ the following limit holds

$$\lim_{n\to\infty} \sup_{x\in C} \mathbb{T}(B_n; x) = 0.$$

Indeed, the uniform countable additivity condition requires only the knowledge of one-step transition probabilities and is simpler to verify. Based on the result of [110], Theorem 3, in order to guarantee that any particular set C is petite for an irreducible Markov chain it suffices to demonstrate that the condition of uniform countable additivity holds for the given set C as well as the following generalised drift condition.

Definition 1.6.12 (*Generalised Drift Towards C*)**.** The drift condition for a Markov chain X on a state space S holds if there exists a real valued measurable function $V : S \to (0, \infty]$ such that for some constant $\beta < \infty$, and a measurable set C,

$$\mathbb{E}[V(X_n)|X_{n-1} = x] - V(x) \leq -1 + \beta \mathbb{I}_C(x), \quad x \in S.$$

Proof. Let $\alpha := \max_{i\in I} \mathbb{E}\big[u_0^{r(\psi_n)/\lambda(\psi_n)} \big| \psi_{n-1} = i\big]$. It follows from (1.6.5) that $\alpha < 1$. Denote $\beta := \max_{i\in I} \frac{h(i)}{r(i)} * \max_{i\in I} \mathbb{E}\big[\big(1 - u_0^{r(\psi_n)/\lambda(\psi_n)}\big) \big| \psi_{n-1} = i\big]$. Introduce $C := I \times [d, c]$ with $c = \frac{1+\beta}{1-\alpha}$ and $d = \frac{1}{1-\alpha}$.

Define the drift function as $V(\psi, \chi) := 1 + \chi$. If $\chi_0 \geq d$ it follows that $\mathbb{I}(\chi_n \geq d) = 1$ for all $n \geq 1$. Also, if $\chi_0 < d$ the process $(\chi_n)_{n\in\mathbb{N}}$ at some point n_0 will become bigger than d and for all $n \geq n_0$ it then holds $\mathbb{I}(\chi_n \geq d) = 1$. So, it is enough to consider the case when $\chi_0 \geq d$. By plugging in the dynamics of χ_n from the recursive equation (1.5.3) and using the above notations of α and β we obtain

$$\begin{aligned}
\mathbb{E}\Big[V(\psi_n, \chi_n)\Big|(\psi_{n-1}, \chi_{n-1}) = (i, x)\Big] &= \mathbb{E}\Big[1 + \chi_n\Big|(\psi_{n-1}, \chi_{n-1}) = (i, x)\Big] \\
&= \mathbb{E}\Big[1 + u_n^{r(\psi_n)/\lambda(\psi_n)} \chi_{n-1} \\
&\quad + \big(1 - u_n^{r(\psi_n)/\lambda(\psi_n)}\big) \frac{h(\psi_n)}{r(\psi_n)} \Big| (\psi_{n-1}, \chi_{n-1}) = (i, x)\Big] \\
&= 1 + \mathbb{E}\Big[u_n^{r(\psi_n)/\lambda(\psi_n)} x + \big(1 - u_n^{r(\psi_n)/\lambda(\psi_n)}\big) \frac{h(\psi_n)}{r(\psi_n)} \Big| \psi_{n-1} = i\Big] \\
&= 1 + x\mathbb{E}\big[u_n^{r(\psi_n)/\lambda(\psi_n)} \big| \psi_{n-1} = i\big] \\
&\quad + \mathbb{E}\Big[\big(1 - u_n^{r(\psi_n)/\lambda(\psi_n)}\big) \frac{h(\psi_n)}{r(\psi_n)} \Big| \psi_{n-1} = i\Big] \\
&\leq 1 + x\mathbb{E}\big[u_0^{r(\psi_n)/\lambda(\psi_n)} \big| \psi_{n-1} = i\big] + \beta \\
&\leq 1 + x\alpha + \beta.
\end{aligned}$$

Next, using the fact that $\alpha = 1 - (1+\beta)/c$ gives

$$\begin{aligned}\mathbb{E}\Big[V(\psi_n,\chi_n)\big|(\psi_{n-1},\chi_{n-1})\Big] &\le 1 + x\Big(1 - \frac{1+\beta}{c}\Big) + \beta \\ &\le 1 + x - x\frac{1}{c} - \beta\frac{x}{c} + \beta \\ &\le 1 + x - x\frac{1}{c} - \beta\frac{x}{c}\mathbb{I}_{C^c}(\psi_{n-1},\chi_{n-1}) - \beta\frac{x}{c}\mathbb{I}_{C}(\psi_{n-1},\chi_{n-1}) + \beta\end{aligned}$$

Further, from the definition of C we know that $C^c = I \times [d,c]^c$, meaning that if $\chi_{n-1} \in C^c$, then we need to consider only the case of $x > c$. As was stated above, once the process χ enters the region $[d,\infty)$, its values never gets smaller than d. Thus, we arrive to the following upper bound in the case of $\mathbb{I}_{C^c}(\psi_{n-1},\chi_{n-1}) = 1$

$$\begin{aligned}\mathbb{E}\Big[V(\psi_n,\chi_n)\big|(\psi_{n-1},\chi_{n-1})\Big] &\le 1 + x - x\frac{1}{c} - \beta\mathbb{I}_{C^c}(\psi_{n-1},\chi_{n-1}) + \beta \\ &= 1 + x - x\frac{1}{c} \\ &= V(i,x) - 1.\end{aligned}$$

In a similar way, when $\mathbb{I}_C(\psi_{n-1},\chi_{n-1}) = 1$, $x \ge \frac{1}{1-\alpha}$ and

$$\begin{aligned}\mathbb{E}\Big[V(\psi_n,\chi_n)\big|(\psi_{n-1},\chi_{n-1})\Big] &\le 1 + x - x\frac{1}{c} - \beta\frac{x}{c}\mathbb{I}_{C}(\psi_{n-1},\chi_{n-1}) + \beta \\ &\le 1 + x - \frac{x}{c}(1+\beta) + \beta\mathbb{I}_{C}(\psi_{n-1},\chi_{n-1}) \\ &= 1 + x - x(1-\alpha) + \beta\mathbb{I}_{C}(\psi_{n-1},\chi_{n-1}) \\ &\le 1 + x - 1 + \beta\mathbb{I}_{C}(\psi_{n-1},\chi_{n-1}) \\ &= V(i,x) - 1 + \beta\mathbb{I}_{C}(\psi_{n-1},\chi_{n-1}),\end{aligned}$$

implying that the generalised drift condition from Definition (1.6.12) holds for the set C defined above.

To prove that the set C is petite, it remains to demonstrate the uniform countable additivity condition is satisfied. For $B_n \in \mathcal{S}$ let $B_n \downarrow \emptyset$. Note that $B_n = (B_n \cap C) \cup (B_n \cap C^c)$ and if $B_n \downarrow \emptyset$, then both $(B_n \cap C) \downarrow \emptyset$ and $(B_n \cap C^c) \downarrow \emptyset$. Next,

$$\begin{aligned}\lim_{n\to\infty}\sup_{(i,x)\in C}\mathbb{T}(B_n;(i,x)) &= \lim_{n\to\infty}\sup_{(i,x)\in C}\mathbb{P}((\psi_{n+1},\chi_{n+1}) \in B_n|(\psi_n,\chi_n) = (i,x)) \\ &\le \lim_{n\to\infty}\sup_{(i,x)\in C}\mathbb{P}((\psi_{n+1},\chi_{n+1}) \in B_n \cap C|(\psi_n,\chi_n) = (i,x)) \\ &+ \lim_{n\to\infty}\sup_{(i,x)\in C}\mathbb{P}((\psi_{n+1},\chi_{n+1}) \in B_n \cap C^c|(\psi_n,\chi_n) = (i,x)) \quad (1.6.7)\end{aligned}$$

For the second component in the above the following holds

$$\begin{aligned}
&\lim_{n\to\infty}\sup_{(i,x)\in C}\mathbb{P}((\psi_{n+1},\chi_{n+1})\in B_n\cap C^c|(\psi_n,\chi_n)=(i,x))\\
&=\lim_{n\to\infty}\sup_{(i,x)\in C}\mathbb{P}\Bigg(\left(\psi_{n+1},\left(u_{n+1}^{r(\psi_{n+1})/\lambda(\psi_{n+1})}x+\left(1-u_{n+1}^{r(\psi_{n+1})/\lambda(\psi_{n+1})}\right)\frac{h(\psi_{n+1})}{r(\psi_{n+1})}\right)\right)\\
&\in B_n\cap C^c|(\psi_n,\chi_n)=(i,x)\Bigg).
\end{aligned}$$

Let $B_n=J_n\times K_n$, where $J_n\subseteq I$ and $K_n\in\mathcal{B}(\mathbb{R})$ for all $n\in\mathbb{N}$. Since $B_n\downarrow\emptyset$, it follows that at least $J_n\downarrow\emptyset$ or $K_n\downarrow\emptyset$. When $J_n\downarrow\emptyset$, the result follows easily. In the case $K_n\downarrow\emptyset$, we have

$$\begin{aligned}
&\sup_{(i,x)\in C}\mathbb{P}\Bigg(\left(\psi_{n+1},\left(u_{n+1}^{r(\psi_{n+1})/\lambda(\psi_{n+1})}x+\left(1-u_{n+1}^{r(\psi_{n+1})/\lambda(\psi_{n+1})}\right)\frac{h(\psi_{n+1})}{r(\psi_{n+1})}\right)\right)\\
&\qquad\in K_n\cap[d,c]^c|(\psi_n,\chi_n)=(i,x)\Bigg)\\
&\leq\sup_{(i,x)\in C}\sum_{j\in I}\mathbb{P}\Bigg(\left(u_{n+1}^{r(\psi_{n+1})/\lambda(\psi_{n+1})}x+\left(1-u_{n+1}^{r(\psi_{n+1})/\lambda(\psi_{n+1})}\right)\frac{h(\psi_{n+1})}{r(\psi_{n+1})}\right)\\
&\qquad\in K_n\cap[d,c]^c\big|(\psi_n,\chi_n)=(i,x),\psi_{n+1}=j\Bigg)p_{ij}\\
&\leq|I|\sup_{(j,x)\in C}\mathbb{P}\Bigg(\left(u_{n+1}^{r(j)/\lambda(j)}x+\left(1-u_{n+1}^{r(j)/\lambda(j)}\right)\frac{h(j)}{r(j)}\right)\in K_n\cap[d,c]^c\Bigg),
\end{aligned}$$

where $|I|$ represents the number of possible states for the Markov chain $(\psi_n)_{n\in\mathbb{N}}$. Note that for any ϵ, there exists $n^*\in\mathbb{N}$ such that

$$\sup_{(j,x)\in C}\mathbb{P}\Bigg(\left(u_{n+1}^{r(j)/\lambda(j)}x+\left(1-u_{n+1}^{r(j)/\lambda(j)}\right)\frac{h(j)}{r(j)}\right)\in K_n\cap[d,c]^c\Bigg)\leq\varepsilon,$$

for all $n\geq n^*$, where $\varepsilon=\epsilon|I|$. Similar arguments can be applied to the first component of (1.6.7). This allows us to conclude that the one-step transition kernel of the joint Markov chain $(\psi_n,\chi_n)_{n\in\mathbb{N}}$ satisfies

$$\lim_{n\to\infty}\sup_{(i,x)\in C}\mathbb{T}(B_n;(i,x))=0.$$

As a result, we conclude that the set C is petite, cf. Theorem 3 in [110].

The next and final step is to check the existence of the geometric drift of $(\psi_n,\chi_n)_{n\in\mathbb{N}}$ towards C, this can be done in similar way to the previous drift condition check. Let

$V_n := V(\psi_n, \chi_n) = \chi_n$, $n \in \mathbb{N}$ then

$$\begin{aligned}
\mathbb{E}\big[V_n\big|\psi_{n-1},\chi_{n-1}\big] - V_{n-1} &\leq \mathbb{E}\Big[u_0^{r(\psi_n)/\lambda(\psi_n)}\Big|\psi_{n-1}\Big]V_{n-1} + E\Big[\Big(1-u_0^{r(\psi_n)/\lambda(\psi_n)}\Big)\frac{h(\psi_n)}{r(\psi_n)}\Big|\psi_{n-1}\Big] \\
&\quad -V_{n-1} \\
&\leq -\Big(1-\alpha\Big)V_{n-1} + \beta\mathbb{I}_C(\psi_{n-1},\chi_{n-1}) + \beta\mathbb{I}_{C^c}(\psi_{n-1},\chi_{n-1}) \\
&\leq \begin{cases} -\Big(1-\alpha\Big)V_{n-1} + \beta\mathbb{I}_C(\psi_{n-1},\chi_{n-1}) & \text{if } (\psi_{n-1},\chi_{n-1}) \in C \\ -(1-\alpha)V_{n-1} + \beta\mathbb{I}_C^c(\psi_{n-1},\chi_{n-1}) & \text{if } (\psi_{n-1},\chi_{n-1}) \in C^c \end{cases} \\
&\leq \begin{cases} -\Big(1-\alpha\Big)V_{n-1} + \beta\mathbb{I}_C(\psi_{n-1},\chi_{n-1}) & \text{if } (\psi_{n-1},\chi_{n-1}) \in C \\ -V_{n-1}\frac{1+\beta}{c} + \beta\mathbb{I}_C^c(\psi_{n-1},\chi_{n-1}) & \text{if } (\psi_{n-1},\chi_{n-1}) \in C^c \end{cases} \\
&\leq \begin{cases} -\Big(1-\alpha\Big)V_{n-1} + \beta\mathbb{I}_C(\psi_{n-1},\chi_{n-1}) & \text{if } (\psi_{n-1},\chi_{n-1}) \in C \\ -1 - V_{n-1}\frac{\beta}{c} + \beta\mathbb{I}_C^c(\psi_{n-1},\chi_{n-1}) & \text{if } (\psi_{n-1},\chi_{n-1}) \in C^c \end{cases} \\
&\leq \begin{cases} -\Big(1-\alpha\Big)V_{n-1} + \beta\mathbb{I}_C(\psi_{n-1},\chi_{n-1}) & \text{if } (\psi_{n-1},\chi_{n-1}) \in C \\ -1 & \text{if } (\psi_{n-1},\chi_{n-1}) \in C^c \end{cases} \\
&\leq -\Big(1-\alpha\Big)V_{n-1} + \beta\mathbb{I}_C(\psi_{n-1},\chi_{n-1}).
\end{aligned}$$

Moreover, since $\alpha < 1$ the coefficient in front of V_{n-1} is positive. Thus, $(\psi_n, \chi_n)_{n\in\mathbb{N}}$ satisfies the geometric drift condition from Definition 1.6.3. We then have that $(\psi_n, \chi_n)_{n\in\mathbb{N}}$ is φ-irreducible, $V(\psi, \chi)$ is everywhere finite and the set C is petite. By Theorem 15.0.1 in [88], we then obtain that $(\psi_n, \chi_n)_{n\in\mathbb{N}}$ is geometrically ergodic.

□

1.7 Central limit theorem

While the estimator $\hat{\pi}$ for the joint law of (Z_0, X_0) enjoys desirable consistency properties, for inference questions we are also in need of distributional properties of $\langle g, \hat{\pi}_n\rangle$, at least asymptotically. A priori, it may not be clear that the limiting distribution is Gaussian, but in fact it is the case and this is the focus of the following result.

Theorem 1.7.1. *Let $(Y_t)_{t\in\mathbb{R}_+}$ be an ergodic continuous-time Markov chain with a finite state space I, $(y_n)_{n\in\mathbb{N}}$ – the associated embedded discrete-time Markov chain. Denote the reversed Markov chain of $(y_n)_{n\in\mathbb{N}}$ as $(\psi_n)_{n\in\mathbb{N}}$ and the transition probabilities of $(\psi_n)_{n\in\mathbb{N}}$ as $(p_{rev,ij})_{(i,j)\in I\times I}$ respectively. Assume that for the functions $r : I \to \mathbb{R}+$ and $\lambda : I \to (0,\infty)$ the following condition is satisfied*

$$\sum_{j\in I} \frac{p_{rev,ij}}{1 + r(j)/\lambda(j)} < 1 \quad \textit{for all } i \in I, \tag{1.7.1}$$

and Assumption 1.3.1 holds for the joint process $(y_n, u_n)_{n\in\mathbb{N}}$*, with* $(u_n)_{n\in\mathbb{N}}$ *being i.i.d. random variables with standard uniform law. As well as*

$$h(i)/r(i) \geq \frac{1}{1-\alpha} \quad \textit{for all } i \in I, \tag{1.7.2}$$

where $\alpha = \max_{i\in I} \mathbb{E}\big[u_0^{r(\psi_n)/\lambda(\psi_n)}\big|\psi_{n-1} = i\big]$*. Then for the empirical measure* $\hat{\pi}$ *on* $\mathcal{S}\otimes\mathcal{B}(\mathbb{R})$*, given by*

$$\hat{\pi}[A] := \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_A(\zeta_i, \chi_i), \qquad A \in \mathcal{S}\otimes\mathcal{B}(\mathbb{R}),$$

a $\sqrt{n}-$*CLT holds for all initial distributions and all* $\mathcal{S}\otimes\mathcal{B}(\mathbb{R})-$*measurable functions* g *whenever* $\mathbb{E}_\pi|g|^{2+\delta} < \infty$ *for some* $\delta > 0$*, i.e.*

$$\sqrt{n}\left(\frac{1}{n}\sum_{n=1}^{n} g(\zeta_i, \chi_i) - \big\langle \pi, g\big\rangle\right) \xrightarrow{\mathrm{d}} \sigma_g \mathcal{N},$$

where $\sigma_g^2 := \mathrm{Var}_\pi\{g(\zeta_0, \chi_0)\} + 2\sum_{i=1}^{\infty} \mathrm{Cov}\{g(\zeta_0, \chi_0), g(\zeta_i, \chi_i)\} \in [0, \infty)$.

Proof. Under assumptions, it follows from Proposition 1.6.9 that the joint process $(\zeta_n, \chi_n)_{n\in\mathbb{N}}$ is geometrically ergodic. The central limit theorem for $(\zeta_n, \chi_n)_{n\in\mathbb{N}}$ follows from Theorem 17.0.1 in [88]. □

Remark 1.7.2. The condition $\mathbb{E}_\pi|g|^{2+\delta} < \infty$ for some $\delta > 0$ in Theorem 1.7.1 can be relaxed for reversible chains to $\mathbb{E}_\pi|g|^2 < \infty$, see [101]. For counterexamples for the general case of non-reversible chains, see [51].

Remark 1.7.3. One can show that for the joint process $(\zeta_n, \chi_n)_{n\in\mathbb{N}}$ not only the law of large numbers holds, but also a strong law in view of the Harris ergodicity of the process.

Corollary 1.7.1. Let g be bounded and $\mathcal{S}\otimes\mathcal{B}(\mathbb{R})$-measurable, and such that $\sup_{s\in S}|g(s,x) - g(s,y)| \leq K|x-y|$ holds for some $K \in [0,\infty)$ and all $(x,y) \in \mathbb{R}^2$. Under the assumptions of Theorem (1.7.1) it holds

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} g(\zeta_i, \chi_i^x) - \big\langle \pi, g\big\rangle\right) \xrightarrow{\mathrm{d}} \sigma_g \mathcal{N} \tag{1.7.3}$$

Remark 1.7.4. Since we assume that the function g is bounded, the assumption $\mathbb{E}_\pi|g|^{2+\delta} < \infty$ for some $\delta > 0$ is automatically fulfilled.

Proof. Following the lines in the proof of Lemma 1.4.7, we can show that

$$\sqrt{n}\Big|\frac{1}{n}\sum_{i=1}^{n} g(\zeta_i, \chi_i^x) - \frac{1}{n}\sum_{i=1}^{n} g(\zeta_i, \chi_i)\Big| \leq \frac{K|x-\chi_0|}{\sqrt{n}} \sum_{i=1}^{\infty}\left(\prod_{0<k\leq i} b(\zeta_k)\right).$$

This implies that the quantity on the left-hand-side above converges $\mathbb{P}$-a.s. to zero as $n \to \infty$, and hence (1.7.3) holds by Slutsky's lemma, with σ_g^2 defined as above. The fact that σ_g stays unchanged can be easily inferred from the fact that for any bounded function g and ergodic process $(\zeta_n, \chi_n)_{n\in\mathbb{N}}$ we have $\sigma_g^2 = \lim_{n\to\infty} \frac{1}{n}\mathbb{E}_\pi\Big(\sum_{i=0}^n \bar{g}(\zeta_i, \chi_i)\Big)^2$, with $\bar{g} = g - \langle \pi, g \rangle$. □

1.8 Illustrative simulations

In this section, we provide an illustrative example which demonstrates the computational efficiency of the MC method over the recursive method. We keep the captions of the figures below to the minimum, providing only the information that was changed with respect to Figure 1.1. Consider

$$Q_0 = \int_0^\infty \exp\left(-\int_0^t r(Y_{u-})\mathrm{d}u\right) h(Y_t)\mathrm{d}t,$$

where $(Y_t)_{t\in\mathbb{R}_+}$ is an ergodic continuous-time Markov chain with finite state space I_1. Define $(\tau_n)_{n\in\mathbb{N}}$ as the jump times of Y (where we set $\tau_0 = 0$), and let $(y_n)_{n\in\mathbb{N}}$ be the associated embedded discrete time Markov chain, defined via $y_n = Y_{\tau_n}$. Let $(\psi_n)_{n\in\mathbb{N}}$ be the time-reversed chain associated to $(y_n)_{n\in\mathbb{N}}$, refer to Section 1.5 for further details, this chain is living then on a finite state space I_1. First, assume that $|I_1| = 4$ and define the transition matrix of $(\psi_n)_{n\in\mathbb{N}}$ as follows

$$P_1 = (p_1(i,j))_{(i,j)\in I\times I} = \begin{bmatrix} 0.1 & 0.2 & 0.2 & 0.5 \\ 0.3 & 0.4 & 0 & 0.3 \\ 0.4 & 0 & 0.4 & 0.2 \\ 0.3 & 0 & 0.2 & 0.5 \end{bmatrix}.$$

Then, the chain $(\psi_n)_{n\in\mathbb{N}}$ is an ergodic discrete-time Markov chain. To find the invariant law of this chain we need to solve the equation $p^* = p^* P_1$, which brings us to the following system of equations

$$\begin{aligned} p^*(1) &= 0.1p^*(1) + 0.3p^*(2) + 0.4p^*(3) + 0.3p^*(4) \\ p^*(2) &= 0.2p^*(1) + 0.4p^*(2) \\ p^*(3) &= 0.2p^*(1) + 0.4p^*(3) + 0.2p^*(4) \\ p^*(4) &= 0.5p^*(1) + 0.3p^*(2) + 0.2p^*(3) + 0.5p^*(4). \end{aligned}$$

Solving this system of equations together with the fact that $\sum_{i\in I} p^*(i) = 1$ gives us the stationary distribution of $(\psi_n)_{n\in\mathbb{N}}$

$$p^* = \Big(\frac{39}{145}, \frac{13}{145}, \frac{33}{145}, \frac{60}{145}\Big).$$

This invariant distribution tells us that the process $(\psi_n)_{n\in\mathbb{N}}$ spends around 40% of its time in the state 4. For the comparison purposes we also consider one more example of the process $(\psi_n)_{n\in\mathbb{N}}$, which resides on a smaller state space I_2 with $|I_2| = 3$ and the transition matrix is defined in the following way

$$P_2 = (p_2(i,j))_{(i,j)\in I\times I} = \begin{bmatrix} 0.5 & 0.25 & 0.25 \\ 0.5 & 0 & 0.5 \\ 0.25 & 0.25 & 0.5 \end{bmatrix},$$

and the corresponding p^* is given by

$$p^* = \Big(0.4, 0.2, 0.4\Big).$$

Next, we introduce the process $(\chi_n)_{n\in\mathbb{N}}$ in the same way as it was done in Section 1.5, i.e. following the dynamics of (1.5.3) with the corresponding functions defined as $\lambda = 1$, $r(i) = 1$, $h(i) = i$ for all $i \in I$, where I can be either I_1 or I_2, depending on where the considered chain $(\psi_n)_{n\in\mathbb{N}}$ resides.

$$\chi_n = u_n^{r(\psi_n)/\lambda(\psi_n)}\chi_{n-1} + (1 - u_n^{r(\psi_n)/\lambda(\psi_n)})\frac{h(\psi_n)}{r(\psi_n)} = u_n\chi_{n-1} + (1-u_n)\psi_n.$$

The goal then becomes to estimate the joint distribution of $(\chi_n, \psi_n)_{n\in\mathbb{N}}$. We are interested in the estimation of the distribution function $\mathbb{P}(\chi_n \le x|\psi_0 = 1, \chi_0 = 0)$, $n \in \mathbb{N}$, and we do it in two ways. First, by using the MC method and then, the recursive approach, which is based on the calculation of the conditional probabilities $F_n(x;i) = \mathbb{P}\left[\chi_n \le x|\psi_n = i\right]$, introduced in Section 1.5

$$\begin{aligned} F_n(x;i) &= \frac{\mathbb{P}\left[\chi_n \le x, \psi_n = i\right]}{\mathbb{P}\left[\psi_n = i\right]} = \frac{1}{p_i^*}\sum_{j\in I}\mathbb{P}\left[\chi_n \le x, \psi_n = i|\psi_{n-1} = j\right]p_j^* \\ &= \sum_{j\in I}\mathbb{P}\left[\chi_n \le x|\psi_n = i, \psi_{n-1} = j\right]\frac{\mathbb{P}\left[\psi_n = i|\psi_{n-1} = j\right]p_j^*}{p_i^*} \\ &= \sum_{j\in I}\mathbb{P}\left[\chi_n \le x|\psi_n = i, \psi_{n-1} = j\right]p_{ij} \\ &= \sum_{j\in I}\mathbb{P}\left[\chi_{n-1} \le u_n^{-r(i)/\lambda(i)}x - (u_n^{-r(i)/\lambda(i)} - 1)\frac{h(i)}{r(i)}\ \Big|\psi_{n-1} = j\right]p_{ij} \end{aligned}$$

$$= \sum_{j \in I} p_{ij} \int_0^1 F_{n-1} \left[u^{-r(i)/\lambda(i)} x - (u^{-r(i)/\lambda(i)} - 1) \left(h(i)/r(i) \right) ; j \right] \mathrm{d}u.$$

For the simulations we use the recursive representation of the conditional laws above with a fixed grid size $\delta = 1/100$ and then compare the results to the MC simulation, for which we compute the empirical distribution along a single path of a given length N. In this simulation study we consider $N = 400$, 2000, 5000 as well as the number of iterations used for the recursive method is $N_{rec} = 4$, 20. We also calculate the Kolmogorov-Smirnov distances between the empirical and true distributions in both methods. The error is a fixed number for each recursive implementation, while in the case of the MC simulation we produce the distribution for the errors by running it over 50 times.

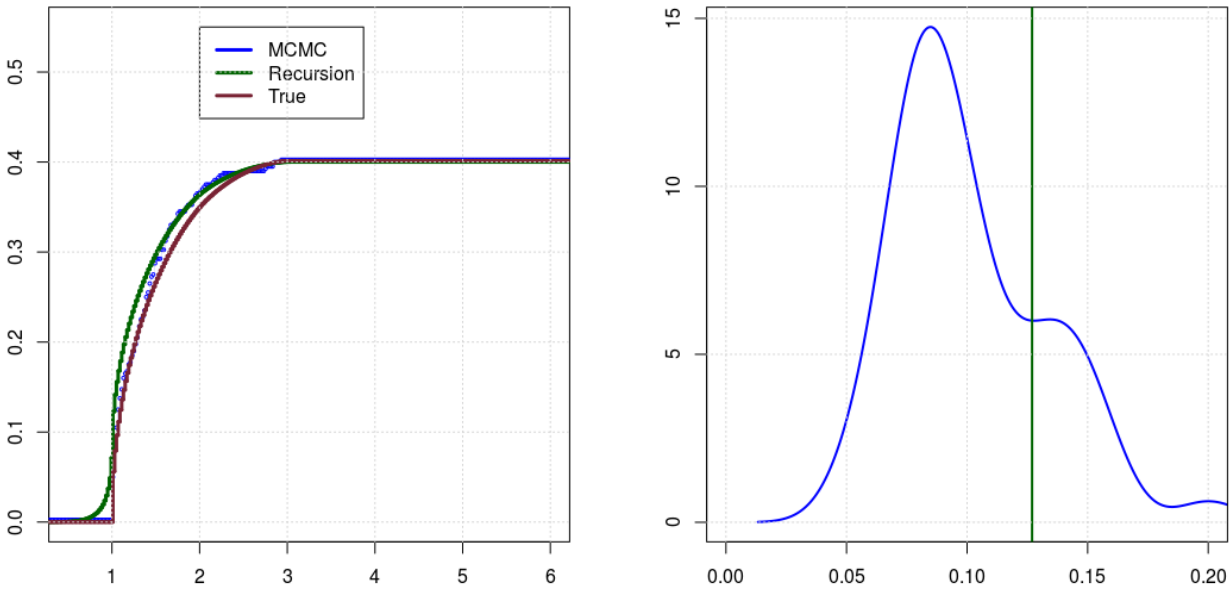


Figure 1.1: On the left: two estimators $\mathbb{P}^{rec}_{est}(\chi_n \leq x | \psi_0 = 1, \chi_0 = 0)$ (green), $\mathbb{P}^{MCMC}_{est}(\chi_n \leq x | \psi_0 = 1, \chi_0 = 0)$ (blue) of the distribution $\mathbb{P}(\chi_n \leq x | \psi_0 = 1, \chi_0 = 0)$ as well as the true distribution itself (purple line). For the recursive method n ranges from 1 to 4, for the MC – from 1 to 400. The y-axis corresponds to the values of $\mathbb{P}^{rec}_{est}$, $\mathbb{P}^{MCMC}_{est}$, and $\mathbb{P}^{rec}_{est}$ respectively, while x-axis represents different values of x. Further, $I = I_2$ and $N = 400$, $N_{rec} = 4$.
On the right: Kolmogorov-Smirnov distances between the empirical and the true distribution. The x-axis corresponds to the error values, while y-axis represents the values of the density function of the errors produced by the MCMC implementation.

Table 1.1: Computational time for the recursive method

Number of iterations	Time in seconds, $\lvert I \rvert = 3$	Time in seconds, $\lvert I \rvert = 4$
4	159	366
20	803	1754

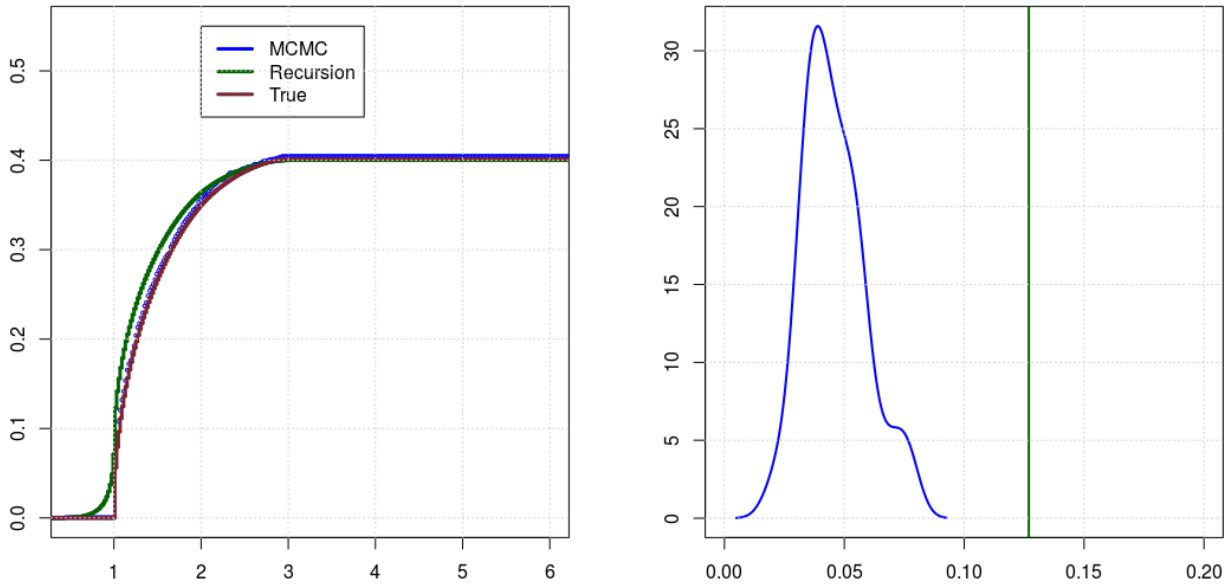


Figure 1.2: $N = 2000$ and $N_{rec} = 4$, $I = I_2$.

Table 1.2: Computational time for the MC method

The length of the path	Time in seconds, $\lvert I\rvert = 3$	Time in seconds, $\lvert I\rvert = 4$
400	0.087	0.110
2000	0.44	0.542
5000	1.083	1.335

First, consider the case when the chain $(\psi_n)_{n\in\mathbb{N}}$ lives on the three state space I_2. We can observe from the right plot in Figure 1.1, where $N = 400$ and $N_{rec} = 4$, that the errors are visually of the same magnitude, however, the recursive method takes about 150 seconds to run with only 4 iterations, while the MC achieves the same quality in about 0.087 seconds, refer to Table 1.1 and 1.2 for details. Moreover, the MC method captures the shape of the true distribution quite early, whereas the estimator from the recursive method is quite "smooth". In particular, this entails the biggest spread for the recursive method at the point 1 in this example. The discretisation technique for recursions can slightly be optimised to achieve the same performance, but it still lags behind the MC method. Figure 1.2 depicts a significant improvement in the accuracy of the MC method with $N = 2000$. If we increase the number of iterations for the recursive method to $N_{rec} = 20$ and take $N = 5000$, see Figure 1.3, the results are close in the accuracy, however, it takes more than 13 minutes for the recursive implementation to run and only around 1 second for the MC method.

When the Markov chain $(\psi_n)_{n\in\mathbb{N}}$ resides on the four state space I_1, the recursive method takes even longer than twice the time it was taking when $I = I_2$. Whereas, the MC

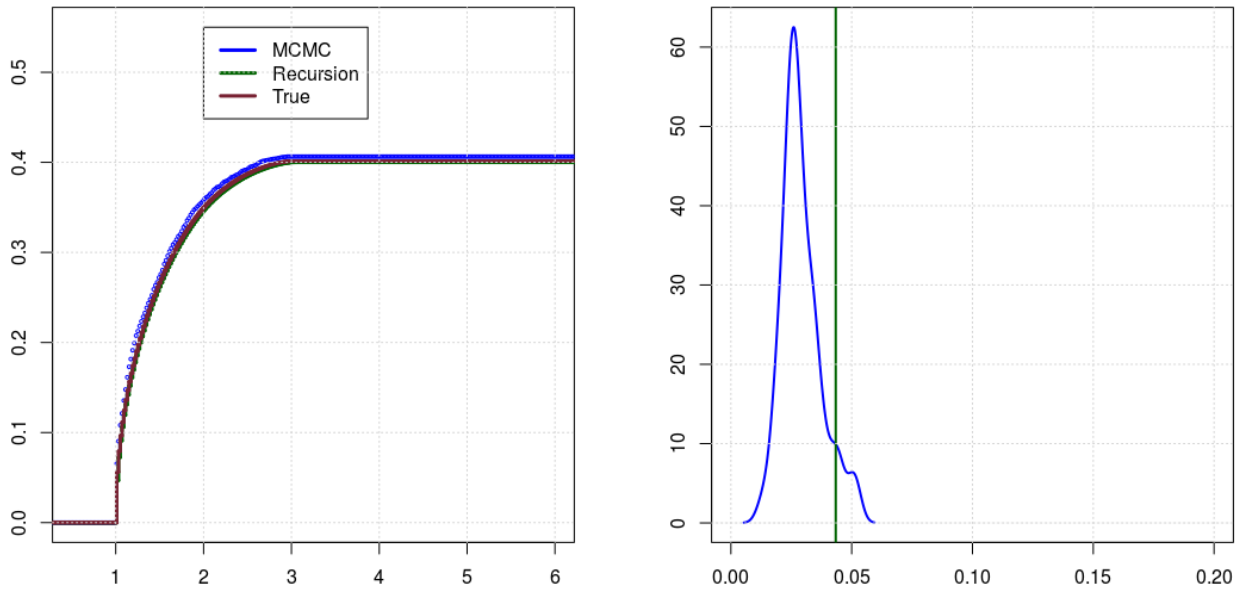


Figure 1.3: $N = 5000$ and $N_{rec} = 20$, $I = I_2$.

simulation is still running considerably fast, see Table 1.2. Also, we observe very similar patterns in terms of the behaviour of the accuracy for both methods, with the results being close to the previous example, when $I = I_2$, see Figures 1.4, 1.5, and 1.6 for more details. Further, we can observe that the blue line starts to capture the true distribution quite early as before.

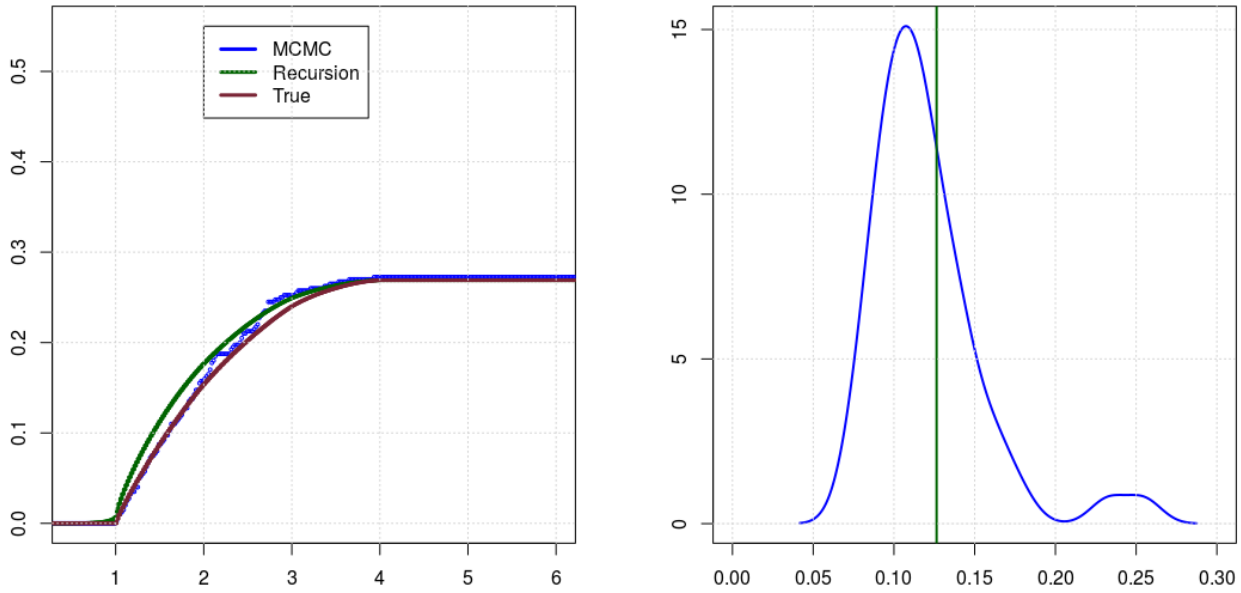


Figure 1.4: $N = 400$ and $N_{rec} = 4$, $I = I_1$.

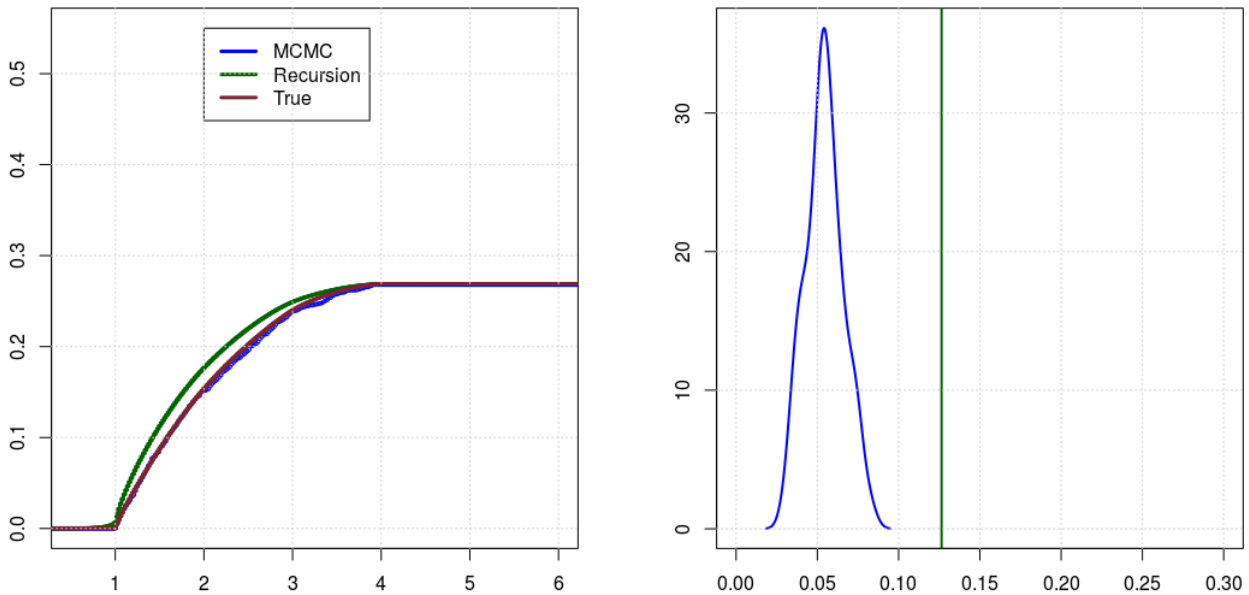


Figure 1.5: $N = 2000$ and $N_{rec} = 4$, $I = I_1$.

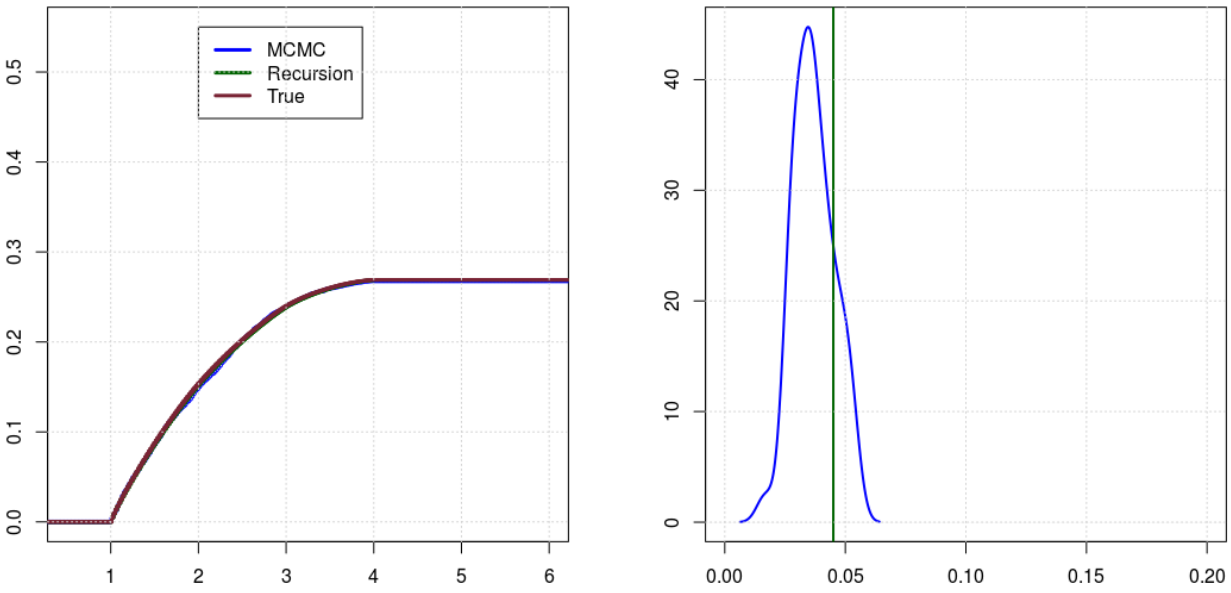


Figure 1.6: $N = 5000$ and $N_{rec} = 20$, $I = I_1$.

Chapter 2

Optimal estimation of functionals in multivariate case

2.1 Introduction to efficient estimation of linear functionals

When working in a discrete-time setting, one can define a regression model in one-dimensional case in the following way

$$\mathscr{Y}_i = g(i/n) + \varepsilon_i, \quad i = 1, \dots, n, \tag{2.1.1}$$

where $\mathscr{Y}_i \in \mathbb{R}$, g is a smooth function and (ε_i) are some random variables that are uncorrelated with mean zero and variance $\sigma^2 > 0$. This type of regression models is well studied in the literature, especially when the error terms are independent and identically distributed random variables, refer to [109]. Consider the discrete time functionals of the following form

$$\vartheta_n = \frac{1}{n}\sum_{i=1}^{n} g(i/n)\omega(i/n), \quad \text{with } \omega : [0,1] \to \mathbb{R} \tag{2.1.2}$$

then, it can be estimated by $\hat{\vartheta}_n = \frac{1}{n}\sum_{i=1}^{n} \mathscr{Y}_i\omega(i/n)$ with no bias and variance of $\frac{\sigma^2}{n^2}\sum_{i=1}^{n}\omega(i/n)^2$.

The corresponding model to (2.1.1) when the errors are uncorrelated with mean zero and variance $\sigma^2 > 0$ in continuous time can be written in the following way

$$dY_t = g(t)dt + \sigma n^{-1/2} dW_t, \quad t \in [0,1], \tag{2.1.3}$$

where W is a Brownian motion on $[0,1]$, the function g is unknown on $[0,1]$ and such that

$g \in L^2([0,1])$, $n \in \mathbb{N}$, see [109] for more details on the connection between Gaussian white noise model (2.1.3) and nonparametric regression (2.1.1). Then, the plug-in estimator $\hat{\vartheta} = \int_0^1 \omega(t)dY(t)$ is an unbiased estimator for the following functional

$$\vartheta = \int_{[0,1]} g(x)\omega(x)dx, \tag{2.1.4}$$

for some $\omega \in L^2([0,1])$. The corresponding variance of the estimator $\hat{\vartheta}$ is given by $\frac{\sigma^2}{n}\int_0^1 \omega(t)^2 dt$.

In this thesis we are interested in a regression model of type (2.1.1) in a multivariate case with one-sided error terms, i.e. when the error terms live on $[0,\infty)$. This type of models is used to estimate a function g when it is a frontier or boundary function of the observations, see [18, 37, 75] for some overview. As such, we assume that for independent and identically distributed error terms $(\varepsilon_i)_{i\in\mathbb{N}}$ the following always holds

$$\varepsilon_i \geq 0, \quad i = 1, \ldots, n.$$

Further, for a mean regression model with one-sided errors we also consider its continuous-time analogous, which is given by observing a Poisson point process (PPP) on $[0,1]^d \times \mathbb{R}$. We introduce these models in Section (2.3.3) properly.

2.1.1 Former results

Estimation of linear functionals of the regression function g is a popular topic in the statistical literature, see [20, 78, 90, 91, 105, 109]. When the error terms ε in the regression model (2.1.1) are uncorrelated random variables with expected value zero and constant variance, the linear functionals (2.1.2) can be estimated by plugging in $\mathscr{Y}_i$ into (2.1.2), i.e. as

$$\hat{\vartheta}_n = \frac{1}{n}\sum_{i=1}^{n} \mathscr{Y}_i \omega(i/n),$$

for $i = 1, \ldots, n$. This estimator is an unbiased estimator with the variance given by $\frac{\sigma^2}{n^2}\sum_{i=1}^n \omega^2(i/n)$. Furthermore, by the Gauß-Markov theorem, see [47] page 146, it has minimal variance among all linear and unbiased estimators. This property of estimators is referred to as BLUE in the statistical literature, meaning that it is the best linear unbiased estimator.

In the continuous-time model (2.1.3), in order to estimate the functional (2.1.4) one can construct a corresponding plug-in estimator as

$$\hat{\vartheta} = \int_{[0,1]} \omega(t)dY_t,$$

which is also unbiased. The variance of this estimator is $\frac{\sigma^2}{n}\int_{[0,1]}\omega(t)^2dt$. By the Riesz representation theorem any linear functional g that is L^2-continuous can be estimated with the parametric rate $n^{-1/2}$, see [74] for a comprehensive summary of this and other related results. Interestingly, as it was noted in [74], an optimal estimator of the regression function itself does not lead to an optimal estimator of the linear functional.

The independence between error terms is a standard assumption for regression models. This condition of independence is of high importance for frontier models when the statistical procedures such as bootstrap are applied, refer to [107, 115] for more detailed information. In the estimation of linear functionals under one-sided errors we assume the independence of the error terms as well. Regression models with one-sided errors underpin many applications where the support rather than the mean properties of the noise are known and where the regression function g describes a frontier or boundary curve. These models arise naturally in analysis of auctions and records, image analysis, or in extreme value analysis with covariates, see, for example, [114]. In [75] and [41] the authors consider a variety of boundary estimation problems with a focus on applications in image recovery. In the prototypical case of $(\varepsilon_i)_{i\in\mathbb{N}}$ are i.i.d. with $\varepsilon_i \geq 0$ and for some intensity parameter $\lambda > 0$

$$\mathbb{P}(\varepsilon_i \leq x) = \lambda x + O(x^2), \quad i = 1, \dots, n, \tag{2.1.5}$$

as $x \downarrow 0$, e.g. $\varepsilon \sim Exp(\lambda)$, the parametric rate for the location model, i.e. the model that assumes g to be a constant function, is n^{-1}, and it is quicker than the regular case. The irregular statistical models are extensively studied in the literature as well. In [5] the authors provide a rate-optimal estimator via the extreme value statistics $\min_i \mathscr{Y}_i$.

When we work in continuous-time settings, we use a PPP model for estimating linear functionals of the form (2.1.4). This idea appears to be quite novel and extremely fruitful, see [62, 87]. Further, in [39] the authors propose a projection based estimator for g together with derived convergence rates and limit distributions. In [87] a regression model with a known bounded support of the errors is considered. The authors show asymptotic equivalence in the strong LeCam sense to a continuous-time PPP model when the error density has a jump at the endpoint of its support.

Both regression-type model with one-sided errors and the PPP model play a major role for the density support estimation or image boundary recovery problems. We also refer to [74] for estimation of the boundary or the frontier function of the support when the regression function belongs to a certain Hölder class. The authors introduce a way to obtain the rate of $n^{-(\beta+1/2)/(\beta+1)}$ for the functional of the form $\int_0^1 g(x)dx$, where β is a constant called Hölder exponent. The derived upper bound is based upon a localisation step and loses a logarithmic factor. They also study the case of monotone or convex regression functions. Estimation of sets with smooth boundaries is considered in [86]. In [35] an estimator of the density support area based on threefold sample splitting is constructed,

and it achieves the minimax optimal rate. For a review on the recent developments in the density support estimation, see [14]. A significant amount of research is dedicated to asymptotic results of the estimation of the expected area of a convex hull. The results for the i.i.d. case can be found in [99]. In [38] the authors introduce a linear programming approach and the corresponding estimators are analysed asymptotically. The work of [86] introduces the connections to the classification problems. In [73] the authors, when estimating the density support set in Hausdorff distance, calculate the asymptotically exact risk. For a regression model when the distribution of errors is one-sided and regularly varying at 0 the authors in [62] propose an estimator for the boundary regression function which adapts to the unknown smoothness of the function.

2.1.2 Organisation of the chapter

The next section collects the main results of this chapter. Section 2.3 provides an introduction into the theory of point process models together with a detailed overview of Poisson point processes. We introduce the main definitions as well as discuss their main applications. In Section 2.3.3, we formulate two statistical models which are the focus of this chapter, a PPP model and a multidimensional regression model with one-sided errors. Section 2.4 is devoted to block-wise estimation methods, when the estimator is an aggregate of local estimators constructed on subsets of $[0,1]^d$. As a first step, in Section 2.4.1 we construct an estimator for the PPP model by using martingale theory and then extend the results to the one-sided regression model in Section 2.4.2. The proposed block-wise estimators estimate the functionals with no bias and at the minimax optimal rate n, which is established in Section 2.4.4. The proof of the minimax optimal rate is based on a Bayesian risk bound, which provides a lower bound for the minimax risk, see [74] for similar approaches. In Section 2.5, we study a nonparametric maximum likelihood approach in the PPP model. Using the fact that a maximum likelihood estimator for the function g is a complete and sufficient statistics, shown in Proposition 2.5.1, we derive an unbiased estimator of minimal variance for the linear functional. Further, we discuss a central limit theorem in Section 2.5.2. In Section 2.5.3, we provide numerical results for the validity of the central limit theorem for the nonparametric maximum likelihood estimator in the PPP model. We illustrate both one-dimensional and two-dimensional scenarios numerically in Section 2.5.3 and Section 2.6. In Section 2.7.1 we provide some auxiliary technical lemmata. We also collect some foundational results from statistical theory that are used throughout this chapter in Section 2.7.2

2.2 Main contributions

The main goal of this chapter is to develop new methods of estimating linear functionals (2.1.2) and (2.1.4) in multidimensional case. As the main contributions we list the following

- We rigorously introduce two statistical models for estimating linear functionals in (2.1.2) and (2.1.4). We first study the PPP model and then the one-sided regression model.
- We construct two types of estimators for each of the models, a block-wise estimator and a maximum likelihood estimator (MLE).
- We further study the statistical properties of the provided estimators. We prove that the block-wise estimator is an unbiased estimator for linear functionals in discrete- and continuous-time settings. We also calculate an upper bound for the variance of the estimator. In the PPP model we provide the minimax optimal rate, and we show that the MLE is UMVU (uniformly of minimum variance among all unbiased estimators).
- Finally, we provide numerical illustrations for the results with simulated data.

2.3 Introduction and background

We start by introducing point process models and discussing a wide range of their applications in statistical modelling. We then focus on the PPP model and also recall the foundational definitions and results from nonparametric statistics which are used throughout the chapter.

2.3.1 Point process models

A point process is a model of indistinguishable points distributed randomly in a space. This space can be quite general, but usually set to be $\mathbb{R}^d$ with $d \geq 1$. In dimensions $d > 1$, point processes are usually referred to as spatial point processes. In the one-dimensional case, there typically exists a natural ordering for the underlying domain of the process, such as for time, unlike in higher dimensions $d > 1$, where ordering doesn't typically exist. A toy example of a one-dimensional point processes is the number of emergency phone calls received at a police station at times T_i, which can be modelled by a Poisson process. A one-dimensional point process can also be seen as a counting process. As such, one can

define for all $t > 0$ (with a slight abuse of notation)

$$\begin{aligned} N_t &= \text{number of points arriving up to time } t \\ &= \sum_{i=1}^{\infty} \mathbb{I}\{T_i \leq t\}. \end{aligned}$$

For special cases of point processes, such as the Poisson process, the increments $N_{t_j} - N_{t_i}$, $0 \leq t_i \leq t_j$ are independent random variables for disjoint intervals. While there is no natural ordering of the points in a multidimensional space, one can generalise the interval counts $N_{t_j} - N_{t_i}$, $0 \leq t_i \leq t_j$ to the subset counts. For a bounded and closed set $A \subseteq \mathbb{R}^d$ we define the following random variable

$$N(A) = \text{number of points falling inside } A.$$

Similarly to the one-dimensional case, for special cases of spatial processes, such as the Poisson point process, whenever the sets A and B are disjoint the random variables $N(B)$ and $N(A)$ are independent.

As the next step, we rigorously formulate the definition of a point process. Let $\mathbb{E}$ in $\mathbb{R}^d$ denote a locally compact Hausdorff space whose topology has a countable base (usually $\mathbb{R}^d$) as a state space and denote by $\mathcal{E}$ its Borel σ-algebra. Let $\mathcal{B}$ be the family of bounded (relatively compact) sets in $\mathcal{E}$. On $(\mathbf{E}, \mathcal{E})$ we define the set of random measures $\mathfrak{M} = \{m \text{ measure on } \mathbf{E} \,:\, m(A) < \infty \; \forall A \in \mathcal{B}\}$ and the set of point measures $\mathfrak{M}_p = \{m \text{ measure on } \mathbf{E} \,:\, m(A) \in \mathbb{N}, \; \forall A \in \mathcal{B}\}$. The σ-algebra on the set $\mathfrak{M}$ can be easily defined in terms of sets as $\mathcal{M} = \sigma(m \mapsto m(A), A \in \mathcal{E})$. In particular, $\mathcal{M}$ is countably generated. The set $\mathfrak{M}_p$ belongs to $\mathcal{M}$, and we define $\mathcal{M}_p = \mathcal{M} \cap \mathfrak{M}_p$.

Definition 2.3.1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be an abstract probability space.
(a) A random measure on $\mathbf{E}$ is a measurable mapping M of $(\Omega, \mathcal{F})$ into $(\mathfrak{M}, \mathcal{M})$;
(b) A point process on $\mathbf{E}$ is a measurable mapping N of $(\Omega, \mathcal{F})$ into $(\mathfrak{M}_\mathfrak{p}, \mathcal{M}_p)$.

A point process N is a random distribution of points in $\mathbf{E}$. Let $N(A)$ be the number of points in A, then, as was mentioned above, there exists a representation of a point process N as the sum of Dirac delta functions δ_{X_i}

$$N = \textstyle\sum_{i=1}^{K} \delta_{X_i},$$

where δ_x is the point mass at x, X_i are measurable mappings of $(\Omega, \mathcal{F})$ into $(\mathbf{E}, \mathcal{E})$, and K is a random variable with values in $\{0, 1, ..., \infty\}$. As such, a point process can be described as the measure that allocates a unit mass a random set of points in $\mathbf{E}$. Furthermore, one can also use finite-dimensional distributions, vacancy probabilities, capacity functional, or generating function for mathematical construction and characterisation of point processes,

refer to [36] for the range of equivalence theorem for spatial processes. However, an easier and more straightforward way is to construct a point process by transforming an already existing point process by thinning, superposition, or clustering, see [108].

Point processes are well studied in probability theory and are mainly used to model and analyse spatial data, refer to [89]. They are applied in a variety of fields starting from epidemiology, and computational neuroscience and finishing with astronomy and agriculture, see [36]. They are also used in applications to machine learning and pattern recognition, however this direction is relatively new, see [12, 17, 26, 76, 113]. Point process models show some links with random fields that are often applied in pattern recognition, for example conditional and Markov random fields, for further detail refer to [6].

2.3.2 Poisson point process

One of the main examples of point processes is the so-called PPP. Poisson point processes play an important role in stochastic analysis. We start by providing the definition of PPP using the notation from the previous section.

Definition 2.3.2. Let μ be an element of $\mathfrak{M}$. A point process N on $\mathbf{E}$ is a Poisson process with intensity measure μ if:

1. For all mutually disjoint sets $A_1, ..., A_n \in \mathcal{E}$, the random variables $N(A_1), ..., N(A_n)$ are independent.

2. For any $A \in \mathcal{E}$, we have $N(A) \sim \text{Poiss}\big(\mu(A)\big)$.

Thus, N has independent increments and for each A the random number of points in A has a Poisson distribution with mean $\mu(A)$. When $\mathbf{E} = R^d$ and $\mu(A) = \lambda\big|A\big|$, where $\big|A\big|$ is the Lebesgue measure of the set A, the process defined above is called a homogeneous Poisson process and λ is known as the intensity of the process. This process is well-defined as it follows from the following theorem.

Theorem 2.3.3. *The Poisson point process exists and is uniquely determined by its intensity measure.*

The aforementioned definition and proof of this result can be found, for example, in [21, 36, 66, 77].

Applications of the Poisson point process

The tractability of the PPP has been acknowledged in many statistical applications. A vast number of realistic models of physical processes give rise to a PPP (often relying on the Poisson limit theorem), see [4, 59, 83], some of which we shortlist here.

1. **Spatial birth-and-death process.** Consider a dynamical system seen as a collection of motionless points in some space at each moment of time. We interpret the points as particles, or individuals. Each particle is characterised by its location. Existing particles may die within time interval Δt with a probability $\alpha\Delta t$. The rate α is called a mortality rate per particle per unit time. Further, in any small region of area Δs in the same time interval new particles may appear with probability $g\Delta s\Delta t$. Independently of the initial state of the system over a long period of time the spatial birth-and-death process arrives to an equilibrium state. At this equilibrium state the pattern of the system is a realisation of a Poisson process with intensity function g/α, for further detail on spatial birth-and-death processes and a proof of the result, see [34, 97].

2. **Random strewing.** Consider a high number of points N that are being scattered randomly in a region E according to a binomial point process, implying that the points are independent of each other and also uniformly distributed over this region E. If we observe this pattern in a subset S, i.e. $S \subseteq E$ and such that the area of S is much smaller than the area of E, then the observed pattern approximately corresponds to a PPP inside S.

3. **Random displacement.** Consider a system represented by a point process X. Suppose that each point of the process X is subjected to a random displacement independently of other points, i.e. a point x_i is moved from its original location to a new location $y_i = x_i + a_i$ where a_i is a random vector. Assuming that a_i are independent random vectors, then, under some additional assumptions, one can show that the resulting point process Y is approximately an inhomogeneous Poisson point process, see [84].

2.3.3 Statistical models and estimation problems

We are now ready to introduce two statistical models for the purpose of efficient estimation of linear functionals of the form (2.1.2) and (2.1.4) respectively which are used throughout this chapter. The PPP model in continuous-time and the regression-type model in discrete-time.

- **Poisson point process model** given by observing a PPP on $[0,1]^d \times \mathbb{R}$ of intensity

$$\lambda_g(x,y) = n\mathbf{1}(y \geq g(x)), \quad x \in [0,1]^d, \;\; y \in \mathbb{R}, \tag{2.3.1}$$

  where g is a smooth function to be specified further.

  See Figure 2.1 for an illustration of the PPP model in one-dimensional case.

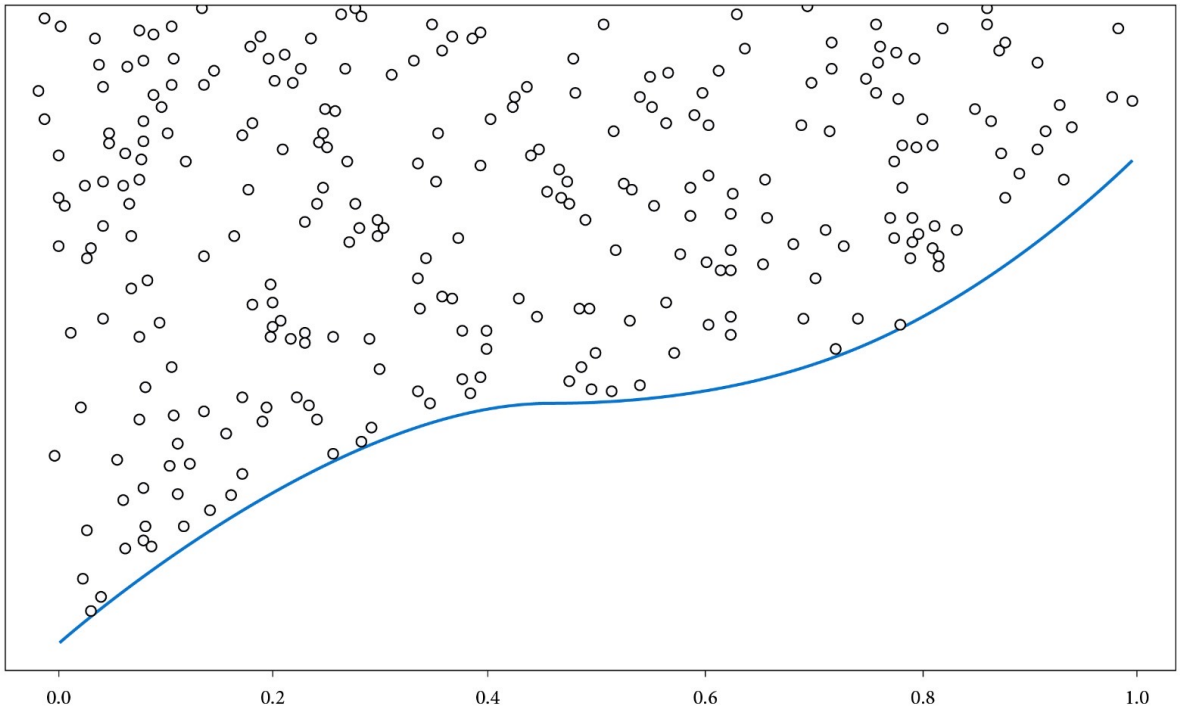


Figure 2.1: PPP model. The circles indicate the PPP observation points (X_i, Y_i) with intensity (2.3.1), the blue line is the function g. The x-axis corresponds to the values of (X_i) and y-axis represents the values of (Y_i).

- **Multidimensional regression model with one-sided errors**

$$\mathscr{Y}_i = g(\mathbf{X}_i) + \varepsilon_i, \quad \mathscr{Y}_i \in \mathbb{R},$$

where we assume that $(\mathbf{X}_i)_{1\le i\le n}$ are spread deterministically over $[0,1]^d$. We further consider the equidistant case, where the data points $(\mathbf{X}_i)_{1\le i\le n}$ form a grid in the following way

$$\begin{pmatrix} n^{-1/d} \\ n^{-1/d} \\ \vdots \\ n^{-1/d} \end{pmatrix}, \begin{pmatrix} 2n^{-1/d} \\ n^{-1/d} \\ \vdots \\ n^{-1/d} \end{pmatrix}, \cdots, \begin{pmatrix} 1 \\ \vdots \\ 1 \\ (n^{1/d}-1)n^{-1/d} \end{pmatrix}, \begin{pmatrix} 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix},$$

with $n^{-1/d} \in \mathbb{N}$. In the one-dimensional case the above becomes $X_i = i/n$ with $i = 1, \ldots, n$. The error terms (ε_i) are assumed to be supported on $[0, \infty)$. The prototypical case is that (ε_i) are independent and identically distributed with $\varepsilon_i \ge 0$ and for some intensity parameter $\lambda > 0$

$$\mathbb{P}(\varepsilon_i \le x) = \lambda x + O(x^2), \quad i = 1, \ldots, n, \tag{2.3.2}$$

as $x \downarrow 0$, e.g. $\varepsilon \sim Exp(\lambda)$.

For an illustrative example of one-sided regression model also in one-dimensional case see Figure 2.2 (where all the observation points $(\mathbf{X}_i, \mathscr{Y}_i)$ lie above the function g). One can see that in this model the points are concentrated mainly near the function g on its upper side.

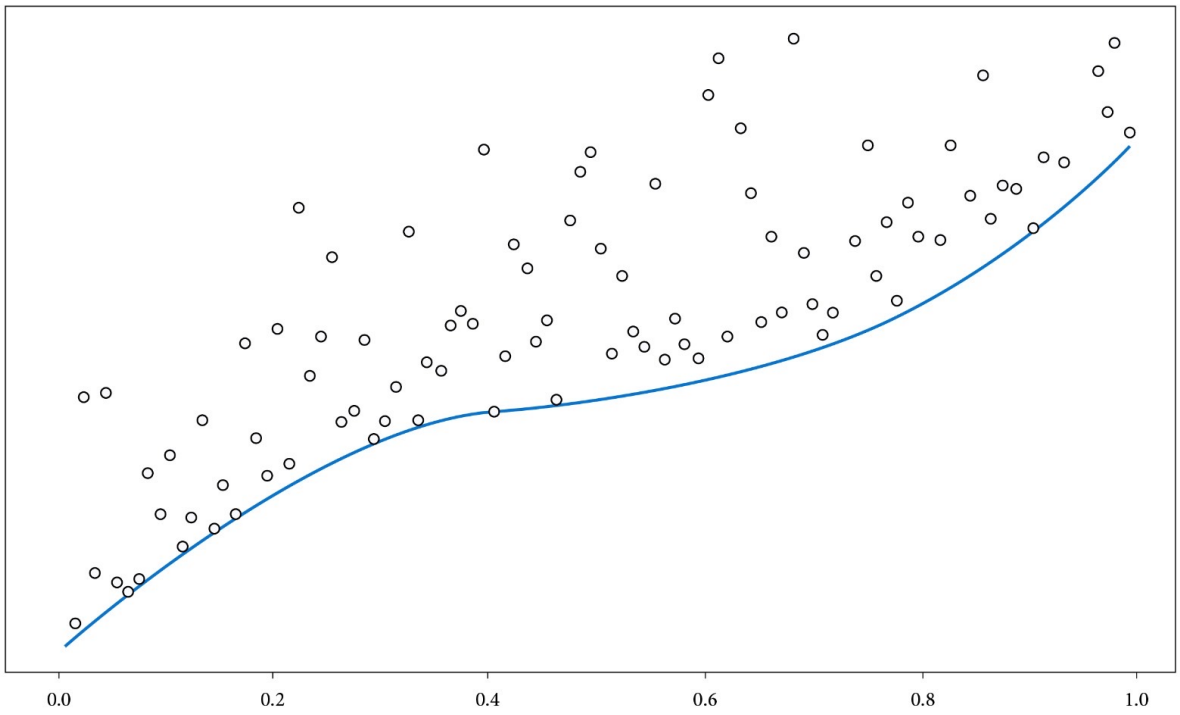


Figure 2.2: One-sided regression model. The circles indicate the observation points $(X_i, \mathscr{Y}_i)$ and the blue line is the function g. The x-axis corresponds to the values of (X_i) and y-axis represents the values of $(\mathscr{Y}_i)$.

One of the main differences between the PPP model (2.3.1) and regression model with one-sided errors (2.3.2) can be seen when doing statistical inferences. A classical approach to the inference is likelihood estimation. It clearly differs for stochastic processes, in particular, for point processes, when the number of observations is a random variable, from estimation procedures for classical inference based on a fixed sample of data.

2.3.4 Technical assumptions

Finally, let us specify the family of probability measures $\mathcal{P} = \{\mathbb{P}_g, g \in G\}$, in our setting. From now on we shall assume that the regression function g in both models (2.3.3) and (2.3.1) belongs to the Hölder class.

Definition 2.3.4. (Hölder property). Let $\beta \in (0, 1]$ and $R > 0$ be two positive numbers. The Hölder class $C^\beta(R)$ is defined as the set of all continuous functions $f : [0, 1]^d \to \mathbb{R}$

satisfying

$$|f(x) - f(y)| \leq R \parallel x - y \parallel_{L^2}^{\beta}, \quad \forall x, y \in [0,1]^d.$$

Remark 2.3.5. The variable β is referred to as the exponent of the Hölder condition. If a function on an interval satisfies the Hölder property with $\beta > 1$ then it is constant. In case of $\beta = 1$ the function satisfies the so-called Lipschitz condition. For any other values of β, the Hölder property implies that the function is uniformly continuous.

The full list of notations can be found in List of Notation, page 13. However, throughout this chapter we are actively using some of those notations. For a better presentation we restate them again here. We agree on the following conventions

1. $a_n \lesssim b_n$ or $a_n = O(b_n)$ whenever a_n is bounded by a constant multiple of b_n
2. $a_n \sim b_n$ means that $a_n \lesssim b_n$ as well as $b_n \lesssim a_n$
3. $a_n = o(b_n)$ meaning that $a_n/b_n \to 0$ and $a_n \asymp b_n$ when $a_n/b_n \to 1$.

2.4 Block-wise estimation in multidimensional space

We start this section with the continuous-time model, i.e. the PPP model.

2.4.1 Poisson point process model

The main goal of this section is to find an efficient way of estimating the following linear functional in dimension d

$$\vartheta = \int_{[0,1]^d} g(x)\omega(x)dx, \quad \text{for some } \omega \in L^2([0,1]^d). \tag{2.4.1}$$

Let $(X_j, Y_j)_{j\geq 1}$ be the observations of the PPP in $\mathbb{R}^d$ ($d \geq 1$) with the intensity function is of the form (2.3.1). To grasp the main idea behind block-wise estimation method that is being introduced in this section, we assume first that $\omega(x) \equiv 1$ holds, then the target of estimation is

$$\vartheta = \int_{[0,1]^d} g(x)dx. \tag{2.4.2}$$

Suppose that we also know a deterministic function $g_0 : [0,1]^d \to \mathbb{R}$ such that the following property $g_0 \geq g$ holds point-wise. The number of PPP observations that lie below the graph of g_0 is distributed according to the Poisson law with the intensity measure equal to n multiplied by the area between g and g_0, i.e.

$$\sum_{j\geq 1} \mathbf{1}(Y_j \leq g_0(X_j)) \sim Poiss\left(n \int_{[0,1]^d} (g_0 - g)(x)dx\right).$$

Hence, by introducing the following estimator $\hat{\vartheta}$ of the functional (2.4.2)

$$\hat{\vartheta} := \int_{[0,1]^d} g_0(x)dx - \frac{1}{n}\sum_{j\geq 1} \mathbf{1}(Y_j \leq g_0(X_j)), \tag{2.4.3}$$

we obtain that the expectation of that estimator (2.4.3) equals to (2.4.2)

$$\mathbb{E}[\hat{\vartheta}] = \int_{[0,1]^d} (g_0 - (g_0 - g))(x)dx = \vartheta,$$

which implies that the estimator $\hat{\vartheta}$ from (2.4.3) is unbiased for ϑ in (2.4.2). We can also compute the variance of the estimator $\hat{\vartheta}$

$$\mathrm{Var}(\hat{\vartheta}) = \frac{1}{n}\int_{[0,1]^d} ((g_0 - g))(x)dx. \tag{2.4.4}$$

From (2.4.4) we can see that the larger the area between the graphs the higher is the variance of $\hat{\vartheta}$.

To find a good candidate for an estimator of a general linear functional (2.4.1) we adapt the same idea as above, when $\omega = 1$. Now the main goal is to identify an empirical substitute for g_0 that keeps the property of unbiasedness by stopping time arguments. A key ingredient here is that the function g can be bounded locally under the Hölder-continuity assumption.

To this end, we divide a d-dimensional unit square $[0,1]^d$ into d-dimensional cubes defined as $I_k^d := [kh, (k+1)h)^d$ and the length of each side of this cube is being equal to h. Denote the block-wise minimum $(X_j, Y_j)_{j\geq 1}$ as Y_k^* then $Y_k^* := \min_{j:X_j \in I_k^d} Y_j$ and it satisfies the following inequality

$$Y_k^* \geq \min_{j:x\in I_k^d} g(x).$$

Since the Hölder property holds for g we can conclude by Definition 2.3.4 that

$$g(x) \leq Y_k^* + R\|z\|_{L^2}^\beta, \quad \forall x \in I_k^d, \tag{2.4.5}$$

where z is a vector in $\mathbb{R}^d$ with all entries equal to h. Thus, $Y_k^* + R\|z\|_{L^2}^\beta$ is a local upper bound for g. As an illustrative example consider a one-dimensional case, i.e., when $d = 1$, see Figure 2.3, where the construction of the estimator $\hat{\vartheta}$ is illustrated. Next, using the fact that the L^2-norm of the vector z is equal to $\sqrt{d}h$ we can rewrite the local upper bound from (2.4.5) as $Y_k^* + R\sqrt{d}h^\beta$. We now can estimate the functional (2.4.1) locally on the cubes introduced in the following way

$$\hat{\vartheta}_k := (Y_k^* + R\sqrt{d}h^\beta)\bar{\omega}_k - \frac{1}{nh^d}\sum_{i\geq 1} \mathbf{1}(X_i \in I_k^d, Y_i \leq Y_k^* + R\sqrt{d}h^\beta)\omega(X_i), \tag{2.4.6}$$

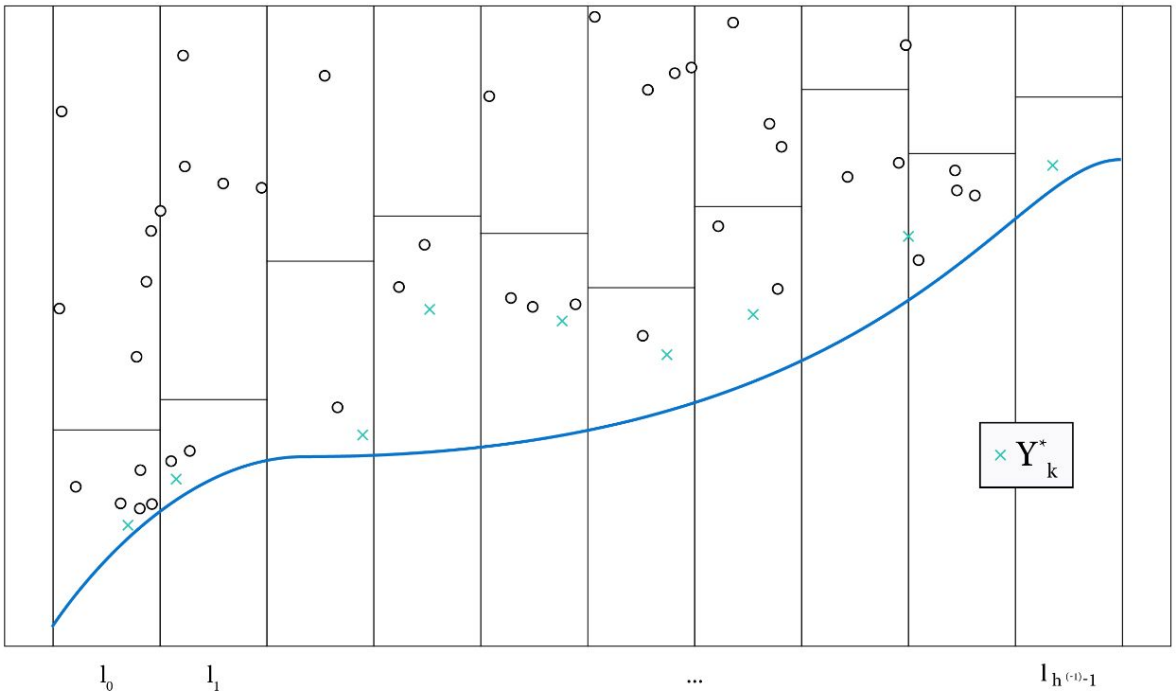


Figure 2.3: Construction of $\hat{\vartheta}^{block}$ in the one-dimensional case. The circles indicate the observation points X_i, Y_i in the PPP model. The crosses correspond to the block-wise minimum Y_k^* and the horizontal grey lines are the upper boundaries $Y_k^* + Rh^\beta$. The blue smooth line represents g.

where $\bar{\omega}_k = h^{-d} \int_{I_k^d} \omega(x)dx$. The set of estimators $(\hat{\vartheta}_k)$ estimate only the following local functionals $\vartheta_k := h^{-d} \int_{I_k^d} g(x)\omega(x)dx$. We further need to introduce a block-wise estimator to estimate the linear functional (2.4.1). We construct it as a weighted sum of the block-wise estimators $(\hat{\vartheta}_k)$, i.e. as

$$\hat{\vartheta}^{block} = \sum_{k=1}^{h^{-d}} \hat{\vartheta}_k h^d$$

Next, we formulate the main properties of this estimator (2.4.1) in the following theorem. Mainly, we show that the estimator $\hat{\vartheta}^{block}$ is unbiased and provide an upper bound for its variance.

Theorem 2.4.1. *Let $(X_i, Y_i)_{i\geq 1}$ be the observations of the PPP in $\mathbb{R}^d$ $(d \geq 1)$ of intensity*

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

with a function $g \in C^\beta(R)$. Then, for the estimator $\hat{\vartheta}^{block} = \sum_{k=1}^{h^{-d}} \hat{\vartheta}_k h^d$ of the functional (2.4.1) with $\hat{\vartheta}_k$ from (2.4.6) the following result holds

$$\mathbb{E}[\hat{\vartheta}^{block}] = \vartheta, \quad \mathrm{Var}(\hat{\vartheta}^{block}) \leq \frac{2R\sqrt{d}h^\beta + (nh^d)^{-1}}{n} \|\omega\|_{L^2}^2,$$

where $\|\omega\|_{L^2}^2 = \int_{[0,1]^d} \omega(x)^2 dx$. *Moreover, for the asymptotically optimal size of the length of the cube* I_k^d *which is given by*

$$h \asymp (n/\sqrt{d})^{-1/(\beta+d)}$$

it holds

$$\limsup_{n\to\infty} \sup_{g\in\mathfrak{C}^\beta(R)} n^{(2\beta+d)/(\beta+d)} \mathrm{Var}(\hat{\vartheta}^{block}) \leq \frac{\beta+d}{\beta} (2\beta R/\sqrt{d})^{d/(\beta+d)} \|\omega\|_{L^2}^2.$$

Proof. Consider the weighted counting process defined on the cube I_k^d in the following way

$$N(t) := \sum_{i\geq 1} \mathbf{1}(X_i \in I_k^d, Y_i \leq t)\omega(X_i), \quad t \in \mathbb{R}.$$

When $w \equiv 1$, $(N(t), t \in \mathbb{R})$ corresponds to the pure counting process $\sum_{i\geq 1} \mathbf{1}(X_i \in I_k^d, Y_i \leq t)$, which is a point process with respect to t and such that the intensity measure $\lambda_t = n\int_{[0,1]^d}(t - g(x))_+ dx$ is deterministic. As a result, one can show that $(N(t), t \in \mathbb{R})$ for a general ω is a process with independent increments satisfying

$$\mathbb{E}[N(t)] = \int_{I_k^d} n(t - g(x))_+ \omega(x) dx, \quad \mathrm{Var}(N(t)) = \int_{I_k^d} n(t - g(x))_+ \omega(x)^2 dx.$$

Moreover, the process $M(t) = N(t) - \mathbb{E}[N(t)]$, which is a centred weighted counting process, is a cádlág martingale with respect to the following filtration

$$\mathcal{F}_t = \sigma((X_i, Y_i)\mathbf{1}(Y_i \leq t), i \geq 1) \quad t \in \mathbb{R}, \tag{2.4.7}$$

which predictable quadratic variation is given by $\langle M \rangle_t = \mathrm{Var}(N(t))$, (see Proposition 2.32 in [66]).

Now define a stopping time τ as $\tau := Y_k^* + R\sqrt{d}h^\beta$, which is an $\mathcal{F}_t$-stopping time. Next, to demonstrate the unbiasedness and estimate the variance of $\hat{\vartheta}_k$, we have to elaborate on the properties of the stopping time τ a bit more. We first show that the expectation of τ is finite. For that we calculate an upper bound for the following probability

$$\begin{aligned} P(\tau \geq t) &= \exp\Big(- n \int_{I_k^d} (t - R\sqrt{d}h^\beta - g(x))_+ dx \Big) \\ &\leq \exp\Big(- nh^d (t - R\sqrt{d}h^\beta - \max_{x\in I_k^d} g(x)) \Big) \end{aligned}$$

for $t \geq \max_{x\in I_k^d} g(x) + R\sqrt{d}h^\beta$. In particular, if we take $T = \max_{x\in I_k^d} g(x) + R\sqrt{d}h^\beta$, for

the expectation of τ we obtain the result

$$\begin{aligned}\mathbb{E}[\tau] &= \int_0^\infty P(\tau \geq t)dt = \int_0^T P(\tau \geq t)dt + \int_T^\infty P(\tau \geq t)dt = \\ &\leq \int_0^T P(\tau \geq t)dt + \exp\left(nh^d(R\sqrt{d}h^\beta + \max_{x\in I_k^d} g(x))\right)\int_0^\infty e^{-nh^d t}dt < \infty. \end{aligned} \tag{2.4.8}$$

Hence, from the fact that the expectation of τ is finite and Theorem 2.7.1 on optional stopping it follows that $\mathbb{E}[M(\tau)] = 0$. From (2.4.8) we have that the expectation of $Y_k^* + R\sqrt{d}h^\beta$ is finite. This implies that we always observe a point in any cube I_k^d of a given length h, suggesting that the estimator is well-defined for any choice of h. However, while reducing h at some point the estimator will result in a higher variance, as on some intervals the upper boundary will move far away for the function of interest g, as Y_k^* might jump relatively high with a relatively small choice of h. Now, for the local estimator $\hat{\vartheta}_k$ by plugging τ into (2.4.6) and using the fact that $\vartheta_k := h^{-d}\int_{I_k^d} g(x)\omega(x)dx$ and $\bar{\omega}_k = h^{-d}\int_{I_k^d}\omega(x)dx$

$$\begin{aligned}\hat{\vartheta}_k &= \tau\bar{\omega}_k - \frac{1}{nh^d}N(\tau) = \tau h^{-d}\int_{I_k^d}\omega(x)dx - \frac{1}{nh^d}\mathbb{E}N(\tau) - \frac{1}{nh^d}M(\tau) \\ &= \tau h^{-d}\int_{I_k^d}\omega(x)dx - h^{-d}\int_{I_k^d}\big(\tau - g(x)\big)\omega(x)dx - \frac{1}{nh^d}M(\tau) \\ &= \vartheta_k - \frac{1}{nh^d}M(\tau), \end{aligned} \tag{2.4.9}$$

that $\mathbb{E}[\hat{\vartheta}_k] = \vartheta_k$. Moreover, we have

$$\mathbb{E}[N(\tau)] = n\int_{I_k^d}\mathbb{E}[(\tau - g(x))_+]\omega(x)dx$$

and since $\langle M\rangle_t = \mathrm{Var}(N(t))$ we obtain

$$\mathrm{Var}[M(\tau)] = \mathbb{E}[\langle M\rangle_\tau] = n\int_{I_k^d}\mathbb{E}[(\tau - g(x))_+]\omega(x)^2dx.$$

The identity

$$\hat{\vartheta}_k = \tau\bar{\omega}_k - \frac{1}{nh^d}N(\tau) = \vartheta_k - \frac{1}{nh^d}M(\tau)$$

and the fact that the estimator is unbiased $\mathbb{E}[\hat{\vartheta}_k] = \vartheta_k$ together with $\tau \geq g(x)$ imply

$$\mathbb{E}[(\tau - g(x))_+] = \mathbb{E}[Y_k^*] + R\sqrt{d}h^\beta - g(x) \quad \text{for all } x \in I_k^d.$$

As a result this brings us to the following equation for the variance of $M(\tau)$

$$\mathrm{Var}[M(\tau)] = n \int_{I_k^d} \mathbb{E}[Y_k^* + R\sqrt{d}h^\beta - g(x)]\omega(x)^2 dx. \tag{2.4.10}$$

From (2.4.9) using the identity for $\hat{\vartheta}_k$ of the form

$$\hat{\vartheta}_k = \vartheta_k - \frac{1}{n} M(\tau)$$

together with (2.4.10) we arrive to the result for the variance of $\hat{\vartheta}_k$

$$\mathrm{Var}(\hat{\vartheta}_k) = \frac{1}{n^2 h^{2d}} \mathrm{Var}[M(\tau)] = \frac{1}{nh^{2d}} \int_{I_k^d} \mathbb{E}[Y_k^* + R\sqrt{d}h^\beta - g(x)]\omega(x)^2 dx.$$

Next, to evaluate the upper bound for the variance of the estimator $\hat{\vartheta}_k$ we need to provide an upper bound for $\mathbb{E}[Y_k^*]$. By using the fact that $Y_k^* - \max_{x \in I_k^d} g(x)$ is stochastically smaller that the minimum of the Poisson point process of the intensity $n\mathbf{1}(x \in I_k^d, y \geq 0)$ in y we conclude the rough universal bound with a random variable $E \sim \mathrm{Exp}(nh^d)$

$$\mathbb{E}[Y_k^*] \leq \mathbb{E}[\max_{x \in I_k^d} g(x) + E] \leq g(x) + R\sqrt{d}h^\beta + (nh^d)^{-1}.$$

This proves that

$$\mathrm{Var}(\hat{\vartheta}_k) \leq \frac{2R\sqrt{d}h^\beta + (nh^d)^{-1}}{nh^{2d}} \int_{I_k^d} \omega(x)^2 dx.$$

Now we are able to conclude that for the final estimator $\hat{\vartheta}^{block} = \sum_{k=1}^{h^{-d}} \hat{\vartheta}_k h^d$ from the property of independence of $(\hat{\vartheta}_k)$ it holds $\mathbb{E}[\hat{\vartheta}^{block}] = \vartheta$, and

$$\mathrm{Var}(\hat{\vartheta}^{block}) \leq \frac{2R\sqrt{d}h^\beta + (nh^d)^{-1}}{n} \sum_{k=1}^{h^{-d}} \int_{I_k^d} \omega(x)^2 dx = \frac{2R\sqrt{d}h^\beta + (nh^d)^{-1}}{n} \|\omega\|_{L^2}^2.$$

To find the asymptotically optimal h we take the derivative of $(2R\sqrt{d}h^\beta + (nh^d)^{-1})/n$ over h and equate it to zero

$$\Big(\frac{2R\sqrt{d}h^\beta + (nh^d)^{-1}}{n}\Big)'_h = \frac{2R\sqrt{d}n\beta h^{\beta+d} - d}{n^2 h^{d+1}} = 0,$$

the solution of this equation is $h \asymp (2\beta Rn/\sqrt{d})^{-1/(\beta+d)}$. Insertion of the asymptotically optimal h yields the variance bound

$$\mathrm{Var}(\hat{\vartheta}^{block}) \leq \frac{\beta + d}{\beta} (2\beta R/\sqrt{d})^{d/(\beta+d)} n^{-(2\beta+d)/(\beta+d)} \|\omega\|_{L^2}^2.$$

□

Example 2.4.2. We now consider an example of how the above results can be applied when a function possess Hölder property along each coordinate axis. For a function $g : [0,1]^2 \to \mathbb{R}$ which is from a Hölder ball of regularity $\beta_1 \in (0,1]$ and radius R_1 for the first component and from a Hölder ball of regularity $\beta_2 \in (0,1]$ and radius R_2 for the second component, the following inequalities hold

$$|g(x_1^1, x^2) - g(x_2^1, x^2)| \le R_1 |x_1^1 - x_2^1|^{\beta_1},$$

$$|g(x^1, x_1^2) - g(x^1, x_2^2)| \le R_1 |x_1^2 - x_2^2|^{\beta_2}.$$

For any two vectors x^1 and x^2 from $\mathbb{R}^2$ it holds

$$\begin{aligned} |g(x_1^1, x_2^1) - g(x_1^2, x_2^2)| &= |g(x_1^1, x_2^1) - g(x_1^2, x_2^1) + g(x_1^2, x_2^1) - g(x_1^2, x_2^2)| \\ &\le R_1 |x_1^1 - x_1^2|^{\beta_1} + R_1 |x_2^1 - x_2^2|^{\beta_2}. \end{aligned}$$

For a function g we use the above theory to estimate linear functions. The analogous block-wise estimator is again $\hat{\vartheta}^{block} = \sum_{k=0}^{(h_1 h_2)^{-1}-1} \hat{\vartheta}_k$ with

$$\hat{\vartheta}_k := (Y_k^* + R_1 h_1^\beta + R_2 h_2^\beta)\bar{\omega}_k - \frac{1}{n h_1^2 h_2^2} \sum_{i \ge 1} \mathbf{1}(X_i \in I_k^2, Y_i \le Y_k^* + R_1 h_1^\beta + R_2 h_2^\beta),$$

and the upper bound of the variance is given by

$$\mathrm{Var}(\hat{\vartheta}^{block}) \le \frac{2R_1 h_1^\beta + 2R_2 h_2^\beta + (n h_1 h_2)^{-1}}{n} \|\omega\|_{L^2}^2.$$

The asymptotically optimal size of the length of the cube's side is a solution of the following equation

$$\left(2R_1 h_1^{\beta_1} + 2R_2 h_2^{\beta_2} + (n h_1 h_2)^{-1}\right)' = 0,$$

under the constraint $h_1^{\beta_1} = h_2^{\beta_2}$ we obtain

$$h_1 \asymp n^{-\beta_2/(\beta_1\beta_2 + \beta_1 + \beta_2)},$$

$$h_2 \asymp n^{-\beta_1/(\beta_1\beta_2 + \beta_1 + \beta_2)}.$$

Introducing $1/\beta = (\beta_1 + \beta_2)/\beta_1\beta_2$ we have $h_1^{\beta_1} = h_2^{\beta_2} \asymp n^{-1/(1+1/\beta)}$.

2.4.2 Regression-type model

Let us recall the regression model with one-sided errors

$$\mathscr{Y}_i = g(\mathbf{X}_i) + \varepsilon_i, \;\; 1 \leq i \leq n,$$

where $(\varepsilon_i)_{1\leq i\leq n}$ are independent and identically distributed with $\varepsilon_i \geq 0$ and for some $\lambda > 0$ it holds

$$P(\varepsilon_i \leq x) = \lambda x + O(x^2) \tag{2.4.11}$$

as $x \downarrow 0$. The main example is when the error terms are exponentially distributed, i.e. $\varepsilon_i \sim Exp(\lambda)$, for $1 \leq i \leq n$, however, any distribution on $\mathbb{R}^d_+$ such that the distribution possesses a Lipschitz continuous density f_ε at zero and $f_\varepsilon(0) = \lambda$ is covered as long as some loose tail boundary condition at infinity holds.

Recall the functional of interest in the discrete-time representation as before satisfies

$$\vartheta^{(n)} = n^{-1} \sum_{i=1}^{n} g(\mathbf{X}_i)\omega(\mathbf{X}_i) \tag{2.4.12}$$

for some function $\omega : [0,1]^d \to \mathbb{R}$. Similarly to the PPP case, we partition $[0,1]^d$ into d-dimensional cubes $\tilde{I}^d_j$ with the length h of each side, such that $1/h^d \in \mathbb{N}$. Next, on each cubes of indices $\tilde{I}^d_k$ we construct an estimator for $\vartheta_k = n^{-1}\sum_{i\in\tilde{I}^d_k} g(\mathbf{X}_i)\omega(\mathbf{X}_i)$ as

$$\tilde{\vartheta}_k := \frac{1}{n} \sum_{i\in\tilde{I}^d_k} \Big(\mathscr{Y}_i \wedge (\mathscr{Y}^*_k + R\sqrt{d}h^\beta) - \lambda^{-1}\mathbf{1}(\mathscr{Y}_i \leq \mathscr{Y}^*_k + R\sqrt{d}h^\beta) \Big) \omega(\mathbf{X}_i). \tag{2.4.13}$$

As before $\mathscr{Y}^*_k = \min_{i\in\tilde{I}^d_k} \mathscr{Y}_i$ corresponds to the minimal observation on each cube. The idea to use a similar estimator as in the regression-type model comes from the fact that the models are asymptotically equivalent in the Le Cam sense, see [79, 80], and so intuitively a similar-type estimator should perform well in both models. The main difference with the PPP-estimator is in the empirical upper bound for g on $\tilde{I}^d_k$. Now this upper bound is given by the minimum of $\mathscr{Y}^*_k + R\sqrt{d}h^\beta$ and $\mathscr{Y}_i$. This difference has a negligible impact on the rate-optimal choice of h. Now let $\|\omega\|_p = \big(\frac{1}{n}\sum_{i=1}^n |\omega(\mathbf{X}_i)|^p\big)^{1/p}$ denote the standardised l^p-norm, then, the block-wise estimator for the functional (2.4.12) is given by

$$\tilde{\vartheta}^{block}_n = \sum_{k=0}^{h^{-d}-1} \tilde{\vartheta}_k.$$

The next result provides upper bounds for both the bias and variance of the estimator $\tilde{\vartheta}^{block}_n$ in case when the error terms satisfy (2.4.11), as well as demonstrates the property of unbiasedness in case of exponentially distributed error terms.

Theorem 2.4.3. *Assume that error variables ε satisfy* (2.4.11) *and also are independent and identically distributed. Then, for $g \in C^{\beta}(R)$ the estimator $\tilde{\vartheta}_n^{block} = \sum_{k=0}^{h^{-d}-1} \tilde{\vartheta}_k$, with each $\tilde{\vartheta}_k$ following* (2.4.13)*, satisfies uniformly in n, h, R, β*

$$|\mathbb{E}[\tilde{\vartheta}_n^{block} - \vartheta^{(n)}]| \lesssim (R\sqrt{d}h^{\beta} + (nh^d)^{-1})^2 \|\omega\|_1, \;\; \mathrm{Var}(\tilde{\vartheta}_n^{block}) \lesssim \frac{R\sqrt{d}h^{\beta} + (nh^d)^{-1}}{n} \|\omega\|_2,$$

when $\sqrt{d}h \to 0$ together with $nh^d \to \infty$. For $d \geq 1$ with the rate-optimal length of the side of the cube

$$h \sim n^{-1/(\beta+d)}$$

it holds

$$\Big(\mathbb{E}[\tilde{\vartheta}_n^{block} - \vartheta^{(n)}]\Big)^2 = o\big(\mathrm{Var}(\tilde{\vartheta}_n^{block})\big), \;\; \mathrm{Var}(\tilde{\vartheta}_n^{block}) \lesssim n^{-(2\beta+d)/(\beta+d)}.$$

Remark 2.4.4. In Theorem 2.4.3 if we consider the case of dimension 1 uniformly over $\beta > 1/2$, $R \leq R_0 < \infty$ we obtain for the rate-optimal length of the cube's side

$$h \sim n^{-1/(\beta+1)}$$

the following result

$$\Big(\mathbb{E}[\tilde{\vartheta}_n^{block} - \vartheta^{(n)}]\Big)^2 = o\big(\mathrm{Var}(\tilde{\vartheta}_n^{block})\big), \;\; \mathrm{Var}(\tilde{\vartheta}_n^{block}) \lesssim R^{1/\beta+1} n^{-(2\beta+1)/(\beta+1)} \|\omega\|_2^2.$$

Remark 2.4.5. For the estimator $\tilde{\vartheta}_n^{block}$ the following result holds

$$\sqrt{\mathbb{E}\big[(\tilde{\vartheta}_n^{block} - \vartheta^{(n)})^2\big]} = O\big(n^{-(\beta+d/2)/(\beta+d)}\big).$$

Moreover, it can be inferred that for the values of d up to $O(\log n)$ the variance of the estimator $\tilde{\vartheta}_n^{block}$ tends to 0 as $n \to \infty$.

Proof. As in the proof of the previous result for the PPP model we first fix the cube $\tilde{I}_k^d$ of index k and for $t \in \mathbb{R}$ introduce the following process

$$M(t) := \sum_{i \in \tilde{I}_k^d} \Big(\mathbf{1}(\mathscr{Y}_i \leq t) + \log \bar{F}_{\varepsilon}(\mathscr{Y}_i \wedge t - g(\mathbf{X}_i))\Big)\omega(\mathbf{X}_i), \tag{2.4.14}$$

where $\bar{F}_{\varepsilon}$ is defined as $\bar{F}_{\varepsilon}(y) = P(\varepsilon_i \geq y)$. This process is formed of two components. The first one $\sum_{i \in \tilde{I}_k^d} \mathbf{1}(\mathscr{Y}_i \leq t)$ represents the counting process and the second component corresponds to it compensator. We can rewrite the equation (2.4.14) by utilising the notion of a cumulative hazard function $H(\bullet)$ for $\mathscr{Y}_i$, which is determined in the following way

$$H(t) \equiv H_{\mathscr{Y}_i}(t) \equiv \int_0^{t-g(\mathbf{X}_i)} [1/\bar{F}_{\varepsilon}(y)]d]\bar{F}_{\varepsilon}(y) = -\log \bar{F}_{\varepsilon}(t - g(\mathbf{X}_i)).$$

As the next step, we plug in the hazard function H into (2.4.14)

$$\begin{aligned} M(t) &= \sum_{i\in\tilde{I}_k^d}\Big(\mathbf{1}(\mathscr{Y}_i \le t) - \int_0^{t-g(\mathbf{X}_i)} \mathbf{1}(\varepsilon_i \ge u) dH(u)\Big)\omega(\mathbf{X}_i) \\ &= \sum_{i\in\tilde{I}_k^d}\Big(\mathbf{1}(\mathscr{Y}_i \le t) - \int_0^{\mathscr{Y}_i\wedge t-g(\mathbf{X}_i)} dH(u)\Big)\omega(\mathbf{X}_i) \\ &= \sum_{i\in\tilde{I}_k^d}\Big(\mathbf{1}(\mathscr{Y}_i \le t) - H(\mathscr{Y}_i \wedge t - g(\mathbf{X}_i))\Big)\omega(\mathbf{X}_i). \end{aligned}$$

The process $(M(t), t \in \mathbb{R})$ represents a martingale with respect to the following filtration

$$\mathcal{F}_t = \sigma\big(\mathscr{Y}_i \mathbf{1}(\mathscr{Y}_i \le t), i \in \tilde{I}_k^d\big),$$

and since $M(0) = 0$, we can conclude that $\mathbf{E}[M(t)] = 0$. The compensator of the following counting process $\sum_{i\in\tilde{I}_k^d} \mathbf{1}(\mathscr{Y}_i \le t)$ equals to the integrated hazard function $\sum_{i\in\tilde{I}_k^d} H\big(\mathscr{Y}_i \wedge t - g(\mathbf{X}_i)\big)$. From the result of Theorem 2.5.1 in [33] we can then infer for the process

$$\tilde{M}(t) = \sum_{i\in\tilde{I}_k^d}\Big(\mathbf{1}(\mathscr{Y}_i \le t) - H\big(\mathscr{Y}_i \wedge t - g(\mathbf{X}_i)\big)\Big)$$

that its quadratic variation equals to

$$\langle\tilde{M}\rangle_t = \sum_{i\in\tilde{I}_k^d} H\big(\mathscr{Y}_i \wedge t - g(\mathbf{X}_i)\big) = \sum_{i\in\tilde{I}_k^d}\Big(-\log\bar{F}_\varepsilon\big(\mathscr{Y}_i \wedge t - g(\mathbf{X}_i)\big)\Big),$$

and thus, the quadratic variation for the process $\langle M\rangle_t$ can be calculated as

$$\langle M\rangle_t = \sum_{i\in\tilde{I}_k^d}\Big(-\log\bar{F}_\varepsilon\big(\mathscr{Y}_i \wedge t - g(\mathbf{X}_i)\big)\Big)\omega(\mathbf{X}_i)^2.$$

Furthermore, by introducing the following random variable $\tau := \mathscr{Y}_k^* + R\sqrt{d}h^\beta$ one can see that it is a stopping time with respect to $(\mathcal{F}_t)$. Taking the following difference

$$\tilde{\vartheta}_k - \vartheta_k = \frac{1}{n\lambda h^d}\Big(\sum_{i\in\tilde{I}_k^d}\big[\lambda\big(\mathscr{Y}_i \wedge \tau - g(\mathbf{X}_i)\big) + \log\bar{F}_\varepsilon\big(\mathscr{Y}_i \wedge \tau - g(\mathbf{X}_i)\big)\big]\,\omega(\mathbf{X}_i) - M(\tau)\Big)$$

from the representation of this difference and the stopping Theorem 2.7.1 together with the fact that the stopping time τ has a finite expectation, i.e. $\mathbb{E}[\tau] < \infty$, which is the

result of the moment bound from Lemma 2.7.2, we obtain the next result

$$\mathbb{E}[\tilde{\vartheta}_k - \vartheta_k] = \frac{1}{n\lambda h^d} \sum_{i \in \tilde{I}_k^d} \mathbb{E}\left[\lambda\big(\mathscr{Y}_i \wedge \tau - g(\mathbf{X}_i)\big) + \log \bar{F}_\varepsilon\big(\mathscr{Y}_i \wedge \tau - g(\mathbf{X}_i)\big)\right] \omega(\mathbf{X}_i),$$

here for the simplicity we introduce the following function $G_\varepsilon(z) := \lambda z + \log \bar{F}_\varepsilon(z)$, then, the equation above can be rewritten as

$$\mathbb{E}[\tilde{\vartheta}_k - \vartheta_k] = \frac{1}{n\lambda h^d} \sum_{i \in \tilde{I}_k^d} \mathbb{E}\left[G_\varepsilon\big(\mathscr{Y}_i \wedge \tau - g(\mathbf{X}_i)\big)\right] \omega(\mathbf{X}_i). \tag{2.4.15}$$

When the error terms are exponentially distributed, i.e. $\varepsilon \sim Exp(\lambda)$, it holds that $\log \bar{F}_\varepsilon(z) = -\lambda z$ and as a result it follows that $G_\varepsilon(z) = 0$. As a result, we can conclude that the estimator is unbiased.
In a more general case, there is some $\delta > 0$ such that $|G_\varepsilon(z)| \lesssim z + |\log \bar{F}_\varepsilon(z)| \mathbf{1}(z > \delta)$. Thus, by plugging in the result into (2.4.15) and introducing the following notation $\mathcal{C}(\tau, \mathbf{X}_i) := \tau - g(\mathbf{X}_i)$ and $\mathcal{K}(\mathscr{Y}_i, \mathbf{X}_i) := \mathscr{Y}_i - g(\mathbf{X}_i)$ we have

$$\begin{aligned}
|\mathbb{E}[\tilde{\vartheta}_k - \vartheta_k]| &\leq \frac{1}{n\lambda h^d} \sum_{i \in \tilde{I}_k^d} \mathbb{E}[|G_\varepsilon(\mathscr{Y}_i \wedge \tau - g(\mathbf{X}_i))|] |\omega(\mathbf{X}_i)| \\
&\lesssim \frac{1}{n\lambda h^d} \sum_{i \in \tilde{I}_k^d} \mathbb{E}[(\mathscr{Y}_i \wedge \tau - g(\mathbf{X}_i))^2 + |\log \bar{F}_\varepsilon(\mathscr{Y}_i \wedge \tau - g(\mathbf{X}_i))| \mathbf{1}(|\log \bar{F}_\varepsilon(\mathscr{Y}_i \wedge \tau - g(\mathbf{X}_i))| > \delta)] |\omega(\mathbf{X}_i)| \\
&\lesssim \frac{1}{n\lambda h^d} \sum_{i \in \tilde{I}_k^d} \mathbb{E}[(\mathcal{C}(\tau, \mathbf{X}_i))^2 + |\log \bar{F}_\varepsilon(\mathcal{K}(\mathscr{Y}_i, \mathbf{X}_i))| \mathbf{1}(|\log \bar{F}_\varepsilon(\mathcal{K}(\mathscr{Y}_i, \mathbf{X}_i))| > \delta)] |\omega(\mathbf{X}_i)|.
\end{aligned}$$

Next, from the Cauchy-Schwarz inequality and the fact that $\bar{F}_\varepsilon(\varepsilon_i) \sim U[0,1]$ we obtain

$$\begin{aligned}
|\mathbb{E}[\tilde{\vartheta}_k - \vartheta_k]| &\lesssim \frac{1}{nh^d} \sum_{i \in \tilde{I}_k^d} \mathbb{E}[(\mathcal{C}(\tau, \mathbf{X}_i))^2] + \mathbb{E}[\log(\bar{F}_\varepsilon(\mathcal{K}(\mathscr{Y}_i, \mathbf{X}_i))^2]^{1/2} P(|\log \bar{F}_\varepsilon(\mathcal{K}(\mathscr{Y}_i, \mathbf{X}_i))| > \delta)^{1/2} |\omega(\mathbf{X}_i)| \\
&\lesssim \frac{1}{nh^d} \sum_{i \in \tilde{I}_k^d} \mathbb{E}[(\mathcal{C}(\tau, \mathbf{X}_i))^2] + \mathbb{E}[\log(\bar{F}_\varepsilon(\varepsilon_i))^2]^{1/2} P(\bar{F}_\varepsilon(\min_{i \in \tilde{I}_k^d} \varepsilon_i + 2R\sqrt{d}h^\beta) < e^{-\delta})^{1/2} |\omega(\mathbf{X}_i)| \\
&\lesssim \frac{1}{nh^d} \sum_{i \in \tilde{I}_k^d} \mathbb{E}[(\mathcal{C}(\tau, \mathbf{X}_i))^2] + P(\min_{i \in \tilde{I}_k^d} \varepsilon_i < \bar{F}_\varepsilon^{-1}(e^{-\delta}) - 2R\sqrt{d}h^\beta)^{1/2} |\omega(\mathbf{X}_i)| \\
&\lesssim \frac{(R\sqrt{d}h^\beta + (n\lambda h^d)^{-1})^2 + \bar{F}_\varepsilon(\bar{F}_\varepsilon^{-1}(e^{-\delta}) - 2R\sqrt{d}h^\beta)^{nh^d/2}}{nh^d} \sum_{i \in \tilde{I}_k^d} |\omega(\mathbf{X}_i)|.
\end{aligned}$$

By setting $h \to 0$ and $nh^d \to \infty$ the second term in the numerator converges geometrically

quick to zero. As a result, we immediately have the assertion for the bias of $\tilde{\vartheta}_n^{block}$

$$|\mathbb{E}[\tilde{\vartheta}_n^{block} - \vartheta^{(n)}]| \lesssim \frac{(R\sqrt{d}h^\beta + (n\lambda h^d)^{-1})^2}{\lambda}\|\omega\|_1.$$

To bound the variance we use the standard upper bound for the variance of a sum of two random variables, i.e. $\mathrm{Var}(A+B) \leq 2\mathrm{Var}(A) + 2\mathrm{Var}(B)$.

$$\begin{aligned}
\mathrm{Var}(\tilde{\vartheta}_k) &= \mathrm{Var}(\tilde{\vartheta}_k - \vartheta_k) \\
&= \frac{1}{(n\lambda h^d)^2}\mathrm{Var}\Big(\sum_{i\in\tilde{I}_k^d}\big[\lambda(\mathscr{Y}_i \wedge \tau - g(\mathbf{X}_i)) + \log\bar{F}_\varepsilon(\mathscr{Y}_i \wedge t - g(\mathbf{X}_i))\big]\,\omega(\mathbf{X}_i) - M(\tau)\Big) \\
&\leq \frac{2}{(n\lambda h^d)^2}\mathrm{Var}\Big(\sum_{i\in\tilde{I}_k^d}[G_\varepsilon(\mathscr{Y}_i \wedge \tau - g(\mathbf{X}_i))]\,\omega(\mathbf{X}_i)\Big) + \frac{2}{(n\lambda h^d)^2}\mathrm{Var}(M(\tau)) \\
&\leq \frac{2}{(n\lambda h^d)^2}\sum_{i\in\tilde{I}_k^d} nh^d\mathbb{E}\Big[\Big(G_\varepsilon(\mathscr{Y}_i \wedge \tau - g(\mathbf{X}_i))\omega(\mathbf{X}_i)\Big)^2\Big] + \frac{2}{(n\lambda h^d)^2}\mathbb{E}[M(\tau)^2].
\end{aligned}$$

From Theorem 2.7.1 we know that $\mathbb{E}[M(\tau)^2] = \mathbb{E}[\langle M\rangle_\tau]$ holds. Hence, we have

$$\mathbb{E}[M(\tau)^2] = \mathbb{E}\Big[\sum_{i\in\tilde{I}_k^d}\Big(-\log\bar{F}_\varepsilon\big(\mathscr{Y}_i \wedge \tau - g(\mathbf{X}_i)\big)\Big)\omega(\mathbf{X}_i)^2\Big].$$

To bound the first term in the upper bound for the variance $\mathrm{Var}(\tilde{\vartheta}_k)$, we use the same technique as above

$$\begin{aligned}
\mathbb{E}\Big[\big(G_\varepsilon(\mathscr{Y}_i \wedge \tau - g(\mathbf{X}_i))\big)^2\Big] &\lesssim \mathbb{E}\Big[\big(\tau - g(\mathbf{X}_i)\big)^4\Big] \\
&\quad + \mathbb{E}\Big[\log\bar{F}_\varepsilon(\varepsilon_i)^4\Big]^{1/2} P\big(\bar{F}_\varepsilon(\min_{i\in\tilde{I}_k^d}\varepsilon_i + 2R\sqrt{d}h^\beta) < e^{-\delta}\big)^{1/2} \\
&\lesssim \big(R\sqrt{d}h^\beta + (n\lambda h^d)^{-1}\big)^4 + \bar{F}_\varepsilon\big(\bar{F}_\varepsilon^{-1}(e^{-\delta}) - 2R\sqrt{d}h^\beta\big)^{nh^d/2} \\
&\lesssim \big(R\sqrt{d}h^\beta + (n\lambda h^d)^{-1}\big)^4.
\end{aligned}$$

Now it is left to find an upper bound for the expectation of quadratic variation $\mathbb{E}[\langle M\rangle_\tau]$ at time τ. To bound it, we also use the same technique as above

$$\begin{aligned}
\mathbb{E}[M(\tau)^2] &\leq \mathbb{E}\Big[\sum_{i\in\tilde{I}_k^d}|\log\bar{F}_\varepsilon\big(\mathscr{Y}_i \wedge \tau - g(\mathbf{X}_i)\big)|\omega(\mathbf{X}_i)^2\Big] \\
&\leq \mathbb{E}\Big[\sum_{i\in\tilde{I}_k^d}\big(|G_\varepsilon\big(\mathscr{Y}_i \wedge \tau - g(\mathbf{X}_i)\big)| + \lambda(\mathscr{Y}_i \wedge \tau - g(\mathbf{X}_i))\big)\omega(\mathbf{X}_i)^2\Big] \\
&\lesssim \mathbb{E}\Big[\sum_{i\in\tilde{I}_k^d}\big((\tau - g(\mathbf{X}_i))^2 + (\tau - g(\mathbf{X}_i)) + |\log\bar{F}_\varepsilon(\mathscr{Y}_i - g(\mathbf{X}_i))|\mathbf{1}(|\log\bar{F}_\varepsilon(\mathscr{Y}_i - g(\mathbf{X}_i))| > \delta)\Big]
\end{aligned}$$

$$\lesssim R\sqrt{d}h^{\beta} + (n\lambda h^d)^{-1} + \bar{F}_{\varepsilon}\big(\bar{F}_{\varepsilon}^{-1}(e^{-\delta}\big) - 2R\sqrt{d}h^{\beta})^{nh^d/2}$$
$$\lesssim R\sqrt{d}h^{\beta} + (n\lambda h^d)^{-1}.$$

Finally,

$$\mathrm{Var}(\tilde{\vartheta}_k) \lesssim \frac{(R\sqrt{d}h^{\beta} + (nh^d)^{-1})^4}{nh^d}\sum_{i\in\tilde{I}_k^d}\omega(\mathbf{X}_i)^2 + \frac{R\sqrt{d}h^{\beta} + (nh^d)^{-1}}{(nh^d)^2}\sum_{i\in\tilde{I}_k^d}\omega(\mathbf{X}_i)^2$$
$$\lesssim \frac{R\sqrt{d}h^{\beta} + (nh^d)^{-1}}{(nh^d)^2}\sum_{i\in\tilde{I}_k^d}\omega(\mathbf{X}_i)^2.$$

For the global estimator $\tilde{\vartheta}_n^{block}$ we infer the result from the independence of $\tilde{\vartheta}_k$

$$\mathrm{Var}(\tilde{\vartheta}_n^{block}) \lesssim h^{2d}\frac{R\sqrt{d}h^{\beta} + (nh^d)^{-1}}{nh^{2d}}\|\omega\|_2^2 \lesssim \frac{R\sqrt{d}h^{\beta} + (nh^d)^{-1}}{n}\|\omega\|_2^2.$$

It can be shown as in the case of PPP model that the optimal length of the side of the cube is $h \sim (Rn)^{-1/(\beta+d)}$. Inserting the rate-optimal choice of h into the bias bound and variance we obtain

$$(\mathbb{E}[\tilde{\vartheta}_n^{block} - \vartheta^{(n)}])^2 \lesssim R^{4d/(\beta+d)}n^{-4\beta/(\beta+d)}\|\omega\|_1^2,$$

$$\mathrm{Var}(\tilde{\vartheta}_n^{block}) \lesssim R^{d/(\beta+d)}n^{-\beta/(\beta+d)}n^{-1} = R^{d/\beta+d}n^{-(2\beta+d)/(\beta+d)}\|\omega\|_2^2.$$

It remains to note that $n^{-4\beta/\beta+1} = o(n^{-(2\beta+1)/(\beta+1)})$ holds for $\beta > 1/2$ in one-dimensional case and that $n^{-(2\beta+d)/(\beta+d)} = o(n^{-4\beta/\beta+d})$ holds for $d > 1$.

□

Corollary 2.4.1. Under the assumption of Theorem 2.4.3 as well as the additional requirement that $\varepsilon_i \sim Exp(\lambda)$, $i = 1, \ldots, n$, for any $\beta \in (0, 1]$, R, $\lambda > 0$ the following more precise result holds

$$\mathbb{E}[\tilde{\vartheta}_n^{block}] = \vartheta, \;\; \mathrm{Var}(\tilde{\vartheta}_n^{block}) \leq \frac{2R\sqrt{d}h^{\beta} + (n\lambda h^d)^{-1}}{n\lambda}\|\omega\|_2^2.$$

Proof. Consider the case of $\varepsilon_i \sim Exp(\lambda)$, $i = 1, \ldots, n$. Then, the following two inequalities hold

$$\mathbb{E}[Y_k^* - \max_i g(\mathbf{X}_i)] \leq (n\lambda h^d)^{-1}$$
$$\mathrm{Var}(\tilde{\vartheta}_k) = \mathbb{E}[\langle M\rangle_{\tau}]/(n\lambda h^d)^2.$$

As a result,

$$\mathrm{Var}(\tilde{\vartheta}_k) \le \sum_{i \in \tilde{I}_k^d} \frac{\mathbb{E}[Y_k^* + R\sqrt{d}h^\beta - \max_i g(\mathbf{X}_i)]}{\lambda(nh^d)^2} \omega(\mathbf{X}_i)^2 \le \frac{2R\sqrt{d}h^\beta + (n\lambda h^d)^{-1}}{\lambda(nh^d)^2} \sum_{i \in \tilde{I}_k^d} \omega(\mathbf{X}_i)^2.$$

which again from the fact that $\tilde{\vartheta}_k$ are independent the claimed bound for $\mathrm{Var}(\tilde{\vartheta}_n^{block})$ follows. □

Remark 2.4.6. The results and proof for $\varepsilon_i \sim Exp(\lambda)$, $i = 1, \dots, n$, are exactly as in the PPP model. For other distributions of $(\varepsilon_i)_{1 \le i \le n}$, the estimator is only asymptotically unbiased. In one-dimensional case, for $\beta > 1/2$ the bias is negligible with respect to the stochastic error.

2.4.3 Adaptive estimation in the regression-type model

We now address the question of choosing the block size h in a data-driven way, not assuming the regularity parameters R and β to be known. For the block estimators $\hat{\vartheta}^{block}$ the construction process primarily depends on the values R and β. A misspecification of β for $\hat{\vartheta}^{block}$, i.e. using $\tilde{\beta} > \beta$ instead, results in the block-wise upper bound $\mathscr{Y}_k + R\sqrt{d}h^{\tilde{\beta}}$ which is producing a bias of the maximal size $R\sqrt{d}(h^\beta - h^{\tilde{\beta}})$. Then with the rate optimal choice of h, which is $n^{-1/(\beta+d)}$, the resulted upper bound for a bias is $O\Big((n^{-\beta/(\tilde{\beta}+d)} - n^{-\tilde{\beta}/(\tilde{\beta}+d)})_+\Big)$, the variance is of the order $n^{-(2\tilde{\beta}+d)/(\tilde{\beta}+d)}$. The bandwidth h is another unknown quantity that needs to be specified when constructing both the block-wise estimator and MLE. See [41, 73] for a more detailed background on the bandwidth selection problem for regression type models under one-sided errors. The data driven approaches are not that well-spreaded in the literature. One of the reasons for this to happen partially due to nonmonotonicity of the approximation error terms. This prevents well-known concepts such as cross-validation being used for mean regression. For more information, refer to [52]. In [19] the authors introduce β-adaptive minimax optimal estimator. They use the Bayesian approach and are relying on the assumption that the law of the error terms is known. Further, in [40] the adaptive methods are studied as well for nonparametric problems. The adaptation problem consists of finding an asymptotically optimal length of the side of the cube when neither the regression function g nor the specific boundary behaviour of the errors is known. More precisely, we consider the question of choosing the length of the cube's side size h, or volumes h^d, in a data-driven way, without an assumption that the regularity parameters R and β are known. We follow the method inaugurated by Lepskii in [82] and consider geometrically growing volumes with $h_m^d \sim h_0^d q^m$, for $m = 1, \dots, M$, $q > 1$, $M \sim \log n$. The basic idea is to increase the volume h_m^d as long as the distance between the estimators is not significantly larger than the usual stochastic fluctuations of the

estimators such that at $\widehat{h}^d$, which is specified below, the bias is not yet dominating. We should also notice that an explicit non-asymptotic risk analysis is evidently possible, we concentrate only on the asymptotic risk here. The results on risk's upper bounds in the non-asymptotic case are covered for both block-wise estimation methods in Section 2.4.1 and Section 2.4.2, refer to Theorem 2.4.1 and Theorem 2.4.3 respectively. Further, in Section 2.5.1 we introduce a maximum likelihood estimator and provide similar results on upper bounds in Theorem 2.5.2. In this section we prove that by the versatility of Lepski's method rate-optimal adaptive estimation is possible up to logarithmic factors in the current non-regular situation.

With a choice $n^{-1}(\log n)^2 \le h_1^d < \ldots < h_M^d \le 1$ of bandwidth h_m with $h_m^{-d}, nh_m^d \in \mathbb{N}$ consider the corresponding block-wise estimators

$$\tilde{\vartheta}^{block}_{n,h_m} := \frac{1}{n}\sum_{k=1}^{h_m^{-d}} \sum_{i\in I_{k,h_m}} \Big(\mathscr{Y}_i \wedge (\mathscr{Y}^*_{k,h_m} + (nh_m^d)^{-1}) - \lambda^{-1}\mathbf{1}(\mathscr{Y}_i \le \mathscr{Y}^*_{k,h_m} + (nh_m^d)^{-1})\Big)\omega(\mathbf{X}_i),$$

where the subscript h_m in I_{k,h_m} represents all quantities depending on the length of the cube side. The quantity $R\sqrt{d}h^\beta$ in (2.4.13) is replaced with $(nh_m^{-d})^{-1}$, which does not depend on the unknown R and β and at the same time asymptotically balanced in size. As the next step we select among $(h_m)_{1\le m\le M}$ the length of the cube side in an adaptive way as

$$\widehat{h}^d := \inf\{h_m^d | \exists m^{'} \le m : |\tilde{\vartheta}^{block}_{n,h_{m^{'}}} - \tilde{\vartheta}^{block}_{n,h_{m+1}}| > \kappa_{m+1} + \kappa_{m'}\} \wedge h_M^d,$$

where (κ_m) denote critical values and they are given by

$$\kappa_m = \sum_i \Big(\mathscr{Y}_i \wedge (\mathscr{Y}^*_{k_i,h_m} + (nh_m^d)^{-1})\frac{H_{\sqrt{c\log n}}(h_m^{d/2}\omega(\mathbf{X}_i))}{n\lambda h_m^{d/2}} + \frac{(Cc\log n)^2 \parallel \omega \parallel_1}{n^2\lambda h_m^{2d}} + \frac{\sqrt{c\log n}}{2n\lambda h_m^{d/2}},$$

where k_i is the cube number k with $i \in \tilde{I}_k^d$. We specify $c > 0$ below, the function $H_x(y) = \frac{\log(1-2x|y|)}{-2x} - |y|$ and the constant $C > 0$ with the following property $|\lambda z + \log \bar{F}_\varepsilon(z)| \le C^2 z^2$ for $z \in [0, \delta]$ are used. In the asymptotic case it holds that $H_x(y) \approx xy^2$ as $xy \to 0$ and $C \approx -(f_\varepsilon^{'}(0) + f_\varepsilon^2(0))$ in the case when the density function f_ε of ε is differentiable around zero (also note that $C = 0$ for $\varepsilon \sim Exp(\lambda)$).

The main technical work is devoted to obtaining explicit critical values, and it is done in Proposition 2.7.3. To this end, the critical values are defined via the compensator of an exponential counting process and are thus itself again stochastic. The proof of the following result for the risk bound can be found in Section 2.7.1.

Theorem 2.4.7. *Let $g \in C^\beta(R)$ and $\sup_x |\omega(x)| < \infty$. Assuming that also $f_\varepsilon/\bar{F}_\varepsilon$ is bounded, then for $\underline{\lambda} \in (0, \lambda)$ and n sufficiently large the adaptive estimator $\tilde{\vartheta}^{block}_n = \tilde{\vartheta}^{block}_{n,\hat{h}}$*

together with

$$h^* := \sup\{h_m | R(\sqrt{d}h_m)^\beta \leq (nh_m^d)^{-1}\} \vee h_1$$

satisfies

$$\mathbb{E}[(\tilde{\vartheta}_n^{block} - \tilde{\vartheta}^{(n)})^2 \mathbf{1}(\widehat{h} < h^*)] \lesssim M(n^{-c} + n^{(1-\underline{\lambda}c)/2}),$$

$$\mathbb{E}[(\tilde{\vartheta}_n^{block} - \tilde{\vartheta}_{n,h^*}^{block})^2 \mathbf{1}(\widehat{h} \geq h^*)] \lesssim \frac{(\log n)^4}{(n(h^*)^d)^4} + \frac{M \log n}{n^2 (h^*)^d}.$$

By choosing $c > 5\lambda^{-1} \vee 2$ *and also asymptotically* $h_0^d \sim (\log n)^2 n^{-1}$, $h_m^d \sim h_0^d q^m$ *for* $m = 1, \ldots, M$, $q > 1$, $M \sim \log n$, *for the asymptotic rate of the adaptive estimator the following result holds*

$$\mathbb{E}[(\tilde{\vartheta}_n^{block} - \tilde{\vartheta}^{(n)})^2] \lesssim (\log n)^2 n^{-(2\beta+d)/(\beta+d)} + (\log n)^4 n^{-4\beta/(\beta+d)}.$$

In particular, for $\beta \geq 1/2$ *the estimator achieves the minimax optimal rate up to a logarithmic factor in the one-dimensional case. When* $d > 1$ *the estimator has the rate* $n^{-4\beta/(\beta+d)}$ *which appears to be not optimal.*

Remark 2.4.8. As it follows from the proof, when $\varepsilon \sim Exp(\lambda)$) not only the critical values, but also the bounds obtain a simpler representation. We have

$$\mathbb{E}[(\tilde{\vartheta}_n^{block} - \tilde{\vartheta}_{n,h^*}^{block})^2 \mathbf{1}(\widehat{h} \geq h^*)] \lesssim \frac{M \log n}{n^2 (h^*)^d}$$

and, as a result, for any $\beta > 0$ the minimax optimal rate is achieved up to a logarithmic factor and in all dimensions.

Remark 2.4.9. In the case of $d = 1$ for the asymptotic rate we just note that the geometric grid of bandwidths needs to achieve the following rate $h^* \sim n^{-1/\beta+1}$ asymptotically, then, by inserting we have the following bound

$$|\mathbb{E}[\tilde{\vartheta}_n^{block} - \vartheta^{(n)}]| \lesssim n^{-(2\beta+1)/(\beta+1)} + n^{-4\beta/(\beta+1)}.$$

From where we can infer that once $n^{-(2\beta+1)/(\beta+1)} \geq n^{-4\beta/(\beta+1)}$ then the estimator achieves the minimax optimal rate up to a logarithmic factor in the one-dimensional case. This is exactly the case when $(2\beta+1)/(\beta+1) \leq 4\beta/(\beta+1)$ what is equivalent to $\beta \geq 1/2$. To achieve the minimax optimal rate in dimensions higher than 1, β would need to be bigger than 1, which is not the case for the class of functions being considered.

Remark 2.4.10. Adaptive procedure is a comprehensive topic, which can be applied to the higher-dimensional setting, because essentially it depends on the errors and the response variables, which are one-dimensional quantities. Note that different approaches to this problem may be considered, such as a theoretically lighter cross validation approach or a bootstrap method, refer to [52] for more details.

2.4.4 Rate optimality for the Poisson point process model

In this section we demonstrate that the rate $R^{d/\beta+d}n^{-(2\beta+d)/\beta+d}$ is optimal in a minimax sense over $C^\beta(R)$. The proof is provided only for the PPP model. The case of the regression-type model with $\varepsilon \sim Exp(\lambda)$ can be treated analogously.

Theorem 2.4.11. *For the problem of estimation $\vartheta = \int_{[0,1]^d} g(x)\omega(x)dx$, where $\omega \in L^2([0,1]^d)$, in the PPP model with parameter class $C^\beta(R)$, with $\beta \in (0,1]$ and $R > 0$ the following result for the asymptotic lower bound holds*

$$\liminf_{n\to\infty} \inf_{\hat{\vartheta}_n} \sup_{g\in C^\beta(R)} R^{-d/\beta+d} n^{-(2\beta+d)/\beta+d} \|\omega\|_{L^2}^{-2} \mathbb{E}_g[(\hat{\vartheta}_n - \vartheta)^2] > 0.$$

This infimum expands over all estimators $\hat{\vartheta}_n$ from the PPP model with the intensity measure provided by (2.3.1).

Proof. The proof is based on the approach provided in [74]. The approach is based on a Bayesian risk bound and it grants a lower bound for the minimax risk. Another possible way to prove the minimax lower bound is based on Yang-Barron version of Fano's method, see again [74]. This method is particularly useful for nonparametric problems, since it obviates the need for constructing a local packing.

First, we start with taking an independent Bernoulli sequence ε_k, i.e. $P(\varepsilon_k = 1) = p$, $P(\varepsilon_k = 0) = 1 - p$ with $p \in (0,1)$ and consider d one-dimensional triangular kernels

$$K_i(x_i) = 2\min(x_i, 1 - x_i)\mathbf{1}_{[0,1]}(x_i),$$

$i = 1, \dots, d$. As the next step we take the product kernel $K(y) = \prod_{i=1}^d K_i(x_i)$, $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ and put

$$g(x) = \textstyle\sum_{k=0}^{h^{-1}-1} \varepsilon_k g_k(x) \text{ with } g_k(x) = cRh^\beta \prod_{i=1}^d K_i((x_i - kh)/h),$$

where $h \in (0,1)$ with $h^{-d} \in \mathbb{N}$ that we specify later.

For $c > 0$ and at the same time sufficiently small, for all h and all realisations of (ε_k) it holds that $g \in C^\beta(R)$. We refer to this representation of g as a prior on $C^\beta(R)$. We use the property of the independence of the prior as well as the observation laws on different cubes. From the Bayes formula for each k we infer the posterior probability given the observations of the PPP in cubes I_k^d

$$\hat{\varepsilon}_k := \mathbb{E}[\varepsilon_k|(X_i, Y_i)_{i\geq 1}] = P(\varepsilon_k = 1|(X_i, Y_i)_{i\geq 1}) = \frac{p(dP_{g_k}/dP_0)(dP_{g_0}/dP_0)^{-1}}{(1-p) + p(dP_{g_k}/dP_0)(dP_{g_0}/dP_0)^{-1}}$$
$$= \frac{pe^{n\int_{I_k^d} g_k(x)}}{1 - p + pe^{n\int_{I_k^d} g_k(x)}} \mathbf{1}(\forall X_i \in I_k^d : Y_i \geq g_k(X_i)).$$

From the fact that each ε_k are either 0 or 1-valued random variables, we obtain the following results for conditional expectation and variance of ε_k

$$\hat{\varepsilon}_k = \mathbb{E}[\varepsilon_k|(X_i, Y_i)_{i\geq 1}], \quad \text{and} \quad \mathrm{Var}(\varepsilon_k|(X_i, Y_i)_{i\geq 1}) = \hat{\varepsilon}_k(1-\hat{\varepsilon}_k).$$

From what it follows the Bayes-optimal estimator of ϑ_2 under squared loss is given by the posterior mean

$$\hat{\varepsilon} = \sum_{k=0}^{h^{-d}-1} \hat{\varepsilon}_k \int_{I_k^d} g_k(x)\omega(x)dx.$$

By the independence and the fact that $\mathbb{E}[\hat{\varepsilon}_k - \varepsilon_k] = 0$, its Bayes risk is calculated in the following way

$$\begin{aligned}
\mathbb{E}[(\hat{\varepsilon}_k - \varepsilon_k)^2] &= \sum_{k=0}^{h^{-d}-1} \mathrm{Var}(\hat{\varepsilon}_k - \varepsilon_k)\Big(\int_{I_k^d} g_k(x)\omega(x)dx\Big)^2 \\
&= \sum_{k=0}^{h^{-d}-1} \mathbb{E}[\mathrm{Var}(\varepsilon_k|(X_i, Y_i)_{i\geq 1})]\Big(\int_{I_k^d} g_k(x)\omega(x)dx\Big)^2 \\
&= \sum_{k=0}^{h^{-d}-1} \frac{p(1-p)}{(1-p+pe^{n\int g_k})^2}\Big(\int_{I_k^d} g_k(x)\omega(x)dx\Big)^2.
\end{aligned}$$

By picking $h = \lceil (cRn)^{1/(\beta+d)} \rceil^{-1}$ such that

$$n\int_{I_k^d} g_k(x)\omega(x)dx = cRh^{\beta+d}n \leq 1$$

holds, the Bayes risk is bounded in order by

$$\mathbb{E}[(\hat{\varepsilon}_k - \varepsilon_k)^2] \gtrsim R^2 h^{2\beta+d} \sum_{k=0}^{h^{-d}-1} \Big(\int_{I_k^d} \frac{K((x-kh)/h)}{\|K((\bullet - h)/h)\|_{L^2}}\omega(x)dx\Big)^2.$$

The same argument over the shifted cubes $\hat{I}_k^d = [(k+1/2)h, (k+3/2)h)^d$ suggests that the minimax risk is bounded by the maximum, and therefore the average, over the respective Bayes risks

$$\inf_{\hat{\vartheta}_n} \sup_{g\in C^\beta(R)} \mathbb{E}[(\hat{\varepsilon}_k - \varepsilon_k)^2] \gtrsim R^2 h^{2\beta+d} \sum_{k=0}^{h^{-d}-1} \Big(\int_{I_k^d} \frac{K((x-kh)/h)}{\|K((\bullet - h)/h)\|_{L^2}}\omega(x)dx\Big)^2.$$

Using generalisation of the orthonormal basis in $\mathbb{R}^d$ it can be shown that the sum in the above inequality is bounded from below by L^2-norm of ω. Insertion of $h \sim (Rn)^{-1/\beta+d}$

gives the result

$$\liminf_{n\to\infty} \inf_{\hat{\vartheta}_n} \sup_{g\in C^\beta(R)} R^{-d/\beta+d} n^{-(2\beta+d)/\beta+d} \|\omega\|_{L^2}^{-2} \mathbb{E}_g[(\hat{\vartheta}_n - \vartheta)^2] > 0.$$

□

2.5 Nonparametric maximum-likelihood

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

In this section we study the nonparametric maximum-likelihood estimator (MLE) in the class $C^\beta(R)$. Consider the PPP model with the intensity $\lambda = n\mathbf{1}(y \geq g(x))$ and denote by P_g the law of the observations in this model. Then the Radon-Nikodym derivative $\frac{dP_g}{dP_{g_0}}$ for $g \geq g_0$ is by (a minor generalisation of) Theorem 1.3 in [77] and the fact that the PPP intensities coincide outside the compact set $[0,1]^d \times [\min g_0, \max g]$ we get the Radon-Nikodym derivative

$$\frac{dP_g}{dP_{g_0}} = \exp\Big(n\int_{[0,1]^d} (g-g_0)(x)dx\Big)\mathbf{1}\Big(\forall j : Y_j \geq g(X_j)\Big). \tag{2.5.1}$$

A simple probability measure dominating all P_g, $g \in C^\beta(R)$ (note that g does need to be bounded from below), is given by the PPP model with intensity $\lambda_0(x,y) = n$ and yields again via Theorem 1.3 in [77] the likelihood

$$\mathcal{L}(g) = \exp\Big(n\int_{[0,1]^d} g(x)dx\Big)\mathbf{1}\Big(\forall j : Y_j \geq g(X_j)\Big). \tag{2.5.2}$$

The MLE over $C^\beta(R)$ is the function $\hat{g}$ that maximises $\int_{[0,1]^d} g$ over all $g \in C^\beta(R)$ with the property that $g(X_j) \leq Y_j$ for all $j \geq 1$. We can write explicitly

$$\hat{g}^{MLE}(x) = \min_{j\geq 1}\Big(Y_j + R\|x - X_j\|_2^\beta\Big), \tag{2.5.3}$$

as the right-hand side of the above equation even maximises $g(x)$ point-wise over the considered class of functions g. Refer to Figure 2.4 for an illustration of the construction of the estimator $\hat{g}^{MLE}(x)$.

Next, in an analogous way the corresponding likelihood function with respect to the n-dimensional Lebesgue measure for the regression-type model (2.3.3) with error terms, independent and identically distributed random variables distributed according to the

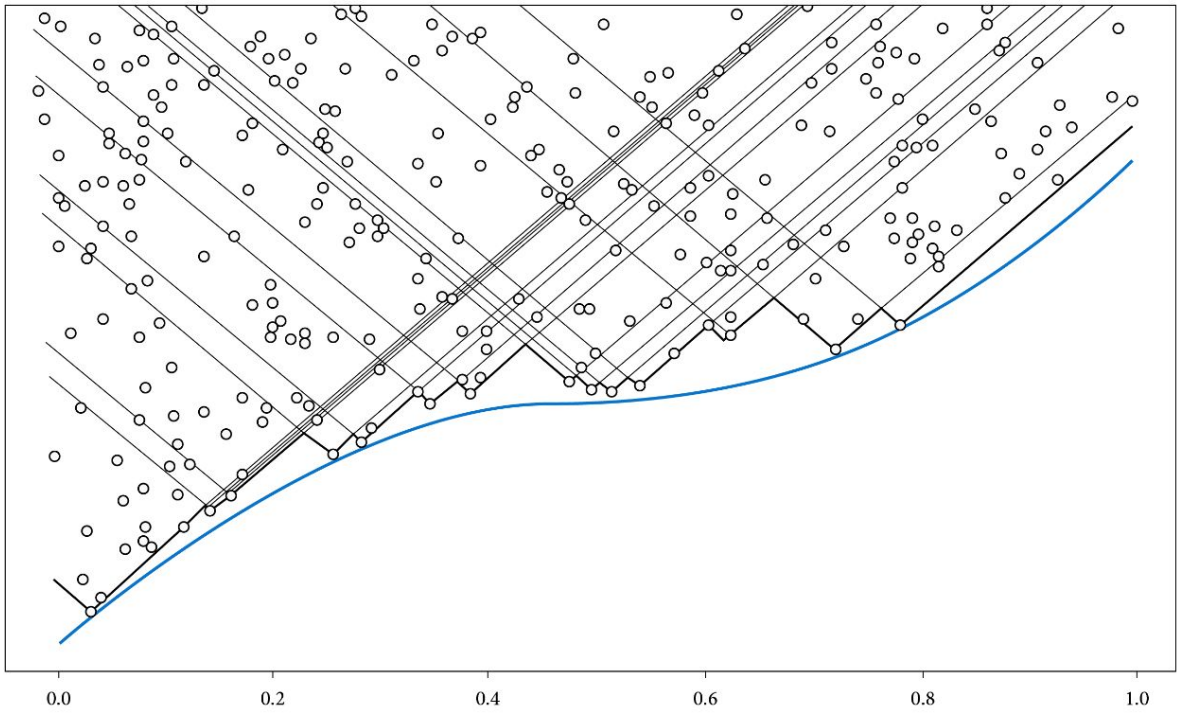


Figure 2.4: Construction of the MLE estimator $\hat{g}^{MLE}(x)$ in the PPP model in the one-dimensional case, $n = 150$. The thin straight lines correspond to $x \to Y_i + R|x - X_i|$. The blue smooth line is the function g.

exponential law $\varepsilon \sim Exp(\lambda)$, can be written in the following way

$$\mathcal{L}^{regr}(g) = \lambda^n \exp\Big(-\lambda \sum_{i=1}^{n} \mathscr{Y}_i\Big) \exp\Big(-\lambda \sum_{i=1}^{n} g_i(\mathbf{X}_i)\Big) \mathbf{1}\Big(\forall i = 1, ..., n : \mathscr{Y}_i \geq g(\mathbf{X}_i)\Big).$$

One can then infer the maximum-likelihood estimator over $C^\beta(R)$

$$\hat{g}^{MLE-regr}(x) = \min_{i=1,...,n} \big(\mathscr{Y}_i + R\|x - \mathbf{X}_i\|_2^\beta\big), \quad x \in [0,1]^d,$$

and $1 \leq i \leq n$.

An illustration of the construction of MLE $\hat{g}^{MLE-regr}(x)$ for the regression model is demonstrated in Figure 2.5.

Both MLE estimators for PPP and the regression-type models are quickly computed numerically. Further, in this section we focus on the MLE in the PPP model and only briefly comment on the results for the regression-type model and in the settings of the exponential noise. The analysis of MLE under non-exponential noise in the regression-type model is completely omitted, as the results must be asymptotic in nature and will be comparable to Theorem 2.4.3. The next Proposition asserts that the introduced above MLE is a sufficient and complete statistic.

Proposition 2.5.1. *The nonparametric MLE* $(\hat{g}^{MLE}(x), x \in [0,1]^d)$ *is a sufficient and complete statistic for the parameter class* $C^\beta(R)$.

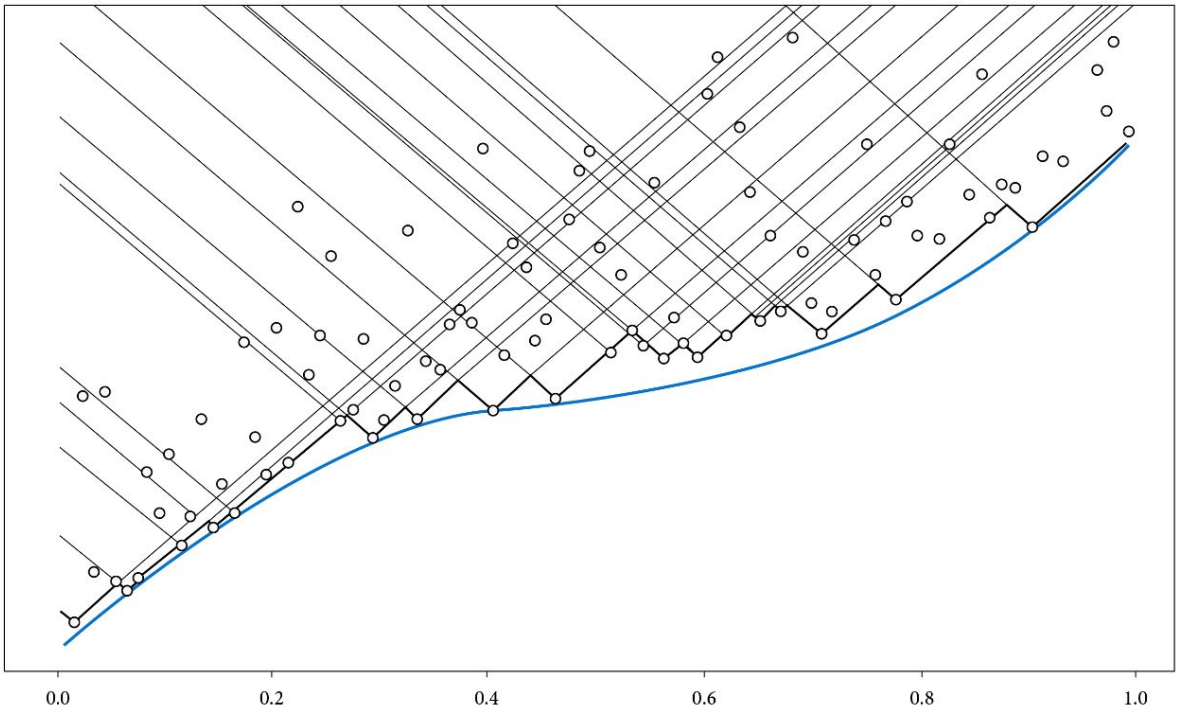


Figure 2.5: Construction of the MLE $\hat{g}^{MLE-regr}(x)$ in the regression-type model. We consider the case of $\varepsilon \sim Exp(1)$ and $\beta = 1$.

Proof. Sufficiency of $\hat{g}^{MLE}(x)$ follows directly by using Neyman factorisation criterion (e.g. [81]) for the likelihood (2.5.2).

Let us first note that by definition $\hat{g}^{MLE}(x)$ is an element of $C^\beta(R)$. As $C^\beta(R)$, equipped with its $C^\beta(R)$-norm, is not separable, we equip it with the Borel σ-algebra generated by the uniform (supremum) norm, which is generated by all point evaluations. We can establish the measurability of the estimator $\hat{g}^{MLE}(x)$ is established since all point evaluations of $\hat{g}^{MLE}(x)$, $x \in [0,1]^d$, are measurable as a minimum of countably many random variables.

To prove the result of completeness we now shall consider any statistic $T : C^\beta(R) \to \mathbb{R}$ that satisfies the following condition

$$\mathbb{E}_g[T(\hat{g}^{MLE})] = 0 \quad \text{for all} \quad g \in C^\beta(R),$$

and which is Borel measurable with respect to the uniform norm. For $g \in C^\beta(R)$ we denote by $[g, \infty) := \{h \in C^\beta(R) | h \geq g\}$ the bracket between g and ∞, that is all the functions that are above g. Next, for some $g \in C^\beta(R)$ and $h \in C^\beta(R)$ taking the intersection $[g, \infty) \cap [h, \infty) = [g \vee h, \infty)$ we see that the maximum $g \vee h$ is again in $C^\beta(R)$. This suggests that the family $\{[g, \infty) | g \in C^\beta(R)\}$ is an $\cap$-stable generator of the uniform Borel $\sigma-$algebra in $C^\beta(R)$. By the Hölder condition we have that $\{h \in C^\beta(R) | h(x_0) \geq y_0\} = [y_0 - R\|\bullet - x_0\|^\beta, \infty)$ for any $x_0 \in [0,1]$ and $y_0 \in \mathbb{R}$ and $\{[y_0, \infty) | y_0 \in \mathbb{R}\}$ generates the Borel $\sigma-$algebra on $\mathbb{R}$.

Using the likelihood function under P_0 and the fact that $\mathbb{E}_g[T(\hat{g}^{MLE})] = 0$ we get

$$e^{n\int(g+1)}\mathbb{E}_0\Big[T(\hat{g}^{MLE})e^{\sum_{j\geq 1}(-Y_j)_+}\mathbf{1}(\hat{g}^{MLE} \in [g,\infty))\Big] = 0.$$

If we split our statistic T into two parts in the following way $T = T^+ - T^-$, where T^+, T^- are both non-negative functions, and introduce the following two measures for Borel sets B as

$$B \to \mathbb{E}_0\Big[T^{\pm}(\hat{g}^{MLE})e^{\sum_{j\geq 1}(-Y_j)_+}\mathbf{1}(\hat{g}^{MLE} \in B)\Big],$$

we can infer that these measures $B \to \mathbb{E}_0\Big[T^{\pm}(\hat{g}^{MLE})e^{\sum_{j\geq 1}(-Y_j)_+}\mathbf{1}(\hat{g}^{MLE} \in B)\Big]$ coincide on $\{[g,\infty)|g \in C^{\beta}(R)\}$ and, as a result, by the uniqueness theorem for all uniform Borel sets B in $C^{\beta}(R)$ and in particular for $B = \{T > 0\}$ and $B = \{T < 0\}$. This in turn implies that

$$T^+(\hat{g}^{MLE})e^{\sum_{j\geq 1}(-Y_j)_+} = T^-(\hat{g}^{MLE})e^{\sum_{j\geq 1}(-Y_j)_+}, \quad P_0\text{-a.s.}$$

and thus, we showed that $T(\hat{g}^{MLE}) = 0$ P_g-a.s. for all $g \in C^{\beta}(R)$, which concludes the proof of the result. □

Similarly to the construction of the block-wise estimator $\hat{\vartheta}^{block}$ for the functional (2.4.1), we introduce here the maximum likelihood estimator of the linear functional in the following way

$$\hat{\vartheta}^{MLE} := \int_{[0,1]^d} \hat{g}^{MLE}(x)\omega(x)dx - \frac{1}{n}\sum_j \mathbf{1}\big(\hat{g}^{MLE}(X_j = Y_j)\big)\omega(X_j). \tag{2.5.4}$$

As can be observed from the representation (2.5.4) the MLE $\hat{\vartheta}^{MLE}$ is obtained by a simple plug-in of the nonparametric MLE $\hat{g}^{MLE}$ into the functional (2.4.1) with a further subtraction of $\frac{1}{n}\sum_j \mathbf{1}\big(\hat{g}^{MLE}(X_j = Y_j)\big)\omega(X_j)$. This subtraction serves as a bias correction term and it counts the relative number of observations on the graph of $\hat{g}^{MLE}$. This produces an unbiased estimator of the functional (2.4.1). However, the main result is that this suggested estimator $\hat{\vartheta}^{MLE}$ is not only unbiased, but also uniformly of minimum variance among all unbiased estimators for the class $C^{\beta}(R)$. We will further refer to this last property of the estimator (2.5.4) as UMVU (uniformly of minimum variance among all unbiased estimators). Now we formulate this result as a part of the next theorem, providing the result for an upper bound and asymptotic upper bound of the variance of the estimator as well.

Theorem 2.5.2. *Let n be the sample size of $\{(X_j, Y_j)\}_{j\geq 1}$. Then, for each finite n the estimator $\hat{\vartheta}^{MLE}$ from* (2.5.4) *is an unbiased estimator of the functional* (2.4.1). *Furthermore, the estimator $\hat{\vartheta}^{MLE}$ is uniformly of minimal variance over the class $C^{\beta}(R)$ and the*

following upper bound holds

$$\begin{aligned}\mathrm{Var}(\hat{\vartheta}^{MLE}) &= \frac{1}{n}\int_{[0,1]^d}\mathbb{E}[\hat{g}^{MLE}(x)-g(x)]\omega(x)^2dx \\ &\leq \Big(\Gamma(\beta/(\beta+d))\beta(2R\beta/(\beta+d))^{d/(\beta+d)}n^{-(2\beta+d)/(\beta+d)} + \frac{1}{n^2}e^{-2\beta Rn/(\beta+d)}\Big)\|\omega\|^2_{L^2}.\end{aligned}$$

In the asymptotic case, when $n\to\infty$ we get

$$\mathrm{Var}(\hat{\vartheta}^{MLE}) \leq (2+o(1))R^{d/(\beta+d)}n^{-(2\beta+d)/(\beta+d)}\|\omega\|^2_{L^2}.$$

Remark 2.5.3. The result for non-asymptotic rates of converges is based on the following deviation inequality that holds for the estimator $\hat{g}^{MLE}(x)$

$$P(\hat{g}^{MLE}(x)-g(x)\geq s) \leq \begin{cases}\exp(-n\frac{2R}{(\beta+d)}(s/2R)^{(\beta+d)/\beta}) & \text{if } s\in[0,2Rd^{\beta/2}),\\ \exp(-n(s-2Rd/(\beta+d))) & \text{if } s>2Rd^{\beta/2}.\end{cases}$$

Thus, we are able to demonstrate the finiteness of the variance of $\hat{\vartheta}^{MLE}$.

Proof. Similar to the proof of Theorem 2.4.1 we first define a weighted counting process in the following way

$$\bar{N}(t) = \textstyle\sum_{j\geq 1}\mathbf{1}\Big(Y_j \leq t\wedge\min(Y_i+R\|X_j-X_i\|_2^\beta)\Big)\omega(X_j),\quad t\in\mathbb{R}.$$

One can notice that for every t the part inside the brackets $\{\min_{i\geq 1}(Y_i+R\|X_j-X_i\|_2^\beta)<t\}$ is identical to $\{\min_{i:Y_i<t}(Y_i+R\|X_j-X_i\|_2^\beta)<t\}$. Again, as in the proof of Theorem 2.4.1, consider a pure counting process, i.e., when $\omega\equiv 1$, then this counting process has the following intensity measure

$$\bar{\lambda}_t = n\int_{[0,1]^d}\int_{[g(x),t]}\mathbf{1}\Big(\min_{i:Y_i<t}(Y_i+R\|x-X_i\|_2^\beta)\geq s\Big)dsdx.$$

Hence, $\bar{N}$ is adapted to the filtration $(\mathcal{F}_t)$, which is defined exactly as in the proof of Theorem 2.4.1, i.e.

$$\mathcal{F}_t = \sigma((X_i,Y_i)\mathbf{1}(Y_i\leq t), i\geq 1)\quad t\in\mathbb{R}.$$

By (Proposition 2.32 in [66]) the following process

$$\bar{M}(t) = \bar{N}(t) - n\int_{[0,1]^d}\int_{[g(x),t]}\mathbf{1}\Big(\min_{i:Y_i<t}(Y_i+R\|x-X_i\|_2^\beta)\geq s\Big)ds\omega(x)dx$$

is an $(\mathcal{F}_t)$-martingale. By taking the difference between the martingale $\bar{M}(t)$ and the

counting process $\bar{N}(t)$ and sending the time t to infinity we obtain

$$\begin{aligned}\lim_{t\to\infty}(\bar{M}(t)-\bar{N}(t)) &= n\int_{[0,1]^d}\int_{g(x)}^{\infty}\mathbf{1}\Big(\min_{i:Y_i<t}(Y_i+R\|x-X_i\|_2^{\beta})\geq s\Big)ds\omega(x)dx\\ &= n\int_{[0,1]^d}\int_{g(x)}^{\infty}\mathbf{1}(\hat{g}^{MLE}(x)\geq s)ds\omega(x)dx\\ &= n\int_{[0,1]^d}\mathbf{1}(\hat{g}^{MLE}(x)-g(x))\omega(x)dx,\end{aligned}$$

and observing that

$$\bar{N}(\infty):=\lim_{t\to\infty}\bar{N}(t)=\sum_{j\geq 1}\mathbf{1}(\hat{g}^{MLE}(x)\geq Y_j)\omega(X_j)=\sum_{j\geq 1}\mathbf{1}(\hat{g}^{MLE}(x)=Y_j)\omega(X_j),$$

we conclude that the counting process counts the weighted number of points (X_j,Y_j) on the graph of $\hat{g}^{MLE}(x)$ and equals to the scaled bias $n(\int_{[0,1]^d}\hat{g}^{MLE}(x)\omega(x)dx-\vartheta)$ up to a martingale term. As a result, we have

$$\hat{\vartheta}^{MLE}=\int_{[0,1]^d}\hat{g}^{MLE}(x)\omega(x)dx-\frac{1}{n}\bar{N}(\infty)=\vartheta-\frac{1}{n}\bar{M}(\infty),$$

where

$$\bar{M}(\infty)=\sum_{j\geq 1}\mathbf{1}(\hat{g}^{MLE}(x)\geq Y_j)\omega(X_j)-\int_{[0,1]^d}\mathbf{1}(\hat{g}^{MLE}(x)-g(x))\omega(x)dx$$

is a.s. an L^2-limit of the L^2 bounded martingale $\bar{M}$. The quadratic variation of $\bar{M}$ is given by

$$\langle\bar{M}\rangle_t=n\int_{[0,1]^d}\int_{g(x)}^{t}\mathbf{1}\Big(\min_{i:Y_i<t}(Y_i+R\|x-X_i\|_2^{\beta})\geq s\Big)ds\omega(x)^2dx.$$

Sending t to the infinity we get

$$\langle\bar{M}\rangle_t\uparrow n\int_{[0,1]^d}(\hat{g}^{MLE}(x)-g(x))\omega(x)^2dx=:\langle\bar{M}\rangle_{\infty}\quad\text{as } t\uparrow\infty.$$

Using the result of Theorem (2.7.1) in the case of $\tau=\infty$ together with the following identity

$$\mathbb{E}[\hat{\vartheta}^{MLE}-\vartheta]=\frac{1}{n}\mathbb{E}[-\bar{M}(\infty)],$$

we can conclude that the estimator (2.5.4) is an unbiased estimator of the functional (2.4.1), i.e., it holds

$$\mathbb{E}[\hat{\vartheta}^{MLE}]=\mathbb{E}[\vartheta].$$

Next, based on the same result of Theorem 2.7.1 in the case of $\tau = \infty$ and the fact that

$$\mathrm{Var}(\hat{\vartheta}^{MLE}) = \frac{1}{n^2}\mathrm{Var}(\bar{M}(\infty)),$$

we obtain the following equation for the variance of $\hat{\vartheta}^{MLE}$

$$\mathrm{Var}(\hat{\vartheta}^{MLE}) = \frac{1}{n^2}\mathbb{E}[\langle \bar{M}\rangle_\infty] = \frac{1}{n}\int_{[0,1]^d} \mathbb{E}[\hat{g}^{MLE}(x) - g(x)]\omega(x)^2 dx. \tag{2.5.5}$$

Hence, by Theorem 2.7.10, see [81], the estimator $\hat{\vartheta}^{MLE}$ derived from a sufficient and complete statistics, is uniformly of minimum variance among all unbiased estimators.

Next for the variance bound we use the following rough deviation bound for $s \geq 0$ and $x \in [0,1]^d$

$$\begin{aligned} P(\hat{g}^{MLE}(x) - g(x) \geq s) &= \exp\Big(- n\int_{[0,1]^d} (s - R\|\xi - x\|_2^\beta + g(x) - g(\xi))_+ d\xi\Big) \\ &\leq \exp\Big(- n\int_{[0,1]^d} (s - 2R\|\xi - x\|_2^\beta)_+ d\xi\Big). \end{aligned}$$

Let us consider the following integral $J = \int_{[0,1]^d} (\frac{s}{2R} - \|\xi - x\|_2^\beta)_+ d\xi$, then

$$J = \int_0^1 ... \int_0^1 \Big(\frac{s}{2R} - \big((\xi_1 - x_1)^2 + ... + (\xi_d - x_d)^2\big)^{\beta/2}\Big)_+ d\xi_1 ... d\xi_d.$$

In d-dimensional spherical coordinates, in which the coordinates consist of a radial coordinate, r which ranges over $[0, r_0]$ with $r_0 = d^{1/2} \geq 1$, and $n-1$ angular coordinates φ_1, φ_2, ..., φ_{d-1} where φ_{d-1} ranges over $[0, 2\pi)$ radians and the other angles range over $[0, \pi]$, the integral has the following form

$$\begin{aligned} J &= \int_0^{2\pi}\int_0^{\pi} ... \int_0^{\pi}\int_0^{r_0} \Big(\frac{s}{2R} - r^\beta\Big)_+ r^{d-1} \sin^{d-2}\sin^{d-2}\varphi_1 \sin^{d-3}\varphi_2 ... \sin\varphi_{d-2} dr d\varphi_1 d\varphi_2 ... d\varphi_{d-1} \\ &= \int\int_0^{\left(s/2R\right)^{\frac{1}{\beta}}} \Big(\frac{s}{2R} - r^\beta\Big) r^{d-1} \Upsilon dr d\upsilon, \quad \text{for } s \in [0, 2Rr_0^\beta), \end{aligned}$$

where we denoted $\sin^{d-2}\sin^{d-2}\varphi_1 \sin^{d-3}\varphi_2 ... \sin\varphi_{d-2}$ as Υ and $d\varphi_1 d\varphi_2 ... d\varphi_{d-1}$ as $d\upsilon$ respectively. Then, taking the integral over r

$$J = \int \Big(\frac{1}{d}\Big(\frac{s}{2R}\Big)^{\frac{\beta+d}{\beta}} - \frac{1}{\beta+d}\Big(\frac{s}{2R}\Big)^{\frac{\beta+d}{\beta}}\Big)\Upsilon d\upsilon = \int \frac{\beta}{d(\beta+d)}\Big(\frac{s}{2R}\Big)^{\frac{\beta+d}{\beta}}\Upsilon d\upsilon.$$

To calculate the integral $\int \Upsilon d\upsilon$ we notice that the volume of the unit d-ball is given by

$$\begin{aligned}V_d &= \int_0^{2\pi}\int_0^{\pi}...\int_0^{\pi}\int_0^{1} r^{d-1}\sin^{d-2}\sin^{d-2}\varphi_1\sin^{d-3}\varphi_2...\sin\varphi_{d-2}drd\varphi_1d\varphi_2...d\varphi_{d-1}\\ &= \frac{1}{d}\int_0^{2\pi}\int_0^{\pi}...\int_0^{\pi}\sin^{d-2}\sin^{d-2}\varphi_1\sin^{d-3}\varphi_2...\sin\varphi_{d-2}drd\varphi_1d\varphi_2...d\varphi_{d-1} = \frac{1}{d}\int \Upsilon d\upsilon.\end{aligned}$$

As a result, we have

$$\int \Upsilon d\upsilon = dV_d = \frac{d\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)} \geq d,$$

since $\frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)} \geq 1$ for $d \geq 0$. The following inequality holds for J when $s \in [0, 2Rr_0^{\beta})$

$$J = \frac{\beta}{d(\beta+d)}\Big(\frac{s}{2R}\Big)^{\frac{\beta+d}{\beta}}\int \Upsilon d\upsilon \geq \frac{\beta}{(\beta+d)}\Big(\frac{s}{2R}\Big)^{\frac{\beta+d}{\beta}}.$$

When $s > 2Rr_0$

$$\begin{aligned}J &= \int_0^{2\pi}\int_0^{\pi}...\int_0^{\pi}\int_0^{r_0}\Big(\frac{s}{2R}-r^{\beta}\Big)r^{d-1}\sin^{d-2}\sin^{d-2}\varphi_1\sin^{d-3}\varphi_2...\sin\varphi_{d-2}drd\varphi_1d\varphi_2...d\varphi_{d-1}\\ &= \Big(\frac{sr_0^d}{2Rd}-\frac{r_0^{\beta+d}}{\beta+d}\Big)\int \Upsilon d\upsilon \geq \frac{sr_0^d}{2R}-\frac{dr_0^{\beta+d}}{\beta+d} \geq \frac{s}{2R}-\frac{d}{\beta+d}.\end{aligned}$$

Now we have the upper bound for the survival function

$$P(\hat{g}^{MLE}(x)-g(x) \geq s) \leq \exp(-2nRJ) \leq \begin{cases}\exp(-n\frac{2R}{(\beta+d)}(s/2R)^{(\beta+d)/\beta}) & \text{if } s \in [0, 2Rr_0^{\beta}),\\ \exp(-n(s-2Rd/(\beta+d))) & \text{if } s > 2Rr_0^{\beta}.\end{cases}$$

Taking the integral of this survival function bound, we get

$$\begin{aligned}\mathbb{E}[\hat{g}^{MLE}(x)-g(x)] &= \int_0^{\infty} P(\hat{g}^{MLE}(x)-g(x) \geq s)\\ &\leq \int_0^{2Rr_0^{\beta}}\exp\Big(-n\frac{2R\beta}{(\beta+d)}(s/2R)^{(\beta+d)/\beta}\Big)ds + \int_{2Rr_0^{\beta}}^{\infty} e^{-n(s-2R/(\beta+d))}ds\\ &\leq \int_0^{2Rr_0^{\beta}}\exp\Big(-n\frac{2R}{(\beta+d)}(s/2R)^{(\beta+d)/\beta}\Big)ds + \int_{2R}^{\infty} e^{-n(s-2R/(\beta+d))}ds\\ &= n^{-\beta/(\beta+d)}\beta(2R/(\beta+d))^{d/(\beta+d)}\int_0^{2Rr_0^{\beta}} e^{-x}x^{-d/(\beta+d)}dx + \frac{1}{n}e^{-2\beta Rn/(\beta+d)}\\ &\leq \Gamma(\beta/(\beta+d))\beta(2R\beta/(\beta+d))^{d/(\beta+d)}n^{-\beta/(\beta+d)} + \frac{1}{n}e^{-2\beta Rn/(\beta+d)}.\end{aligned}$$

Thus, we derive the bound for the variance in (2.5.5)

$$\begin{aligned}\mathrm{Var}(\hat{\vartheta}^{MLE}) &= \frac{1}{n}\int_{[0,1]^d}\mathbb{E}[\hat{g}^{MLE}(x)-g(x)]\omega(x)^2dx\\ &\le \Big(\Gamma(\beta/(\beta+d))(2R\beta/(\beta+d))^{d/(\beta+d)}n^{-1-\beta/(\beta+d)}+\frac{1}{n^2}e^{-2\beta Rn/(\beta+d)}\Big)\|\omega\|^2_{L^2}\\ &= \Big(\Gamma(\beta/(\beta+d))(2R\beta/(\beta+d))^{d/(\beta+d)}n^{-(2\beta+d)/(\beta+d)}+\frac{1}{n^2}e^{-2\beta Rn/(\beta+d)}\Big)\|\omega\|^2_{L^2}\end{aligned}$$

Sending n to infinity we obtain $\mathrm{Var}(\hat{\vartheta}^{MLE}) \le (2+o(1))R^{d/(\beta+d)}n^{-(2\beta+d)/(\beta+d)}\|\omega\|^2_{L^2}$. □

The construction of the estimator $\hat{\vartheta}^{MLE}$ also relies on the values of R and β. From what follows that in the case of using wrong β, i.e. $\tilde{\beta}$, such that $\tilde{\beta} > \beta$ one will end up with an estimation bias. As was mentioned in Section 2.4.3 the size of the bias is bounded above by $R(\tilde{\beta}-\beta)$. Then, since the rate optimal choice for h is $n^{-1/\tilde{\beta}+d}$, it produces a bias upper bound $O\big(\max(0, n^{-\beta/(\beta+d)} - n^{-\tilde{\beta}/(\tilde{\beta}+d)})\big)$ and the variance is of the order $n^{-(2\tilde{\beta}+d)/\tilde{\beta}+d}$.

Example 2.5.4. One of the most important applications of the estimation of linear functionals is called orthogonal series estimators, see [60]. Consider an orthonormal basis in $L^2([0,1])$ defined as $(\psi_k)_{k\ge1}$ as well as the estimator for $g \in C^\beta(R)$, which is defined as $\hat{g}^K = \sum_{k=1}^K \hat{\vartheta}_k\psi_k$ with $\omega = \psi_k$ and $\hat{\vartheta}_k$ estimating the coefficient $\langle g, \psi_k\rangle_{L^2}$. If we use the estimators for g of Hölder continuity β we get the upper bound for the L^2-risk as follows

$$\mathbb{E}\Big[\|\hat{g}_K - \mathbb{E}[\hat{g}_K]\|^2_{L^2}\Big] = \sum_{k=1}^K \mathrm{V}ar(\hat{\vartheta}_k) \lesssim Kn^{-(2\beta+d)/(\beta+d)}.$$

2.5.2 Discussion on central limit theorem

The knowledge of the limiting distribution of an estimator allows us to construct confidence intervals for the estimator. For the block-wise estimator, it can be shown, using independence of the estimators between cubes and applying Lindeberg's theorem, see [2, 3, 13, 100] for a detailed review of the CLT results for martingales, that the limiting distribution is Gaussian. Unfortunately, the same technique cannot be applied for the MLE estimator to derive the central limit theorem, as there is no independence structure in the construction of the estimator. The idea would be to divide the unit cube $[0,1]^d$ into some cubes and on each cube introduce a block-wise MLE, such that the sequence of MLEs are independent random variables. As such, we could establish CLT for the block-wise MLE showing that the Lyapunov condition holds. Then, we could consider the difference between MLE and block-wise MLE and demonstrate that it is of small stochastic order and via Slutsky's lemma the CLT result would follow for the MLE.

However, the proof of this CLT result in multidimensional case, i.e., when $d > 1$ appears to be intricate. Essentially, the main difficulty is to extract geometrical properties

of the estimators so that they could be used to control the difference between the block-wise MLE and MLE. The geometry of the problem is getting more complicated as the dimension grows, and some other methods should be exploited to prove the CLT result in multidimensional case.

2.5.3 Finite sample behaviour

The aim of this section is to provide numerical confirmation of the central limit theorem for the block-wise and nonparametric maximum likelihood estimators in the PPP model. We focus on a numerical study of the following central limit theorem results

$$\chi_n^{(1)} = \frac{n^{1/2}(\widehat{\vartheta}_n^{MLE} - \vartheta)}{\Big(\int_0^1 (\widehat{g}_n^{MLE} - g(x))\omega(x)^2 dx \Big)^{1/2}} \Rightarrow \mathcal{N}(0,1),$$

$$\chi_n^{(2)} = \frac{n^{1/2}(\hat{\vartheta}_n^{block} - \vartheta)}{\Big(\sum_{k=1}^{h^{-d}} \frac{1}{nh^d} \int_{I_k^d} (Y_k^* + R\sqrt{d}h^\beta - g(x))\omega(x)^2 dx \Big)^{1/2}} \Rightarrow \mathcal{N}(0,1).$$

For the details on the notations used, refer to the previous sections. The numerical results are obtained on the simulated data and we place $\omega = 1$. First, we construct the MLE estimators $\widehat{g}^{MLE}(x)$ based on the observations of the PPP on $[0,1] \times \mathbb{R}$ with intensity $\lambda_g(x,y) = n\mathbf{1}(g(x) \le y < c)$, $x \in [0,1]$ for two support functions

- $g_1(x) = \sqrt{x}$, $\beta = 0.5$, $R = 1$,
- $g_2(x) = 0.5\sin(2\pi x) + 4x$, $\beta = 1$, $R = 7.5$,

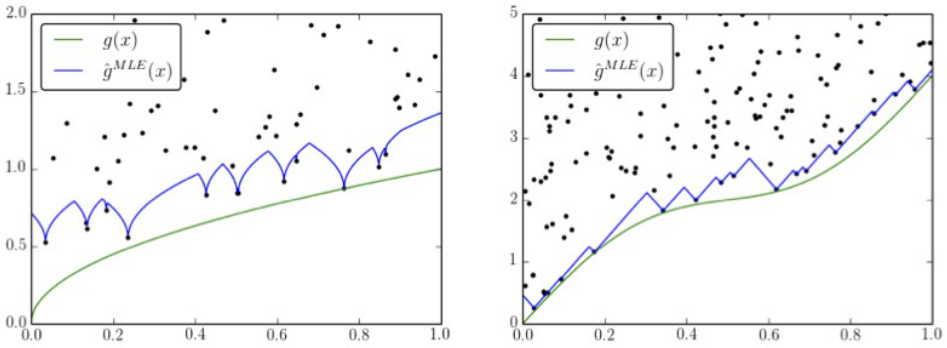


Figure 2.6: Nonparamatric MLE estimators for two functions: $g_1(x) = \sqrt{x}$ on the left and $g_2(x) = 0.5\sin(2\pi x) + 4x$ on the right.

see Figure 2.6. We take $c = 2$ for the function g_1 and $c = 5$ for the function g_2. For both functions, we consider $n = 50, 150, 500, 1000, 2000, 5000$ and construct a histogram $hist_n(x)$

and an empirical density function $\widehat{f}_n(x)$ for both $\chi_n^{(1)}$ and $\chi_n^{(2)}$ based on 100 iterations to illustrate the central limit theorem result. The numerical results for g_1 are depicted in Figure 2.7 and for g_2 – in Figure 2.8 for the statistics $\chi_n^{(1)}$. For the corresponding plots of $\chi_n^{(2)}$ refer to Figures 2.9 and 2.10. The simulations are made using various scipy.stats packages from [63] and packages from [57]: the histogram is made using matplotlib.pyplot.hist in [57] and the empirical density function is constructed applying a smoothing Gaussian kernel, the function scipy.stats.gaussian kde in [63], to the histogram.

Next, we study $\chi_n^{(1)}$ and $\chi_n^{(2)}$ for the PPP on $[0,1]^2 \times \mathbb{R}$ with intensity $\lambda_g(x_1, x_2, y) = n\mathbf{1}(g(x_1, x_2) \leq y < c)$, $(x_1, x_2) \in [0,1]^2$ and $c = 2$ for a support function

- $g(x_1, x_2) = \sqrt{x_1^2 + x_2^2}$, $\beta = 0.5$, $R = 1$.

Refer to Figure 2.13 for an illustration of $\widehat{g}^{MLE}$. Again as before, we consider $n = 50$, 150, 500, 1000, 2000, 5000 and construct histograms $hist_n(x)$ and the empirical density functions $\widehat{f}_n(x)$ for the quantities $\chi_n^{(1)}$ and $\chi_n^{(2)}$ based on 100 iterations to illustrate the central limit theorem result, see Figures 2.11 and 2.12.

Finally, as the last example, we examine $\chi_n^{(1)}$ with a support function

- $g(x_1, x_2, x_3, x_4, x_5) = 0.5\sin(2\pi(x_1 + x_2 + x_3 + x_4 + x_5)) + 4(x_1 + x_2 + x_3 + x_4 + x_5)$,

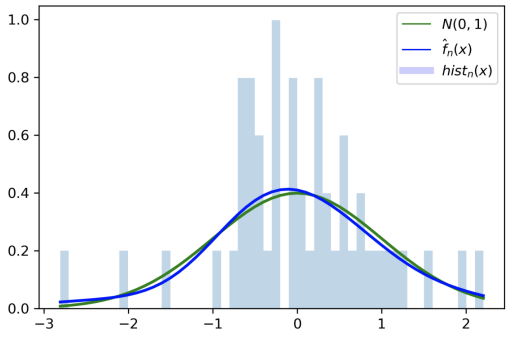
for the PPP on $[0,1]^5 \times \mathbb{R}$ of intensity $\lambda_g(x_1, x_2, x_3, x_4, x_5, y) = n\mathbf{1}(g(x_1, x_2, x_3, x_4, x_5) \leq y < c)$, $(x_1, x_2, x_3, x_4, x_5) \in [0,1]^5$ and $c = 5$. See Figure 2.14 for the results.

2.6 Further simulation studies

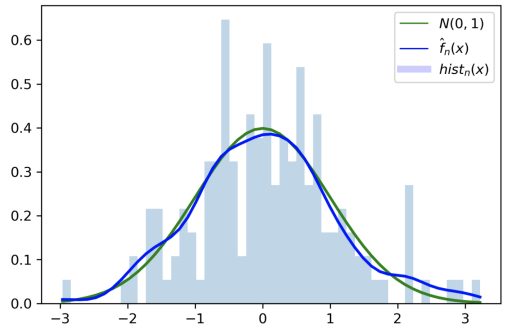
In this section we investigate the behaviour of the block-wise estimator and the MLE for a linear functional in both $\mathbb{R}$ and $\mathbb{R}^2$ spaces on finite samples. We simulate both the PPP model and the regression-type model. For the last model we consider the error terms being exponentially distributed with parameter 1. We consider two different regression functions g specified below

- $g(x) = \sqrt{x}$, $\beta = 0.5$, $R = 1$,
- $g(x_1, x_2) = \sqrt{x_1^2 + x_2^2}$, $\beta = 0.5$, $R = 1$,

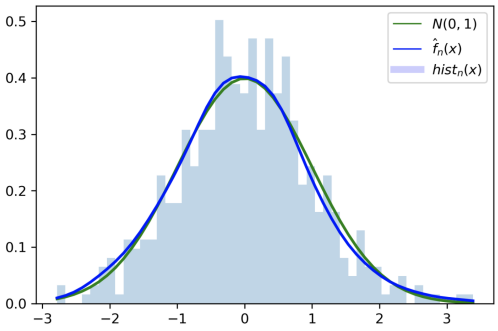
the intensity measure λ is defined in the same way as in the previous section, and estimate the root mean squared error (RMSE) in $M = 100$ Monte Carlo repetitions. What can be inferred from the plots is that both the block-wise estimator and the MLE work well enough even for small sized samples, i.e. starting from $n = 100$. The performance for the PPP model is slightly better than in the regression-type model.



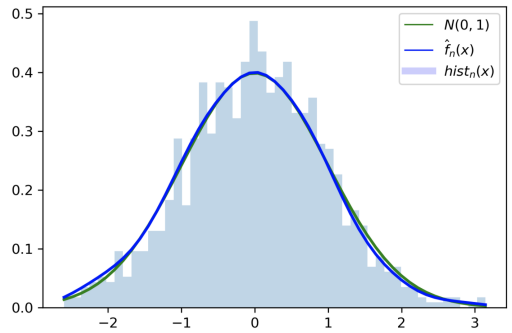
(a) $n = 50$



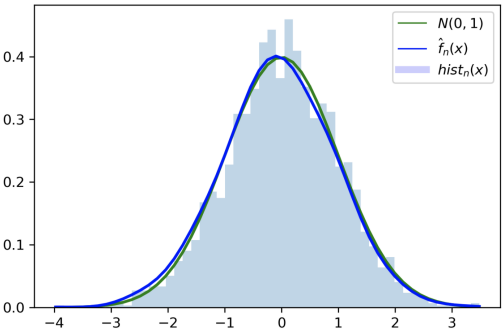
(b) $n = 150$



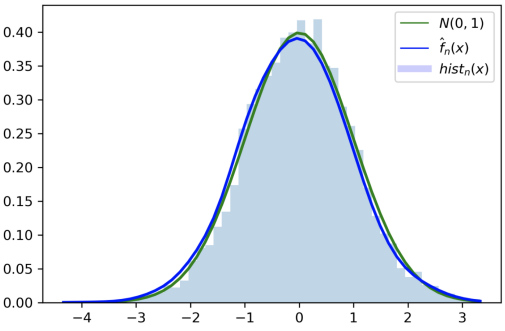
(c) $n = 500$



(d) $n = 1000$

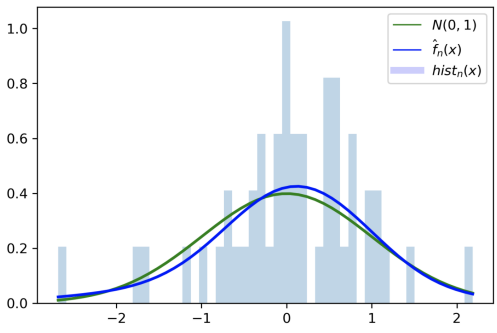


(e) $n = 2000$

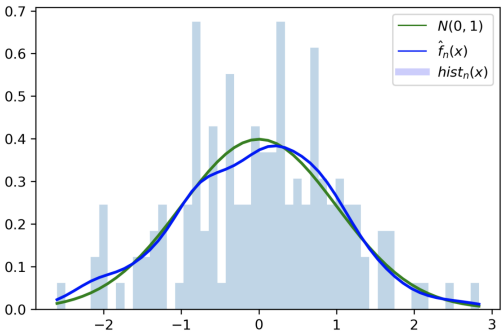


(f) $n = 5000$

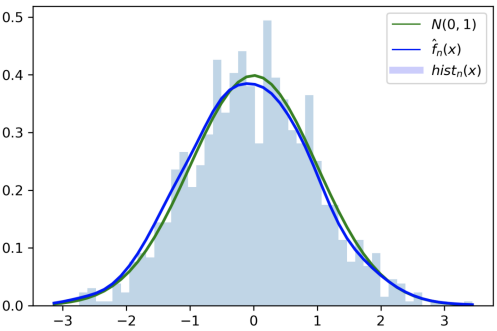
Figure 2.7: Histograms and empirical densities for the quantity $\chi_n^{(1)}$ for the function $g_1(x) = \sqrt{x}$.



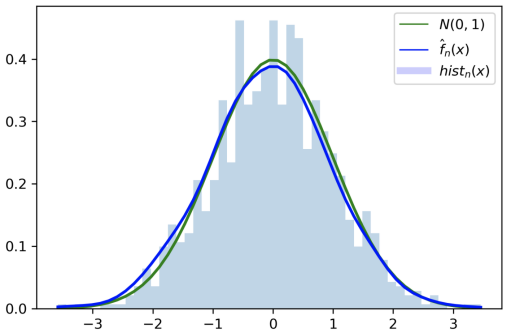
(a) $n = 50$



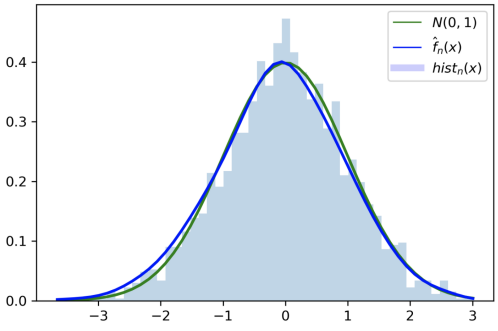
(b) $n = 150$



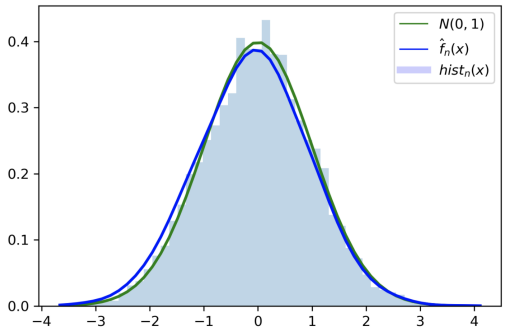
(c) $n = 500$



(d) $n = 1000$

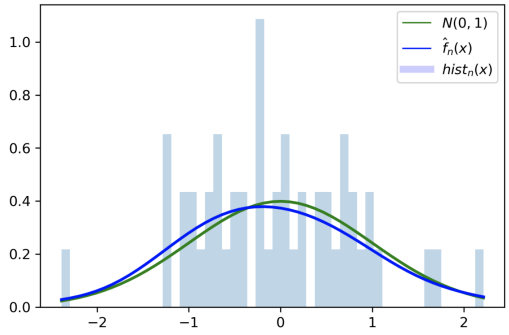


(e) $n = 2000$

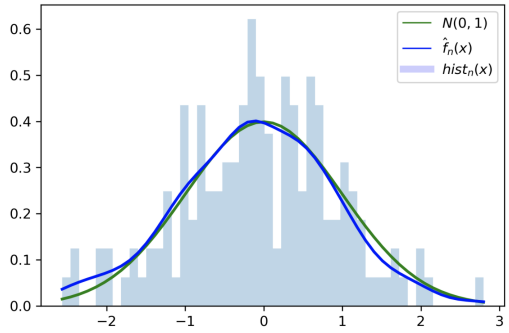


(f) $n = 5000$

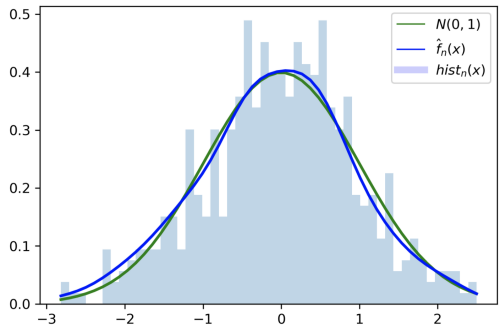
Figure 2.8: Histograms and empirical densities for the quantity $\chi_n^{(1)}$ for the function $g_2(x) = 0.5\sin(2\pi x) + 4x$.



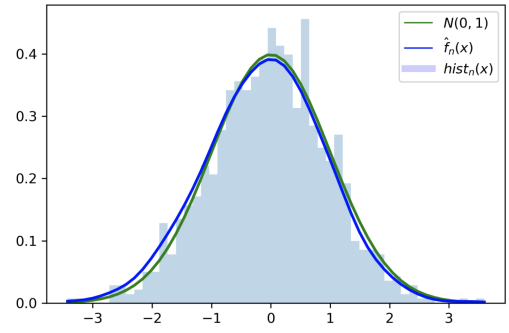
(a) $n = 50$



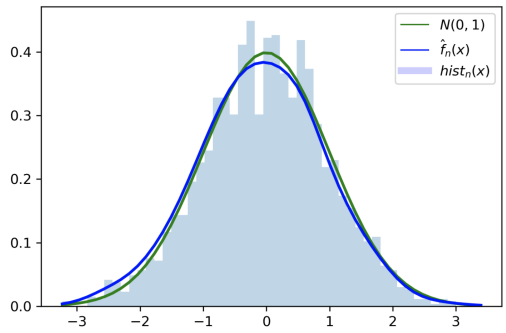
(b) $n = 150$



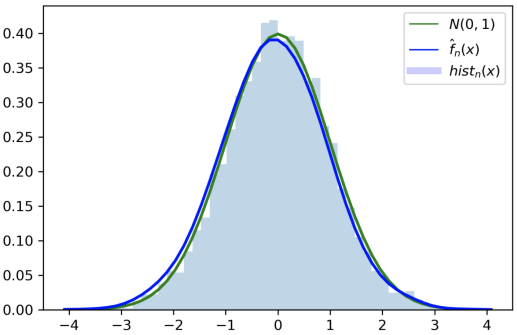
(c) $n = 500$



(d) $n = 1000$

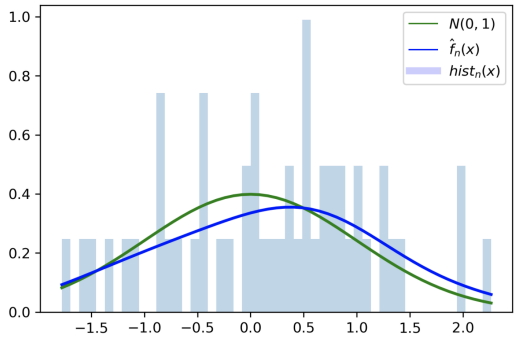


(e) $n = 2000$

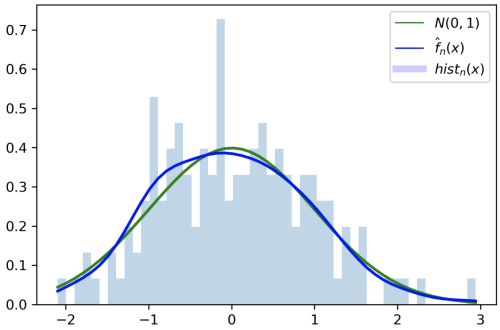


(f) $n = 5000$

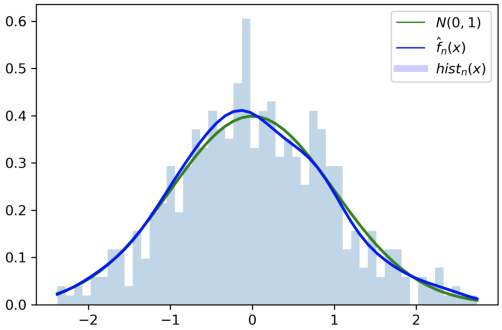
Figure 2.9: Histograms and empirical densities for the quantity $\chi_n^{(2)}$ for the function $g_1(x) = \sqrt{x}$.



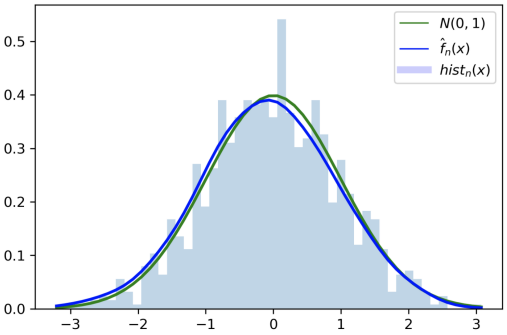
(a) $n = 50$



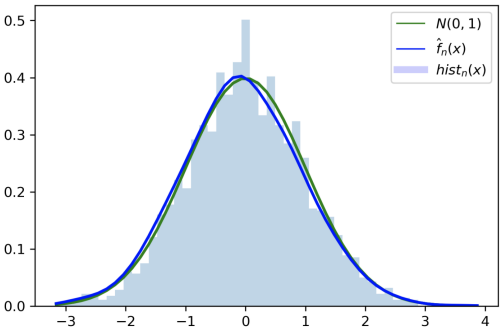
(b) $n = 150$



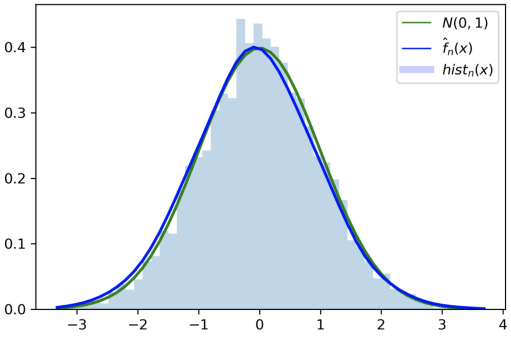
(c) $n = 500$



(d) $n = 1000$

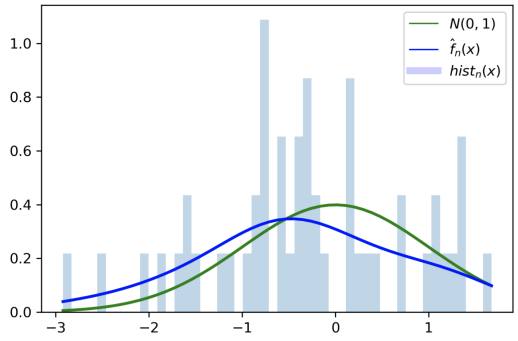


(e) $n = 2000$

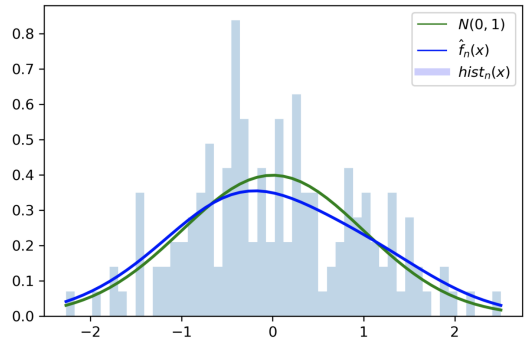


(f) $n = 5000$

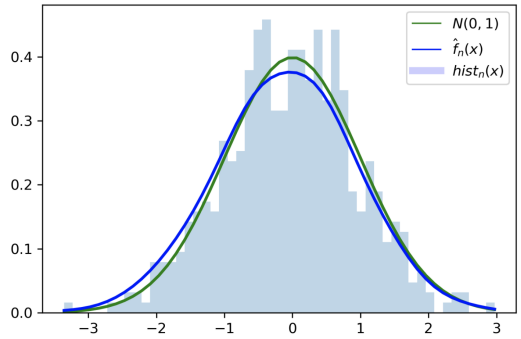
Figure 2.10: Histograms and empirical densities for the quantity $\chi_n^{(2)}$ for the function $g_2(x) = 0.5\sin(2\pi x) + 4x$.



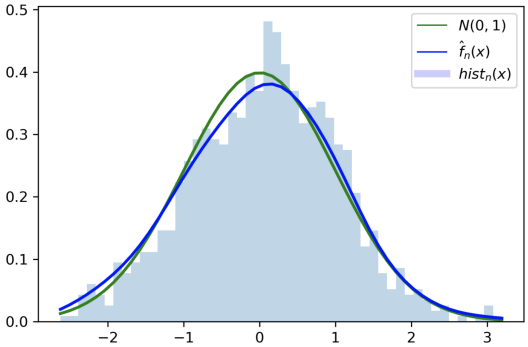
(a) $n = 50$



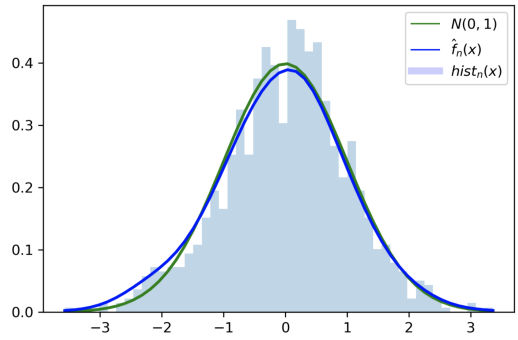
(b) $n = 150$



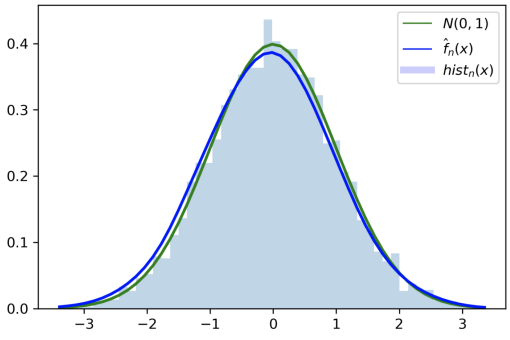
(c) $n = 500$



(d) $n = 1000$

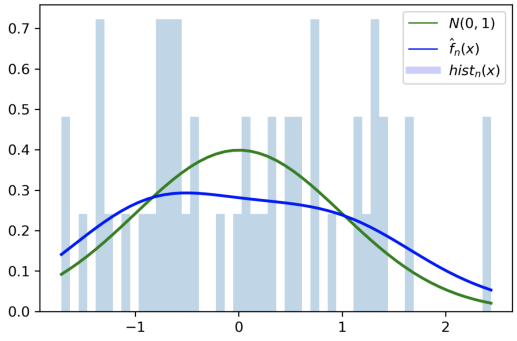


(e) $n = 2000$

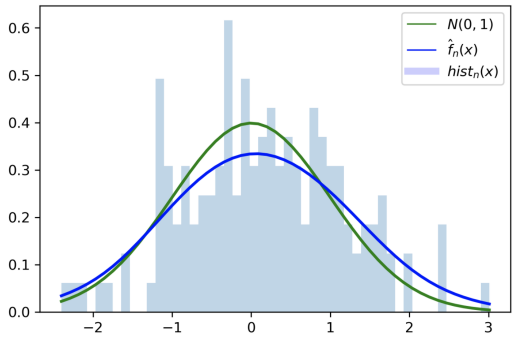


(f) $n = 5000$

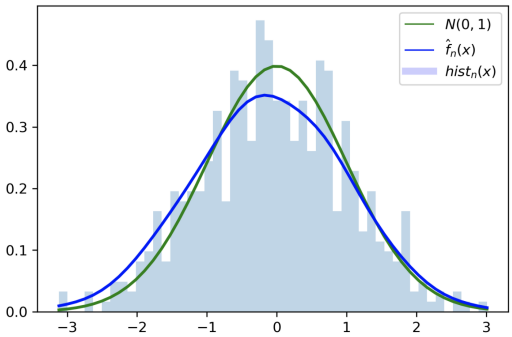
Figure 2.11: Histograms and empirical densities for the quantity $\chi_n^{(1)}$ for the function $g(x_1, x_2) = \sqrt{x_1^2 + x_2^2}$.



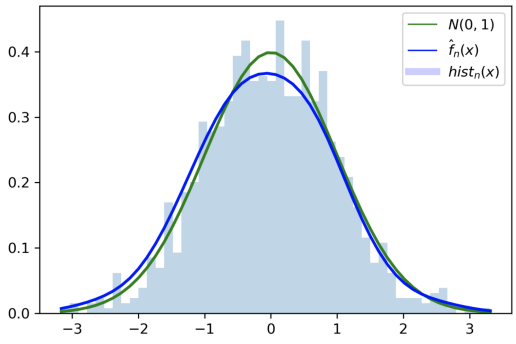
(a) $n = 50$



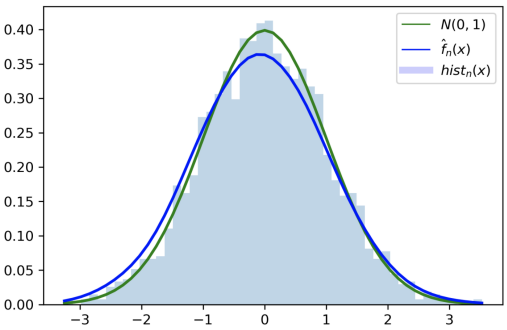
(b) $n = 150$



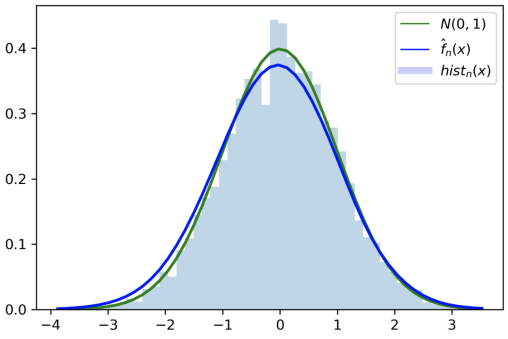
(c) $n = 500$



(d) $n = 1000$



(e) $n = 2000$



(f) $n = 5000$

Figure 2.12: Histograms and empirical densities for the quantity $\chi_n^{(2)}$ for the function $g(x_1, x_2) = \sqrt{x_1^2 + x_2^2}$.

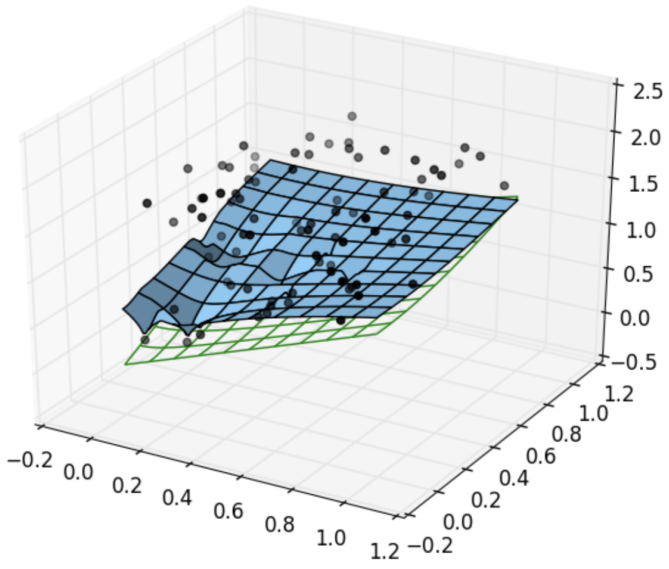


Figure 2.13: Nonparamatric MLE estimator (blue) for the function $g(x_1, x_2) = \sqrt{x_1^2 + x_2^2}$ (green).

On Figure 2.15 we see the results for one-dimensional case and on Figure 2.16 the results are for two-dimensional function g. Similar plots can be produced with respect to h for a range of n. Then, one can select h based upon where RMSE attains its minimum. However, the results are quite close to the theoretically optimal value of h.

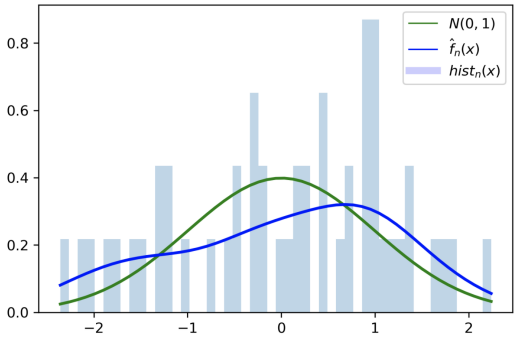
We can additionally observe from the plots with RMSE results that the block-wise estimator does not perform way worse than the MLE, what can also be expected from the theoretical results of the previous sections. If we consider the ratio of the non-asymptotic variances for $\hat{\vartheta}^{MLE}$ as well as $\hat{\vartheta}^{block}$ we obtain the following equation

$$\frac{\Gamma(\beta/(\beta+d))\beta(\beta/(\beta+d))^{d/(\beta+d)}}{\beta^{-\beta/(\beta+d)}(\beta+d)} = \Gamma(\beta/(\beta+d))\beta^{(2\beta+d)/(\beta+d)}(\beta+d)^{-(\beta+2d)/(\beta+d)}. \tag{2.6.1}$$

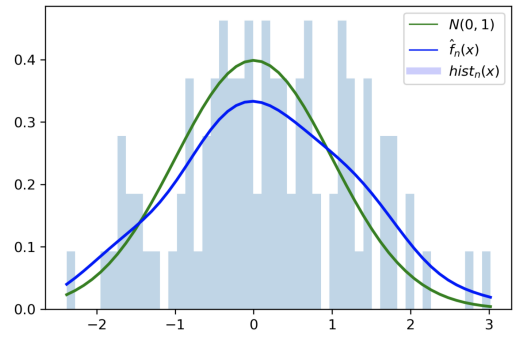
If we consider the case of dimension one, the above ratio will be equal to the following function of β

$$\Gamma(\beta/(\beta+1))\beta^{(2\beta+1)/(\beta+1)}(\beta+1)^{-(\beta+2)/(\beta+1)}.$$

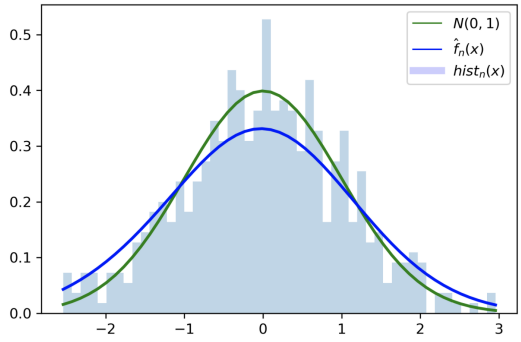
This function attains its minimum when $\beta \approx 0.473$ which is 0.54, after that it increases to approximately 0.6 as β approaches 1, as β goes to zero the function approaches 1. Refer to Figure 2.17 for the illustration of the described dynamics in dimension one. If we consider the behaviour of the ratio from (2.6.1) in the case of higher dimension, we get similar results. As an example, see Figure 2.18 for the illustration of the case of dimensionality 10.



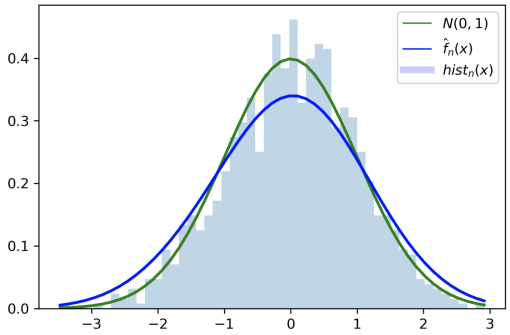
(a) $n = 50$



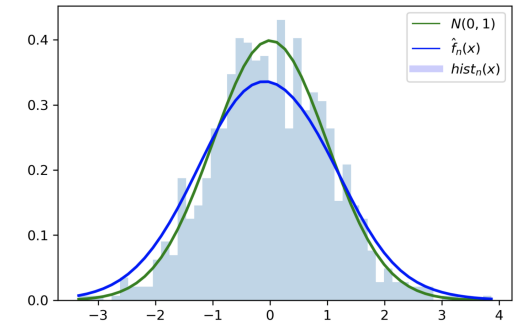
(b) $n = 150$



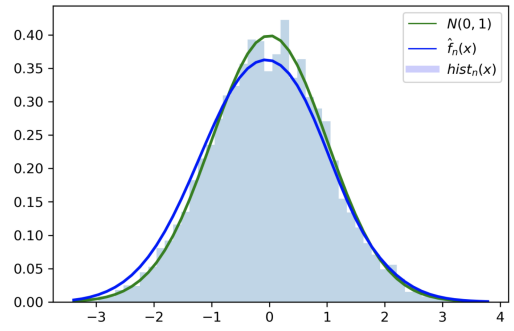
(c) $n = 500$



(d) $n = 1000$



(e) $n = 2000$



(f) $n = 5000$

Figure 2.14: Histograms and empirical densities for the quantity $\chi_n^{(1)}$ for the function $g(x_1, x_2, x_3, x_4, x_5) = 0.5 \sin(2\pi(x_1 + x_2 + x_3 + x_4 + x_5)) + 4(x_1 + x_2 + x_3 + x_4 + x_5)$.

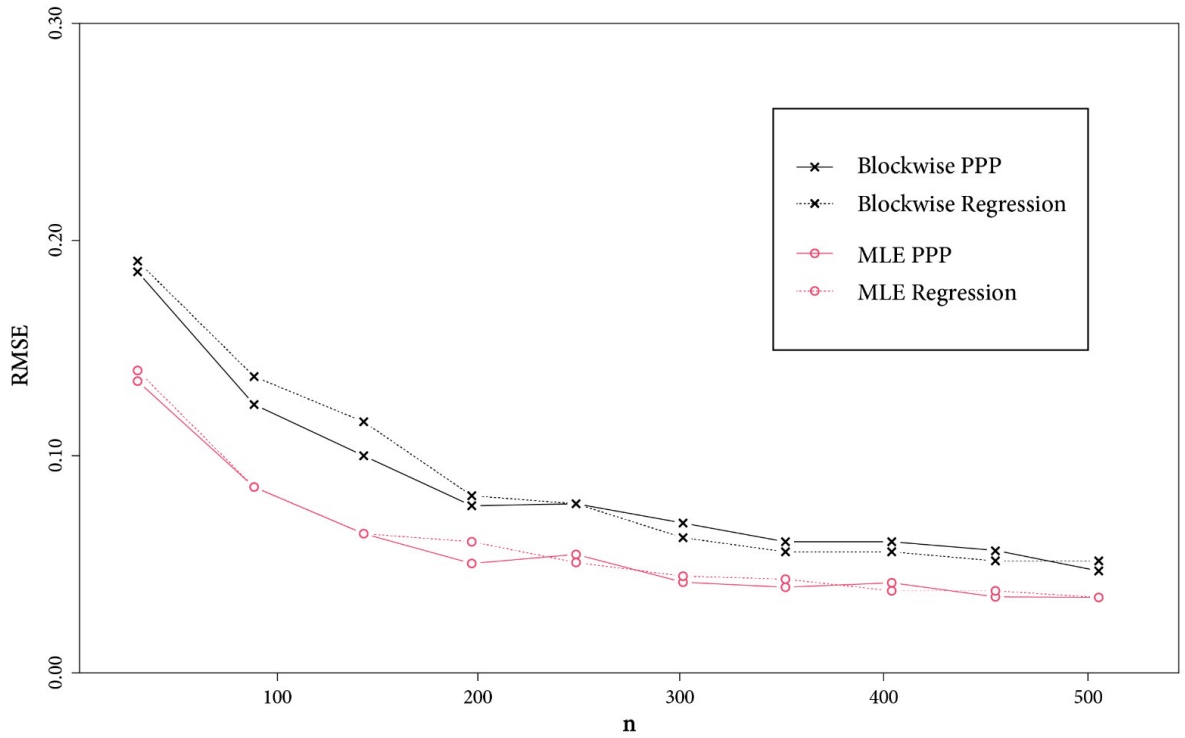


Figure 2.15: Monte-Carlo based RMSE for the function $g(x) = \sqrt{x}$.

Comparing the MLE and simple block-wise estimator the latter is computationally faster, which is highly applicable for any adaptive estimator. On top of that it is theoretically simpler to analyse than the MLE. These reasons make us conclude that both approaches can be appealing and as such should be considered in their own accord.

2.7 Appendix

2.7.1 Technical results

The main results of the previous chapters heavily rely upon the martingale approach, with martingale stopping arguments for counting processes playing a key role. One way to construct a martingale from a stochastic process is via the Doob decomposition theorem, which allows us to derive a martingale by subtracting from a point process its compensator. To get an idea how the compensator of a point process looks like we consider a simple example of a point process of a single jump. Let T be a random variable and the process N have a single jump at the random time T, that is, $N(t) = \mathbf{1}(T \leq t)$ and let F be the distribution function of T. Then, the compensator $A(t)$ of $N(t) = \mathbf{1}(T \leq t)$ is given by

$$A(t) = \int_0^{t \wedge T} \frac{dF(s)}{1 - F(s-)}.$$

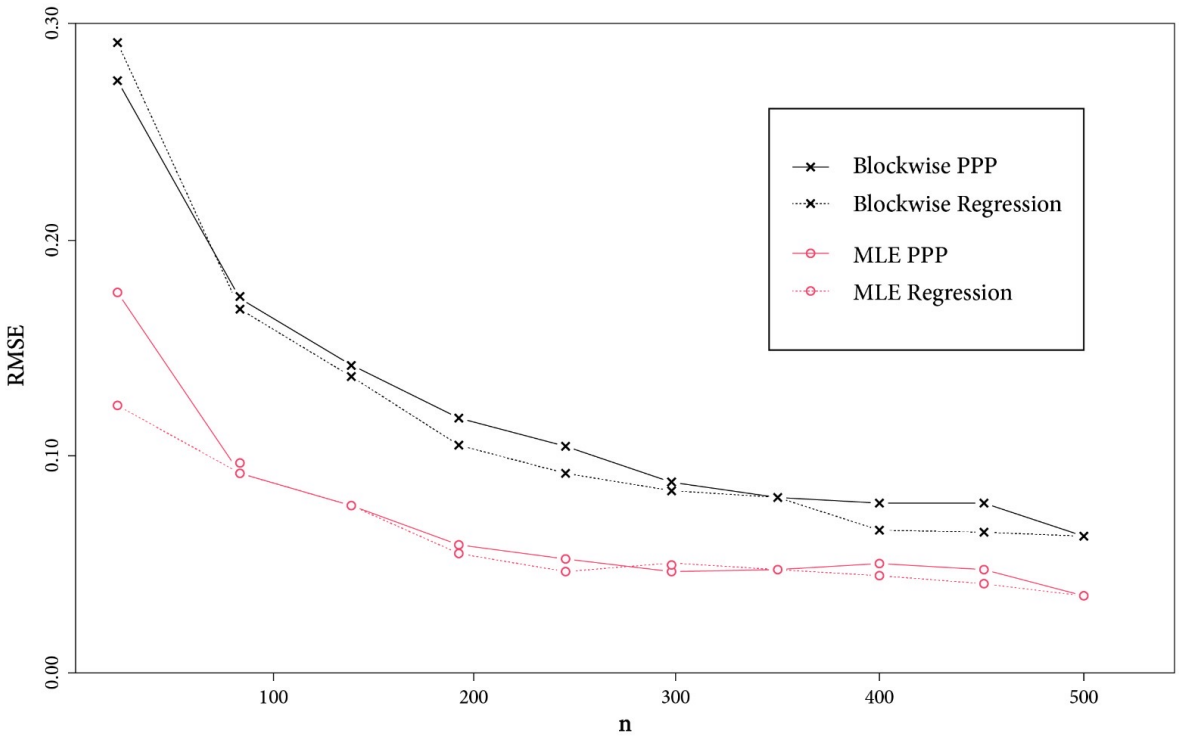


Figure 2.16: Monte-Carlo based RMSE for the function for the function $g(x_1, x_2) = \sqrt{x_1^2 + x_2^2}$.

$A(t)$ is clearly predictable, and to show that $N(t) - A(t)$ is a martingale it suffices to show that $\mathbb{E}[N(\tau)] = \mathbb{E}[A(\tau)]$ for any stopping time τ. One can show, see [66], that there exists an $\mathcal{F}_0$-measurable (i.e. almost surely constant) random variable ξ such that

$$\{\tau \geq T\} = \{\tau \wedge T = T\} = \{\xi \wedge T = T\} = \{T \leq \xi\}.$$

Therefore,

$$\begin{aligned}
\mathbb{E}[N(\tau)] &= \mathbb{P}(\tau \geq T) = \mathbb{P}(T \leq \xi) \\
&= \int_0^{\xi} dF(t) = \int_0^{\xi \wedge T} \frac{\mathbb{P}(T \geq t)}{1 - F(t-)} dF(t) \\
&= \mathbb{E}\Big(\int_0^{\xi} \frac{\mathbf{1}(T \geq t)}{1 - F(t-)} dF(t) \Big) = \mathbb{E}\Big(\int_0^{\xi \wedge T} \frac{1}{1 - F(t-)} dF(t) \Big) \\
&= \mathbb{E}\Big(\int_0^{\tau \wedge T} \frac{1}{1 - F(t-)} dF(t) \Big) = \mathbb{E}[A(\tau)].
\end{aligned}$$

Similarly, one can derive a compensator for a general counting process. Note that if the distribution F is continuous with $F(0) = 0$ then, by changing variables we have

$$\int_0^T \frac{dF(t)}{1 - F(t-)} = \int_0^T \frac{dF(t)}{1 - F(t)} = -\log(1 - F(T)).$$

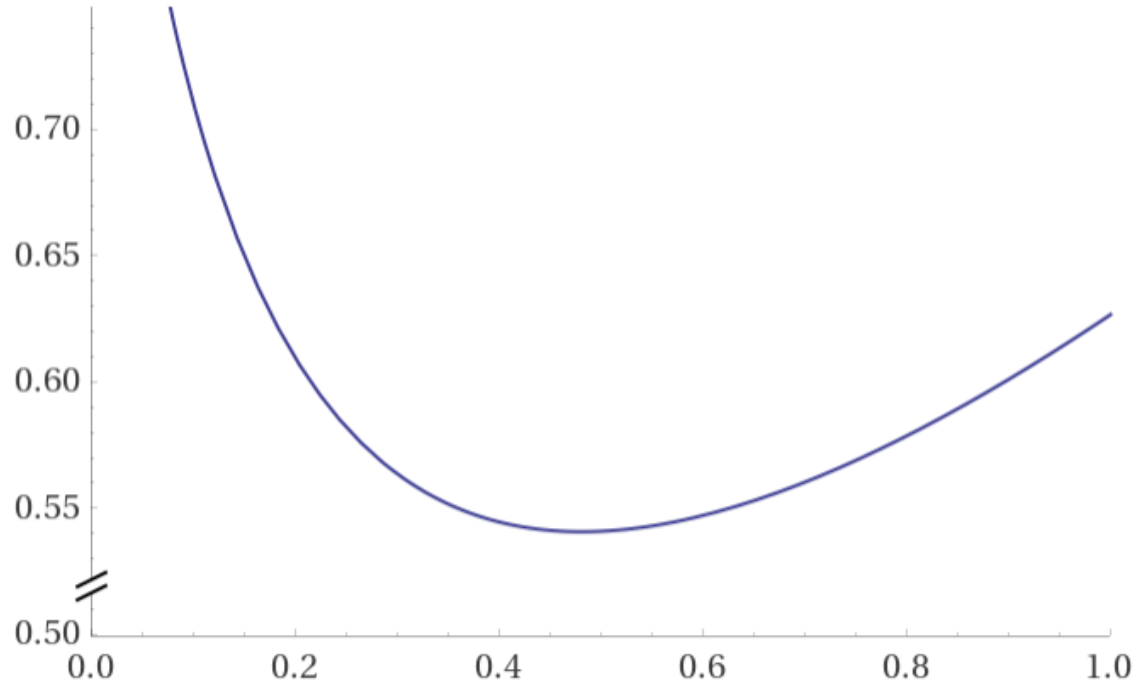


Figure 2.17: The ratio of the non-asymptotic variances of $\hat{\vartheta}^{MLE}$ and $\hat{\vartheta}^{block}$. The vertical axis represents to the values of the function $\Gamma(\beta/(\beta+1))\beta^{(2\beta+1)/(\beta+1)}(\beta+1)^{-(\beta+2)/(\beta+1)}$ while x-axis corresponds the values of β.

Next, we formulate a stopping-time theorem for continuous-time martingales. This result we provide with the proof as well as it is quite concise.

Theorem 2.7.1. *Let $(M(t), t \geq t_0)$ be a cadlag martingale with $M(t_0) = 0$ and let τ be a stopping time with values in $[t_0, \infty]$, both on some filtered probability space. If $\mathbb{E}[\langle M \rangle_\tau]$ is finite, then $\mathbb{E}[M(\tau)] = 0$ and $\mathbb{E}[M(\tau)^2] = \mathbb{E}[\langle M \rangle_\tau]$ hold.*

Proof. From the Burkholder-Davis-Gundy inequality (Theorem 26.12 in [64]) and the identity $\mathbb{E}[[M]_\tau] = \mathbb{E}[\langle M \rangle_\tau]$ we can infer that

$$\mathbb{E}[\sup_{t \geq t_0} M^2_{t \wedge \tau}] \lesssim \mathbb{E}[\langle M \rangle_\tau].$$

Hence, $(|M_{t\wedge\tau}|^p)_{t \leq t_0}$ for $p \in 1, 2$ is uniformly integrable. By optional stopping it follows that $\mathbb{E}[[M]_\tau] = \lim_{t\to\infty} \mathbb{E}[M_{t\wedge\tau}] = 0$ as well as $[M^2_\tau] = \mathbb{E}[[M]_\tau] = \mathbb{E}[\langle M \rangle_\tau]$. □

Lemma 2.7.2. *Under the assumptions of Theorem 2.4.3 we have for $\tau = Y^*_k + R\|h\|^\beta$*

$$\mathbb{E}[(\tau - g(\mathbf{X}_i))^p]^{1/p} \lesssim R\|h\|^\beta + (n\lambda h^d)^{-1}$$

as $nh^d \to \infty$ for any $p > 0$.

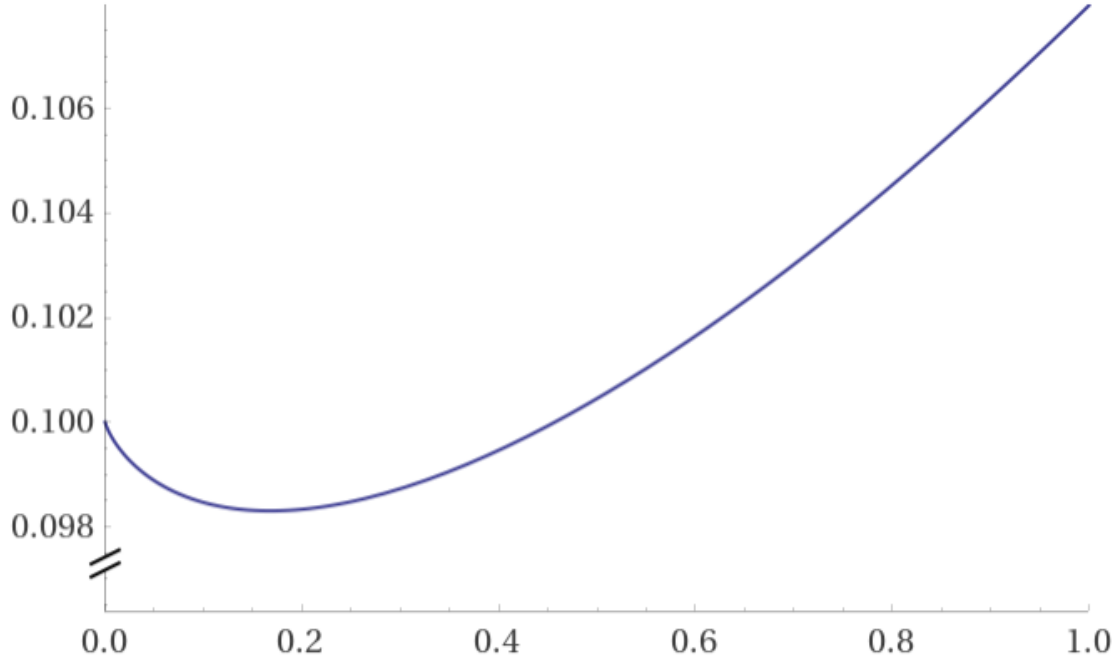


Figure 2.18: The ratio of the non-asymptotic variances of $\hat{\vartheta}^{MLE}$ and $\hat{\vartheta}^{block}$. The vertical axis represents the values of the function $\Gamma(\beta/(\beta+10))\beta^{(2\beta+10)/(\beta+10)}(\beta+10)^{-(\beta+20)/(\beta+10)}$ while x-axis corresponds to different values of β.

Proof. The fact that $Y_k^* \leq \max_{i\in\tilde{I}_k^d} g(\mathbf{X}_i) + \min_{i\in\tilde{I}_k^d} \varepsilon_i$ provides that as $nh^d \to \infty$

$$P(n\lambda h^d(Y_k^* - \max_{i\in\tilde{I}_k^d} g(\mathbf{X}_i)) \geq z) \leq \bar{F}_\varepsilon(z/n\lambda h^d)^{nh^d} = e^{nh^d \log \bar{F}_\varepsilon(z/n\lambda h^d)} \to e^{-z}.$$

Since $\bar{F}_\varepsilon(z/n\lambda h^d)^{nh^d} \lesssim (1+z/nh^d)^{-nh^d\rho}$, for any $p \geq 0$ we can establish that the following limit is zero

$$\lim_{R\to\infty} \sup_{n,h} \int_R^\infty z^{p-1} P(n\lambda h^d(Y_k^* - \max_{i\in\tilde{I}_k^d} g(\mathbf{X}_i)) \geq z)dz = 0$$

such that by uniform integrability the result of

$$\limsup_{nh^d\to\infty} \mathbb{E}\Big[\Big(n\lambda h^d|Y_k^* - \max_{i\in\tilde{I}_k^d} g(\mathbf{X}_i)|\Big)^p\Big] \leq \int_0^\infty z^p e^{-z} dz < \infty$$

follows.

Next, from the Hölder condition we infer that g varies at most by $R(\sqrt{d}h)^\beta$ on each cube and thus, the result $\mathbb{E}[(\tau - g(\mathbf{X}_i))^p]^{1/p} \lesssim R(\sqrt{d}h)^\beta + (n\lambda h^d)^{-1}$ holds. □

Now we ready to address the proof of Theorem 2.4.7.

Proof. For $h_m < h^*$ we also have that $h_m^d < (h^*)^d$ and thus, we obtain from the deviation

bound in Proposition 2.7.3, stated below, that

$$\begin{aligned}\mathbb{P}(\widehat{h}^d = h_m^d) = \mathbb{P}(\widehat{h} = h_m) &\leq \sum_{m'}^{m-1} \Big(\mathbb{P}(|\tilde{\vartheta}^{block}_{n,h'_m} - \vartheta^{(n)}| > \kappa'_m) + \mathbb{P}(|\tilde{\vartheta}^{block}_{n,h_{m+1}} - \vartheta^{(n)}| > \kappa_{m+1}) \Big) \\ &\leq \sum_{m'}^{m-1} \Big(4n^{-2c} + n\bar{F}_\varepsilon((c\log n - 2)/nh^d_{m'})^{nh^d_{m'}} + n\bar{F}_\varepsilon((c\log n - 2)/nh^d_{m+1})^{nh^d_{m+1}} \Big) \\ &\lesssim M(n^{-2c} + n^{1-\underline{\lambda}c})\end{aligned}$$

for some $\underline{\lambda} \in (0, \lambda)$ and n is being sufficiently large. Applying Cauchy-Schwarz inequality we get the following upper bound

$$\mathbb{E}[(\tilde{\vartheta}^{block}_n - \tilde{\vartheta}^{(n)})^2 \mathbf{1}(\widehat{h} < h^*)] \lesssim \mathbb{E}[(\tilde{\vartheta}^{block}_n - \tilde{\vartheta}^{(n)})^4 \mathbf{1}(\widehat{h} < h^*)]^{1/2} M(n^{-c} + n^{(1-\underline{\lambda}c)/2})^{1/2}.$$

Combining the result of Theorem 2.7.1 together with Lemma 2.7.2 we infer that the fourth moment of the error is bounded so that the first inequality follows. Moreover, it even tends to zero.

From the construction it follows that

$$|\tilde{\vartheta}^{block}_n - \tilde{\vartheta}^{block}_{n,h^*}| < \kappa_{\hat{m}} + \kappa_{m^*}$$

for $h_{\hat{m}} := \hat{h} > h^* =: h_{m^*}$ holds. Moreover, since $H_x(y) \approx xy^2$ and $h^d_m \geq (\log n)^2 n^{-1}$, we can conlclude that the sum of $\kappa_{\hat{m}}$ and κ_{m^*} is bounded by

$$\kappa_{\hat{m}} + \kappa_{m^*} \lesssim \frac{(\log n)^2}{(n(h^*)^d)^2} + \frac{\sqrt{(\log n)^2}}{(n(h^*)^{d/2})} \Big(1 + \max_{h_m \geq h^*} h^d_m \sum_{k=1}^{h_m^{-d}} \#\{i \in I_{k,h_m} : \mathscr{Y}_i \leq (\mathscr{Y}^*_{k,h_m} + (nh^d_m)^{-1}\} \Big).$$

Now for every fixed h_m by compensation of the block-wise counting process we conclude that

$$\mathbb{E}\Big[\Big(\sum_{k=1}^{h_m^{-d}} \#\{i \in I_{k,h_m} : \mathscr{Y}_i \leq (\mathscr{Y}^*_{k,h_m} + (nh^d_m)^{-1}\} \Big)^2 \Big] \leq h_m^{-d} sum_{k=1}^{h_m^{-d}} \mathbb{E}[A_k^2 + A_k]$$

where A_k is defined as follows

$$A_k = \sum_{I_{k,h_m}} \int \mathbf{1}\Big(\mathscr{Y}^*_{k,h_m} < s + g(\mathbf{X}_i) \leq \mathscr{Y}^*_{k,h_m} + (nh^d_m)^{-1} \Big) \frac{f_{\varepsilon(s)}}{\bar{F}_\varepsilon(s)} ds \leq \|f_\varepsilon / \bar{F}_\varepsilon\|_\infty \sim 1.$$

Using the fact that $\#\{i \in I_{k,h_m} : \mathscr{Y}_i \leq (\mathscr{Y}^*_{k,h_m} + (nh^d_m)^{-1}\} = 1$ a.s. by the definition, we obtain

$$\mathbb{E}\Big[h^d_m \sum_{k=1}^{h_m^{-d}} \#\{i \in I_{k,h_m} : \mathscr{Y}_i \leq (\mathscr{Y}^*_{k,h_m} + (nh^d_m)^{-1}\} \Big)^2 \Big] \lesssim 1.$$

Hence, we obtain the second inequality

$$\mathbb{E}[(\tilde{\vartheta}_n^{block} - \tilde{\vartheta}_{n,h^*}^{block})^2 \mathbf{1}(\widehat{h}' \geq h^*)] \lesssim \frac{(\log n)^4}{(n(h^*)^d)^4} + \frac{M \log n}{n^2 (h^*)^d}.$$

To show the result for the asymptotic rate we just note that the geometric grid of bandwidths needs to achieve the following rate $h^* \sim n^{-1/\beta+d}$ asymptotically, then by inserting we have the following bound

$$|\mathbb{E}[\tilde{\vartheta}_n^{block} - \vartheta^{(n)}]| \lesssim n^{-(2\beta+d)/(\beta+d)} + n^{-4\beta/(\beta+d)}$$

and the result follows from the triangle inequality. Also note that the risk is negligible on $\hat{h} < h^*$. □

The proof of the result of Theorem 2.4.7 is based upon the following proposition that provides the result of deviation inequality.

Proposition 2.7.3. *For any $h, x, \kappa > 0$ that satisfy the following conditions*

$$\begin{aligned} R(\sqrt{d}h)^\beta &\leq (nh^d)^{-1}, \\ 2xh^{d/2}\|\omega\|_\infty &< 1, \\ \kappa &< (\delta - 2R(\sqrt{d}h)^\beta) nh^d \end{aligned}$$

the following bound holds

$$n\lambda h^{d/2}\tilde{\vartheta}_n^{block} - \vartheta^{(n)} \leq \sum_i \mathbf{1}(\mathscr{Y}_i \leq \tau^{(i)}) H_x(h^{d/2}\omega(\mathbf{X}_i)) + C^2(\kappa+2)^2 n^{-1} h^{3d/2}\|\omega\|_1 + x,$$

with probability at least $1 - 2e^{2x^2} - h^{-d}\bar{F}_\varepsilon(z/n\lambda h^d)^{nh^d}$.

Proof. We start with the martingale $M(t)$ from (2.4.14) that we constructed in the proof of Theorem 2.4.3 and consider as well the associated stopping time τ which was introduced there too. By the substitution rule, refer to (Theorem 26.7 in [64]), we obtain the following exponential martingale

$$\begin{aligned} \mathcal{E} &= \exp\Big(\sum_i \mathbf{1}(\mathscr{Y}_i \leq t) \log(1 + \gamma\omega(\mathbf{X}_i)) + \log(\bar{F}_\varepsilon(\mathscr{Y}_i \wedge t - g(\mathbf{X}_i)))\gamma\omega(\mathbf{X}_i) \Big) \\ &= \exp\Big(\gamma M(t) - \sum_i \mathbf{1}(\mathscr{Y}_i \leq t) \Big(\gamma\omega(\mathbf{X}_i) - \log(1 + \gamma\omega(\mathbf{X}_i)) \Big) \Big), \end{aligned}$$

for $\gamma > -1/\|\omega\|_\infty$.

From the fact that for all $p > 1$ the following random variable $\bar{F}_\varepsilon(\varepsilon_i)$, that is uniformly distributed on $[0, 1]$, has finite p-moments and by Theorem 2.7.1 we obtain the following

stopping result

$$\mathbb{E}\Big[\exp\Big(\gamma M(\tau)-\sum_i \mathbf{1}(\mathscr{Y}_i\leq\tau)\Big(\gamma\omega(\mathbf{X}_i)-\log(1+\gamma\omega(\mathbf{X}_i))\Big)\Big)\Big]=1.$$

The independence between cubes provides $\mathbb{E}[e^{\gamma Z_\gamma}]=1$, where Z_γ is defined in the following way

$$\begin{aligned}Z_\gamma &:= n\lambda(\tilde{\vartheta}^{block}_{n,h}-\vartheta^{(n)})-\sum_i G_\varepsilon(\mathscr{Y}_i\wedge\tau^i-g(\mathbf{X}_i))\omega(\mathbf{X}_i)\\ &-\sum_i \mathbf{1}(\mathscr{Y}_i\leq\tau^{(i)})\Big(\omega(\mathbf{X}_i)-\frac{\log(1+\gamma\omega(\mathbf{X}_i))}{\gamma}\Big).\end{aligned}$$

We now pick γ as $|\gamma|=2xh^{d/2}$ and infer by using the result of Markov inequality the following two upper bounds

$$\begin{aligned}\mathbb{P}(h^{d/2}Z_{2xh^{d/2}}\geq x)&\leq e^{-2x^2},\\ \mathbb{P}(h^{d/2}Z_{2xh^{d/2}}\leq -x)&\leq e^{2x^2}\end{aligned}$$

such that

$$h^{d/2}|n\lambda(\tilde{\vartheta}^{block}_{n,h}-\vartheta^{(n)})-\Xi|\leq\sum_i \mathbf{1}(\mathscr{Y}_i\leq\tau^{(i)})H_x(h^{d/2}\omega(\mathbf{X}_i))+x, \tag{2.7.1}$$

with probability $1-2e^{-2x^2}$. We use the following notation in (2.7.1)

$$\Xi=\sum_i G_\varepsilon(\mathscr{Y}_i\wedge\tau^i-g(\mathbf{X}_i))\omega(\mathbf{X}_i).$$

From the fact that for $z\in[0,\delta]$ the following bound holds $|G_\varepsilon(z)|\leq C^2z^2$ and also $R(\sqrt{d}h)^\beta\leq(nh^d)^{-1}$ we obtain

$$\begin{aligned}\mathbb{P}\Big(\max_{i\in I_k^d}|G_\varepsilon(\mathscr{Y}_i\wedge\tau^i-g(\mathbf{X}_i))|>\frac{C^2(\kappa+2)^2}{(nh^d)^2}\Big)&\leq\mathbb{P}\Big(\max_{i\in I_k}(\mathscr{Y}_i\wedge\tau^i-g(\mathbf{X}_i))>\frac{(\kappa+2)}{nh^d}\Big)\\ &\leq\mathbb{P}\Big(\min_{i\in I_k}\varepsilon_i+2(nh^d)^{-1})>\frac{(\kappa+2)}{nh^d}\Big)\\ &=\bar{F}_\varepsilon\Big(\kappa/nh^d\Big)^{nh^d}.\end{aligned}$$

We thus infer the following upper bound

$$\Big|\sum_{i\in I_k}G_\varepsilon(\mathscr{Y}_i\wedge\tau^i)-g(\mathbf{X}_i))\omega(\mathbf{X}_i)\Big|\leq\frac{C^2(\kappa+2)^2}{(nh^d)^2}\sum_{i\in I_k}|\omega(\mathbf{X}_i)|$$

with probability $1 - \bar{F}_\varepsilon(\kappa/nh^d)^{nh^d}$.

Summing over the h^{-d} cubes in the result above we obtain that

$$h^{d/2}\Big|\sum_{i\in I_k} G_\varepsilon(\mathscr{Y}_i \wedge \tau^i) - g(\mathbf{X}_i))\omega(\mathbf{X}_i)\Big| \leq \frac{C^2(\kappa+2)^2}{n} h^{-3d/2} \parallel \omega \parallel_1,$$

holds with probability $1 - h^{-d}\bar{F}_\varepsilon(\kappa/nh^d)^{nh^d}$. From (2.7.1) then the result follows. □

2.7.2 Basic statistical framework and minimax risk

We here collect foundational statistical tools which are used throughout, and we prefer to follow the universal notation. We assume that the data $\mathbf{Y}$, are given by a random vector on $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values in a measurable space $(S, \mathcal{S})$ and coming from some true measure $\mathbb{P}_*$. This measure belongs to a family of probability measures $\mathcal{P} = \{\mathbb{P}_g, g \in G\}$, defined on the space $(S, \mathcal{S})$, where G is a parameter space. We denote ϑ as a functional on the space $\mathcal{P}$, taking values in a measurable space (Θ, Ξ) – that is, a mapping $\mathbb{P} \to \vartheta(\mathbb{P}_g) \in \Theta$ (usually $\Theta = \mathbb{R}^d$ equipped with its Borel sigma-algebra $\mathcal{B}_{\mathbb{R}^d}$ for some $d \geq 1$).

Our goal is to estimate $\vartheta(\mathbb{P}_g)$ based on the data $\mathbf{Y}$ we observe. We call a statistic any measurable function u of the data $\mathbf{Y}$ and usually denote it as T or u. In order to estimate the functional $\vartheta(\mathbb{P}_g)$, we construct a statistic $\widehat{\vartheta} : (S, \mathcal{S}) \to (\Theta, \Xi)$, which we call an estimator. One of the most desired properties of an estimator is to be unbiased and at the same time of minimal variance. This property is referred to as UMVU (uniformly minimum variance unbiased). We formulate a precise definition of this property.

Definition 2.7.4. For a sample $\mathbf{Y}$ that is taken from an unknown measure $(\mathbb{P}_g \in \mathcal{P})$ an unbiased estimator $\widehat{\vartheta}(\mathbf{Y})$ of ϑ is called UMVU estimator if and only if $\mathrm{Var}(\widehat{\vartheta}(\mathbf{Y})) \leq \mathrm{Var}(u(\mathbf{Y}))$ for any $(\mathbb{P}_g \in \mathcal{P})$ and any other unbiased estimator $u(\mathbf{Y})$ of ϑ.

We now introduce two essential classes of statistics which allow us to conduct further statistical inference.

Definition 2.7.5. A measurable function $u(\mathbf{Y})$ of the data is called a sufficient statistic for $\mathbf{Y}$, if conditionally on $u(\mathbf{Y})$ the distribution of the data $\mathbf{Y}$ does not depend on an unknown function of interest ϑ, i.e. $\mathbb{P}(\mathbf{Y}|u(\mathbf{Y}) = v)$ is independent of ϑ, provided that a regular conditional probability exists.

Broadly speaking, a statistic is sufficient with respect to a statistical model and its associated unknown function ϑ if no other statistic that can be calculated from the same data provides any additional information as to the value of the function ϑ.

Another important class of statistics we use in this chapter is closely related to the identifiability of a model, and is called complete statistic.

Definition 2.7.6. A measurable function $u(\mathbf{Y})$ of the data $\mathbf{Y}$ is called a complete statistic for the distribution of $\mathbf{Y}$, if for every measurable function T the following implication holds

$$\forall g: \quad \mathbb{E}_g[T(u(\mathbf{Y}))] = 0 \quad \text{implies that } \; T(u(\mathbf{Y})) = 0 \quad \mathbb{P}_g - a.s. \tag{2.7.2}$$

One of the reasons why these two classes of statistics are of interest is that given an unbiased estimator of the target parameter ϑ and a complete sufficient statistic u one can derive the unique unbiased estimator of uniformly minimal variance among all unbiased estimators. This striking result can be inferred from the two following prominent theorems.

Theorem 2.7.7. *(Rao-Blackwell) Let* $\mathbf{Y}$ *be a sample from an unknown measure* $\mathbb{P}_g \in \mathcal{P}$, $\widehat{\vartheta}$ *be an estimator of* ϑ *with* $\mathbb{E}[\widehat{\vartheta}^2] < \infty$ *and* u *be a sufficient statistic for* $\mathcal{P}$*. Then, for* $(\mathbb{P}_g \in \mathcal{P})$ *the following holds*

$$\mathbb{E}_g[(\mathbb{E}_g[\widehat{\vartheta}|u] - \vartheta(\mathbb{P}_g))^2] \leq \mathbb{E}_g[(\widehat{\vartheta} - \vartheta(\mathbb{P}_g))^2].$$

Remark 2.7.8. In fact, Theorem 2.7.7 can also be applied to any convex loss function and not just to a quadratic one.

Remark 2.7.9. Note that sufficiency only ensures that $E_g[\widehat{\vartheta}|u]$ is independent of ϑ, i.e. that it is an estimator.

Theorem 2.7.10. *(Lehmann-Scheffe) Let* $\mathbf{Y}$ *be a sample from an unknown measure* $\mathbb{P}_g \in \mathcal{P}$ *and* $\widehat{\vartheta}$ *be an unbiased estimator of* ϑ *with* $\mathbb{E}[\widehat{\vartheta}^2] < \infty$*. If* u *is a complete and sufficient statistic for* $\mathcal{P}$*, then* $E_g[\widehat{\vartheta}|u]$ *is an UMVU estimator for* ϑ*. Furthermore,* $E_g[\widehat{\vartheta}|u]$ *is the unique UMVU estimator.*

Proof. By Rao-Blackwell theorem, for any $\mathbb{P}_g \in \mathcal{P}$

$$\mathrm{Var}_g(\mathbb{E}_g[\widehat{\vartheta}|u]) \leq \mathrm{Var}_g(\widehat{\vartheta}).$$

Let $\tilde{\vartheta}$ be any other unbiased estimator of ϑ. Then,

$$\mathbb{E}_g(\mathbb{E}_g[\widehat{\vartheta}|u] - \mathbb{E}_g[\tilde{\vartheta}|u]) = \mathbb{E}_g(\mathbb{E}_g[(\widehat{\vartheta} - \tilde{\vartheta})|u]) = 0,$$

thus, by the definition of complete statistic we have that for all $\mathbb{P}_g \in \mathcal{P}$ the following holds $\mathbb{P}_g(\mathbb{E}_g[\widehat{\vartheta}|u] = \mathbb{E}_g[\tilde{\vartheta}|u]) = 1$. Hence, $\mathbb{E}_g[\widehat{\vartheta}|u]$ is the unique UMVU estimator. □

A comprehensive discussion of these results can be found in [81]. From the Lehmann-Scheffe theorem it follows that determining the UMVU is simply a matter of finding a complete sufficient statistic for the family $\mathcal{P}$ and conditioning any unbiased estimator on it. In some models, however, biased estimators have lower mean-squared-error (MSE) because they have a smaller variance than any unbiased estimator, e.g. James-Stein estimator in

the Gaussian model that can be found in the paper of [61]. Notwithstanding, there might be an unbiased estimator with a smaller variance. We concentrate on UMVU estimators in this chapter. If the UMVU estimator exists it might be inadmissible, i.e. there might be another estimator which has uniformly smaller risk but which is biased. It is worth mentioning, that in Bayesian statistics the so-called Bayes estimators are virtually always admissible. The Bayesian approach is when the unknown parameter ϑ is viewed as a random variable, that is being endowed with some prior distribution. We can then take expectations over the risk function with respect to that prior. Nonetheless, while Bayes estimators, as a class of estimators, possess important properties (see e.g. [111]), in any given situation the assumption of the existence of an *a priori* distribution may not be valid, or, if such a distribution exists, it may be unknown.

A closely related minimax approach, which we focus on in this chapter, is to model the choice of ϑ in an adversarial manner, and to compare estimators based on their worst-case performances. More precisely, for each estimator $\widehat{\vartheta}$, we compute its worst-case risk $\sup_{\mathbb{P}_g \in \mathcal{P}} \mathbb{E}_g[\widehat{\vartheta} - \vartheta]^2$ and rank estimators according to this property. The estimator, which is optimal in the minimax sense, achieves the so-called minimax risk that is being defined in the following way

$$\mathcal{M}(\vartheta(\mathcal{P})) = \inf_{\widehat{\vartheta}} \sup_{\mathbb{P}_g \in \mathcal{P}} \mathbb{E}_g[(\widehat{\vartheta} - \vartheta(\mathbb{P}_g))^2],$$

where the infimum extends over all possible estimators. We call the minimax optimal rate the rate of convergence of the minimax risks. The minimax approach ensures that an estimator performs best in the worst possible case allowed in the problem. We refer to [86] for a comprehensive summary of results about minimax estimation.

Bibliography

[1] Alsmeyer, G., and F. Buckmann (2017): "Stability of perpetuities in Markovian environment," *Journal of Difference Equations and Applications*, 23(4), 699–740.

[2] Ash, R. B., B. Robert, C. A. Doleans-Dade, and A. Catherine (2000): *Probability and measure theory*. Academic Press.

[3] Athreya, K. B., and S. N. Lahiri (2006): *Measure theory and probability theory*. Springer Science & Business Media.

[4] Baddeley, A., I. Bárány, and R. Schneider (2007): "Spatial point processes and their applications," *Stochastic Geometry: Lectures Given at the CIME Summer School Held in Martina Franca, Italy, September 13–18, 2004*, pp. 1–75.

[5] Baraud, Y., L. Birgé, and M. Sart (2017): "A new method for estimation and model selection: ρ-estimation," *Inventiones Mathematicae*, 207(2), 425–517.

[6] Barber, D. (2012): *Bayesian reasoning and machine learning*. Cambridge University Press.

[7] Baron, M., and A. L. Rukhin (2001): "Perpetuities and asymptotic change-point analysis," *Statistics & probability letters*, 55(1), 29–38.

[8] Beare, B. K., and A. A. Toda (2017): "Geometrically stopped Markovian random growth processes and Pareto tails," *arXiv preprint arXiv:1712.01431*.

[9] Bédard, D., and D. Dufresne (2001): "Pension funding with moving average rates of return," *Scandinavian Actuarial Journal*, 2001(1), 1–17.

[10] Blanchet, J. H., and K. Sigman (2011): "On exact sampling of stochastic perpetuities," *Journal of Applied Probability*, 48A, 165–182.

[11] Brandt, A. (1986): "The stochastic equation $Y_{n+1} = A_n Y_n + B_n$ with stationary coefficients," *Advances in Applied Probability*, 18(1), 211–220.

[12] Briand, L. C., V. R. Basili, and W. M. Thomas (1992): "A pattern recognition approach for software engineering data analysis," *IEEE Transactions on Software Engineering*, 18(11), 931–942.

[13] Brown, B. M. (1971): "Martingale central limit theorems," *The Annals of Mathematical Statistics*, 42(1), 59–66.

[14] Brunel, V. E. (2014): "Non parametric estimation of convex bodies and convex polytopes," Ph.D. thesis.

[15] Buraczewski, D., and M. Letachowicz (2012): "Limit theorems for stochastic recursions with Markov dependent coefficients," 128(2), 263–276.

[16] Campbell, J. Y., A. W.-C. Lo, and A. C. MacKinlay (1997): *The econometrics of financial markets*. princeton University press.

[17] Carbonell, J. G., R. S. Michalski, and T. M. Mitchell (1983): "An overview of machine learning," in *Machine Learning*, pp. 3–23. Elsevier.

[18] Chernozhukov, V., and H. Hong (2004): "Likelihood estimation and inference in a class of nonregular econometric models," *Econometrica*, 72(5), 1445–1480.

[19] Chichignoud, M. (2012): "Minimax and minimax adaptive estimation in multiplicative regression: locally bayesian approach," *Probability Theory and Related Fields*, 153(3-4), 543–586.

[20] Cuevas, A., M. Febrero, and R. Fraiman (2002): "Linear functional regression: the case of fixed design and functional response," *Canadian Journal of Statistics*, 30(2), 285–300.

[21] Daley, D. J., and D. Vere-Jones (2003): "An introduction to the theory of point processes, volume 1: Elementary theory and methods," *Verlag New York Berlin Heidelberg: Springer*.

[22] Dassios, A., Y. Qu, and J. W. Lim (2019): "Exact simulation of generalised Vervaat perpetuities," *Journal of Applied Probability*, 56(1), 57–75.

[23] de Saporta, B. (2005): "Tail of the stationary solution of the stochastic equation $Y_{n+1} = a_n Y_n + b_n$ with Markovian coefficients," *Stochastic Processes and their Applications*, 115(12), 1954–1978.

[24] Decamps, M., A. D. Schepper, M. Goovaerts, and W. Schoutens (2005): "A note on some new perpetuities," *Scandinavian Actuarial Journal*, 2005(4), 261–270.

[25] Devroye, L., and O. Fawzi (2010): "Simulating the Dickman distribution," *Statistics & probability letters*, 80(3-4), 242–247.

[26] Dubois, J.-P., and O. M. Abdul-Latif (2005): "Detection of Ultrasonic Images in the Presence of a Random Number of Scatterers: A Statistical Learning Approach.," pp. 326–329.

[27] Dufresne, D. (1990): "The distribution of a perpetuity, with applications to risk theory and pension funding," *Scandinavian Actuarial Journal*, 1990(1), 39–79.

[28] ——— (1996): "On the stochastic equation L (X)= L [B (X+ C)] and a property of gamma distributions," *Bernoulli*, 2(3), 287–291.

[29] ——— (1998): "Algebraic properties of beta and gamma distributions, and applications," *Advances in Applied Mathematics*, 20(3), 285–299.

[30] ——— (2010): "G distributions and the beta-gamma algebra," *Electronic Journal of Probability*, 15, 2163–2199.

[31] Embrechts, P., and C. M. Goldie (1994): "Perpetuities and random equations," *Asymptotic statistics (Prague, 1993)*, pp. 75–86.

[32] Fill, J., and M. Huber (2010): "Perfect simulation of Vervaat perpetuities," *Electronic Journal of Probability*, 15, 96–109.

[33] Fleming, T. R., and D. P. Harrington (2011): *Counting processes and survival analysis*, vol. 169. John Wiley & Sons.

[34] Garcia, N. L., and T. G. Kurtz (2006): "Spatial birth and death processes as solutions of stochastic equations," *ALEA Lat. Am. J. Probab. Math. Stat*, p. 2249658.

[35] Gayraud, G. (1997): "Estimation of functionals of density support," *Mathematical Methods of Statistics*, 6(1), 26–46.

[36] Gelfand, A. E., P. Diggle, P. Guttorp, and M. Fuentes (2010): *Handbook of spatial statistics*. CRC press.

[37] Gijbels, I., E. Mammen, B. U. Park, and L. Simar (1999): "On estimation of monotone and concave frontier functions," *Journal of the American Statistical Association*, 94(445), 220–228.

[38] Girard, S., A. Iouditski, and A. V. Nazin (2005): "L 1-optimal nonparametric frontier estimation via linear programming," *Automation and Remote Control*, 66(12), 2000–2018.

[39] Girard, S., and P. Jacob (2003): "Projection estimates of point processes boundaries," *Journal of Statistical Planning and Inference*, 116(1), 1–15.

[40] Goldenshluger, A., and O. Lepski (2011): "Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality," *The Annals of Statistics*, 39(3), 1608–1632.

[41] Goldenshluger, A., and A. Zeevi (2006): "Recovering convex boundaries from blurred and noisy observations," *The Annals of Statistics*, 34(3), 1375–1394.

[42] Goldie, C. M. (1991): "Implicit renewal theory and tails of solutions of random equations," *The Annals of Applied Probability*, pp. 126–166.

[43] Goldie, C. M., and R. Grübel (1996): "Perpetuities with thin tails," *Advances in Applied Probability*, 28(2), 463–480.

[44] Goldie, C. M., and R. A. Maller (2000): "Stability of perpetuities," *Annals of Probability*, pp. 1195–1218.

[45] Gollier, C. (2004): "Maximizing the expected net future value as an alternative strategy to gamma discounting," *Finance Research Letters*, 1(2), 85–89.

[46] ——— (2010): "Expected net present value, expected net future value, and the Ramsey rule," *Journal of Environmental Economics and Management*, 59(2), 142–148.

[47] Graybill, F. A. (1976): *Theory and application of the linear model*, vol. 183. Duxbury press North Scituate, MA.

[48] Grey, D. (1994): "Regular variation in the tail behaviour of solutions of random difference equations," *The Annals of Applied Probability*, 4(1), 169–183.

[49] Grincevičjus, A. (1975): "On a limit distribution for a random walk on lines," *Litovsk. Mat. Sb*, 15(4), 79–91.

[50] Grübel, R., and U. Rösler (1996): "Asymptotic distribution theory for Hoare's selection algorithm," *Advances in Applied Probability*, 28(1), 252–269.

[51] Häggström, O. (2005): "On the central limit theorem for geometrically ergodic Markov chains," *Probability theory and related fields*, 132(1), 74–82.

[52] Hall, P., and B. U. Park (2004): "Bandwidth choice for local polynomial estimation of smooth boundaries," *Journal of Multivariate Analysis*, 91(2), 240–261.

[53] Harrison, J. M. (1977): "Ruin problems with compounding assets," *Stochastic Processes and their Applications*, 5(1), 67–79.

[54] Hitczenko, P., and J. Wesołowski (2009): "Perpetuities with thin tails revisited," *The Annals of Applied Probability*, 19(6), 2080–2101.

[55] Hoare, C. A. (1961): "Algorithm 65: find," *Communications of the ACM*, 4(7), 321–322.

[56] Hunt, G. A. (1957): "Markoff processes and potentials I," *Illinois J. Math.*, 1(1), 44–93.

[57] Hunter, J. D. (2007): "Matplotlib: A 2D graphics environment," *Computing in Science & Engineering*, 9(3), 90–95.

[58] Hwang, H.-K., and T.-H. Tsai (2002): "Quickselect and the Dickman function," *Combinatorics, Probability and Computing*, 11(4), 353–371.

[59] Illian, J. B., and D. F. Burslem (2017): "Improving the usability of spatial point process methodology: an interdisciplinary dialogue between statistics and ecology," *AStA Advances in Statistical Analysis*, 101(4), 495–520.

[60] Jacob, P., and C. Suquet (1995): "Estimating the edge of a Poisson process by orthogonal series," *Journal of Statistical Planning and Inference*, 46(2), 215–234.

[61] James, W., and C. Stein (1961): "Estimation with quadratic loss," in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 361–379.

[62] Jirak, M., A. Meister, and M. Reiss (2014): "Adaptive function estimation in nonparametric regression with one-sided errors," *The Annals of Statistics*, 42(5), 1970–2002.

[63] Jones, E., T. Oliphant, and P. Peterson (2001): "SciPy: Open source scientific tools for Python," .

[64] Kallenberg, O. (1997): *Foundations of modern probability*, vol. 2. Springer.

[65] Kardaras, C., and S. Robertson (2017): "Continuous-time perpetuities and time reversal of diffusions," *Finance and Stochastics*, 21(1), 65–110.

[66] Karr, A. (2017): *Point processes and their statistical inference*. Routledge.

[67] Kesten, H. (1973): "Random difference equations and renewal theory for products of random matrices," *Acta Mathematica*, 131(1), 207–248.

[68] Kirschenhofer, P., and H. Prodinger (1998): "Comparisons in Hoare's Find algorithm," *Combinatorics, Probability and Computing*, 7(1), 111–120.

[69] Knape, M., and R. Neininger (2008): "Approximating perpetuities," *Methodology and Computing in Applied Probability*, 10(4), 507–529.

[70] Knuth, D. E. (1972): "Mathematical analysis of algorithms, Information Processing 71," *Invited Papers-Foundations North-Holland Publishing Company*.

[71] Kolmogoroff, A. (1936): "Zur Theorie der Markoffschen Ketten," *Mathematische Annalen*, 112, 155–160.

[72] ——— (1937): "Zur Umkehrbarkeit der statistischen Naturgesetze," *Mathematische Annalen*, 113(1), 766–772.

[73] Korostelëv, A. P., L. Simar, and A. B. Tsybakov (1995): "Efficient estimation of monotone boundaries," *The Annals of Statistics*, 23(2), 476–489.

[74] Korostelev, A. P., and A. B. Tsybakov (1993): "Estimation of the density support and its functionals," *Problemy Peredachi Informatsii*, 29(1), 3–18.

[75] Korostelev, A. P., and A. B. Tsybakov (2012): *Minimax theory of image reconstruction*, vol. 82. Springer Science & Business Media.

[76] Kulesza, A., and B. Taskar (2012): "Determinantal point processes for machine learning," *Foundations and Trends® in Machine Learning*, 5(2–3), 123–286.

[77] Kutoyants, Y. A. (2012): *Statistical inference for spatial Poisson processes*, vol. 134. Springer Science & Business Media.

[78] Laurent, B., C. Ludena, and C. Prieur (2008): "Adaptive estimation of linear functionals by model selection," *Electronic Journal of Statistics*, 2, 993–1020.

[79] Le, L. (1964): "Sufficiency and approximate sufficiency," *The Annals of Mathematical Statistics*, pp. 1419–1455.

[80] Le Cam, L., and G. L. Yang (2012): *Asymptotics in statistics: some basic concepts*. Springer Science & Business Media.

[81] Lehmann, E. L., and G. Casella (2006): *Theory of point estimation*. Springer Science & Business Media.

[82] Lepskii, O. (1991): "On a problem of adaptive estimation in Gaussian white noise," *Theory of Probability & Its Applications*, 35(3), 454–466.

[83] Luca, S. E., M. A. Pimentel, P. J. Watkinson, and D. A. Clifton (2018): "Point process models for novelty detection on spatial point patterns and their extremes," *Computational Statistics & Data Analysis*, 125, 86–103.

[84] Lund, J., and M. Rudemo (2000): "Models for point processes observed with noise," *Biometrika*, 87(2), 235–249.

[85] Mahmoud, H. M., R. Modarres, and R. T. Smythe (1995): "Analysis of quickselect: An algorithm for order statistics," *RAIRO-Theoretical Informatics and Applications*, 29(4), 255–276.

[86] Mammen, E., and A. B. Tsybakov (1995): "Asymptotical minimax recovery of sets with smooth boundaries," *The Annals of Statistics*, pp. 502–524.

[87] 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.

[88] Meyn, S. P., and R. L. Tweedie (2012): *Markov chains and stochastic stability*. Springer Science & Business Media.

[89] Moyeed, R., and A. J. Baddeley (1991): "Stochastic approximation of the MLE for a spatial point pattern," *Scandinavian Journal of Statistics*, pp. 39–50.

[90] Müller, U. U. (2009): "Estimating linear functionals in nonlinear regression with responses missing at random," *The Annals of Statistics*, 37(5A), 2245–2277.

[91] Müller, U. U., A. Schick, and W. Wefelmeyer (2004): "Estimating linear functionals of the error distribution in nonparametric regression," *Journal of Statistical Planning and Inference*, 119(1), 75–93.

[92] Nyrhinen, H. (2001): "Finite and infinite time ruin probabilities in a stochastic economic environment," *Stochastic Processes and Their Applications*, 92(2), 265–285.

[93] Paulsen, J. (1993): "Risk theory in a stochastic economic environment," *Stochastic Processes and Their Applications*, 46(2), 327–361.

[94] ——— (1998): "Sharp conditions for certain ruin in a risk process with stochastic return on investments," *Stochastic Processes and Their Applications*, 75(1), 135–148.

[95] Paulsen, J., and A. Hove (1999): "Markov chain Monte Carlo simulation of the distribution of some perpetuities," *Advances in Applied Probability*, 31(1), 112–134.

[96] Pollak, M., and D. Siegmund (1985): "A diffusion process and its applications to detecting a change in the drift of Brownian motion," *Biometrika*, 72(2), 267–280.

[97] Preston, C. (1975): "Spatial birth and death processes," *Advances in Applied Probability*, 7(3), 465–466.

[98] Rachev, S. T., and G. Samorodnitsky (1995): "Limit laws for a stochastic process and random recursion arising in probabilistic modelling," *Advances in a Applied Probability*, 27(1), 185–202.

[99] Rényi, A., and R. Sulanke (1964): "Über die konvexe Hülle von n zufällig gewählten Punkten. II," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 3(2), 138–147.

[100] Resnick, S. (2003): *A probability path*. Birkhauser Verlag AG.

[101] Roberts, G., and J. Rosenthal (1997): "Geometric ergodicity and hybrid Markov chains," *Electronic Communications in Probability*, 2, 13–25.

[102] Roitershtein, A., et al. (2007): "One-dimensional linear recursions with Markov-dependent coefficients," *The Annals of Applied Probability*, 17(2), 572–608.

[103] Schrödinger, E. (1931): *Über die Umkehrung der Naturgesetze*, Sitzungsberichte der Preussischen Akademie der Wissenschaften. Physikalisch-mathematische Klasse. Akad. d. Wissenschaften.

[104] Sericola, B. (2013): *Markov chains: theory and applications*. John Wiley & Sons.

[105] Severini, T. A., and G. Tripathi (2012): "Efficiency bounds for estimating linear functionals of nonparametric regression models with endogenous regressors," *Journal of Econometrics*, 170(2), 491–498.

[106] Shalizi, C., and A. Kontorovich (2007): "Almost None of the Theory of Stochastic Processes A Course on Random Processes, for Students of Measure-Theoretic Probability, with a View to Applications in Dynamics and Statistics," .

[107] Simar, L., and P. W. Wilson (1998): "Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models," *Management Science*, 44(1), 49–61.

[108] Stoyan, D., and H. Stoyan (1994): *Fractals, random shapes, and point fields: methods of geometrical statistics*, vol. 302. John Wiley & Sons Inc.

[109] Tsybakov, A. B. (2008): *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edn.

[110] Tweedie, R. (2001): "Drift conditions and invariant measures for Markov chains," *Stochastic Processes and Their Applications*, 92(2), 345–354.

[111] van der Pas, S., B. Szabó, and A. van der Vaart (2017): "Uncertainty quantification for the horseshoe (with discussion)," *Bayesian Analysis*, 12(4), 1221–1274.

[112] Vervaat, W. (1979): "On a stochastic difference equation and a representation of non–negative infinitely divisible random variables," *Advances in Applied Probability*, 11(4), 750–783.

[113] Vo, B.-N., N. Dam, D. Phung, Q. N. Tran, and B.-T. Vo (2018): "Model-based learning for point pattern data," *Pattern Recognition*, 84, 136–151.

[114] Wang, S. (2003): "Adaptive non-parametric efficiency frontier analysis: a neural-network-based model," *Computers & Operations Research*, 30(2), 279–295.

[115] Wilson, P. W. (2003): "Testing independence in models of productive efficiency," *Journal of Productivity Analysis*, 20(3), 361–390.

[116] Yor, M. (2001): "Bessel processes, Asian options, and perpetuities," in *Exponential Functionals of Brownian Motion and Related Processes*. Springer.